**大 连 理 工 大 学 本 科 外 文 翻 译**

**全景分割**

**Panoptic Segmentation**

学 部（院）： 软件学院

专 业： 数字媒体技术

学 生 姓 名： 王丹丹

学 号： 201692354

指 导 教 师： 陆坤

完 成 日 期： 2020年2月

大连理工大学

**全景分割**

Alexander Kirillov1,2 Kaiming He1 Ross Girshick1 Carsten Rother2 Piotr Dollár1

1Facebook人工智能研究院 2HCI/IWR，德国海德堡大学

# 摘要

我们提出并研究了一个我们命名为全景分割（PS）的任务。全景分割结合了语义分割（为每个像素指定一个类标签）和实例分割（检测并分割每个对象实例）这两个典型的不同任务。提出的任务要求生成一个丰富而完整的连贯场景分割，这是迈向现实世界视觉系统的重要一步。虽然计算机视觉的早期工作解决了相关的图像/场景解析任务，但这些任务目前并不流行，可能是由于缺乏适当的度量或相关的识别挑战。为了解决这个问题，我们提出了一种新的全景质量（PQ）度量，它以可解释和统一的方式捕获所有类（stuff和thing）的性能。使用所提出的度量，我们在三个现有的数据集上对PS的人类和机器性能进行严格的研究，揭示了关于此任务的有趣的见解。我们的工作目标是在一个更统一的图像分割的观点中恢复社区的兴趣。

关键词：计算机视觉，语义分割，实例分割

目录

[1 背景介绍](#_Toc17439_WPSOffice_Level1) [1](#_Toc17439_WPSOffice_Level1)

[2 相关工作](#_Toc916_WPSOffice_Level1) [4](#_Toc916_WPSOffice_Level1)

[3 全景分割格式](#_Toc19748_WPSOffice_Level1) [6](#_Toc19748_WPSOffice_Level1)

[4 全景分割指标](#_Toc2156_WPSOffice_Level1) [7](#_Toc2156_WPSOffice_Level1)

[4.1 分割匹配](#_Toc916_WPSOffice_Level2) [7](#_Toc916_WPSOffice_Level2)

[4.2 PQ计算](#_Toc19748_WPSOffice_Level2) [8](#_Toc19748_WPSOffice_Level2)

[4.3 与现有指标的比较](#_Toc2156_WPSOffice_Level2) [9](#_Toc2156_WPSOffice_Level2)

[5 全景分割数据集](#_Toc22300_WPSOffice_Level1) [11](#_Toc22300_WPSOffice_Level1)

1 背景介绍

在计算机视觉的早期，thing——如人、动物、工具等可计数的物体——获得了主要的注意力。对这一趋势的智慧提出质疑后，Adelson[1]提升了研究识别物体的系统的重要性——识别具有相似纹理或材料的无定形区域，如草地、天空、道路。这种stuff与thing的二分法一直持续到今天，反映在视觉识别任务的划分和为stuff与thing任务开发的专门算法中。

研究stuff通常被称为语义分割，见图1b。由于素材是无定形的、不可数的，这项任务被定义为简单地为图像中的每个像素分配一个类标签（注意，语义分割将thing类视为stuff）。相比之下，研究thing通常被表述为对象检测或实例分割任务，目标是检测每个对象，并分别用边界框或分割掩码对其进行描绘，见图1c。这两个视觉识别任务的数据集、详细信息和度量标准看似相关，但实际上是非常不同的。

图1.1 对于给定的(a)图像，我们展示了以下真实值：(b)语义分割（每个像素类标签），(c)实例分割（每个对象掩码和类标签），(d)提出的全景分割任务（每个像素类+实例标签）。PS任务：(1)包含stuff和thing类，(2)使用简单但通用的格式，以及(3)为所有类引入统一的评估度量。全景分割将语义分割和实例分割结合起来，我们期望统一的任务将带来新的挑战，并带来新的创新方法。

语义分割和实例分割之间的分裂导致了用于这些任务的方法中的平行裂痕。stuff分类器通常建立在具有扩张[52、5]的全卷积网络[30]上，而物体检测器通常使用物体提议[15]并基于区域[37、14]。在过去的十年中，这些任务的总体算法进展令人难以置信，但是，孤立地专注于这些任务可能会忽略一些重要的事情。

一个自然的问题浮出水面：stuff与thing之间能否实现和解？能够生成丰富而连贯的场景分割的统一视觉系统最有效的设计是什么？考虑到它们与诸如自动驾驶或增强现实之类的现实应用相关性，这些问题尤其重要。

有趣的是，尽管语义和实例分割在当前工作中占主导地位，但在深度学习之前的时代，人们对使用各种名称描述的联合任务感兴趣，诸如场景解析[42]，图像解析[43]或整体场景理解[51]。尽管具有实际意义，但可能由于缺乏适当的度量标准或识别挑战而导致该一般方向目前并不流行。

在我们的工作中，我们旨在复兴这一方向。我们提出一项任务：（1）同时包含stuff和thing类；（2）使用简单但通用的输出格式；（3）引入统一的评估指标。为了清楚地与先前的工作消除歧义，我们将产生的任务称为全景分割（PS）。 “全景”的定义是“包括在一个视图中可见的所有内容”，在我们的上下文中，全景是指统一的全局细分视图。

我们用于全景分割的**任务格式**很简单：必须为图像的每个像素分配一个语义标签和一个实例ID。具有相同标签和ID的像素属于同一物体；对于stuff标签，实例ID被忽略。参见图1d的可视化。以前已经采用了这种格式，特别是通过产生不重叠的实例分段的方法[18、28、2]。我们将其用于包含stuff和things的联合任务中。

全景分割的基本方面是用于评估的任务指标。尽管许多现有度量标准都可以用于语义或实例分割，但是这些度量标准分别最适合于stuff或things，但不适用于两者。我们认为使用不相交的度量标准是社区通常单独研究stuff和thing分割的主要原因之一。为了解决这个问题，我们在§4中引入了全景质量（PQ）指标。 PQ简单且内容丰富，最重要的是，PQ可用于以统一的方式衡量stuff和things的性能。我们希望，我们提出的的联合指标将有助于更广泛地采用联合任务。

全景分割任务既包括语义分割又包括实例分割，但引入了新的算法挑战。与语义分割不同，它需要区分各个对象实例；这对全卷积网络构成了挑战。与实例分割不同，对象分割必须是不重叠的。这给在每个对象上独立运行的基于区域的方法提出了挑战。生成可解决stuff和things之间矛盾的连贯图像分割是迈向现实应用的重要一步。

由于PS的真实值和算法格式必须采用相同的形式，因此我们可以对全景分割中的人类一致性进行详细的研究。这使我们可以更详细地了解PQ指标，包括识别与细分，stuff与things性能的详细细分。此外，测量人类PQ有助于我们了解机器性能。这很重要，因为它将使我们能够监视PS的各种数据集上的性能饱和。

最后，我们对PS的机器性能进行了初步研究。为此，我们定义了一种简单且可能不太理想的启发式方法，该方法通过一系列合并它们的输出的后处理步骤（实质上是一种非最大抑制的复杂形式），来将两个独立系统的输出合并起来以进行语义和实例分割 。我们的启发式方法为PS建立了基线，并让我们深入了解了PS所带来的主要算法挑战。

我们在同时具有stuff和things注释的三个流行的分割数据集上研究了人类和机器的性能。这包括Cityscapes [6]，ADE20k [55]和Mapillary Vistas [35]数据集。对于这些数据集，我们直接从挑战组织者那里获得了最新技术方法的结果。将来，我们将把分析范围扩展到正在进行stuff注解[4]的COCO [25]。我们在这些数据集上的结果一起为研究人类和机器在全景分割方面的性能奠定了坚实的基础。

COCO [25]和Mapillary Vistas [35]都将全景分割任务作为其在ECCV 2018上的识别挑战中的轨道之一。我们希望在这些流行的识别数据集上将PS与实例和语义分割轨道一起显示，将有助于广泛采用提出的联合任务。

2 相关工作

在整个计算机视觉的历史中，新颖的数据集和任务都起着关键作用。它们有助于促进进步并实现我们领域的突破，同样重要的是，它们有助于我们衡量和认识社区正在取得的进步。例如，ImageNet [38]帮助推动了深度学习技术在视觉识别领域的最新普及[20]，并举例说明了数据集和任务可以具有的潜在变革能力。我们引入全景分割任务的目标是相似的：挑战我们的社区，以新颖的方向推动研究，并实现预期和意想不到的创新。接下来，我们将回顾相关任务。

**物体检测任务**。早期使用ad-hoc数据集进行人脸检测的工作（例如[44，46]）有助于普及边界框物体检测。后来，行人检测数据集[8]帮助推动了该领域的进步。 PASCAL VOC数据集[9]在更具挑战性的图像上将任务升级为一组更多样化的常规对象类。最近，COCO数据集[25]将检测推向了实例分割的任务。通过框架化此任务并提供高质量的数据集，COCO帮助定义了一个新的令人振奋的研究方向，并在实例分割方面带来了许多近期突破[36，24，14]。我们对全景分割的总体目标是相似的。

**语义分割任务**。语义分割数据集具有悠久的历史[39，26，9]，并有助于推动关键的创新（例如，使用[26，9]开发了全卷积网络[30]）。这些数据集包含stuff和thing类，但不区分单个对象实例。最近，该领域已经看到了许多新的分割数据集，包括Cityscapes [6]，ADE20k [55]和Mapillary Vistas [35]。这些数据集实际上既支持语义分割又支持实例分割，并且每个数据集都选择对这两项任务有单独的跟踪。重要的是，它们包含PS所需的所有信息。换句话说，全景分割任务可以在这些数据集上启动，而无需任何新的数据收集。

**多任务学习**。随着深度学习在许多视觉识别任务中的成功应用，人们对具有广泛能力并可以在一个框架中解决多种视觉问题的多任务学习方法产生了浓厚的兴趣[19、32、34]。例如，UberNet [19]使用单个网络解决了多个低级到高级可视任务，包括对象检测和语义分割。尽管在这个领域有很大的兴趣，但我们强调全景全景图不是多任务问题，而是图像分割的统一视图。具体来说，多任务设置允许事物的独立且可能不一致的输出，而PS需要单个一致的场景分割。

**联合分割任务**。在前深度学习时代，人们对生成连贯的场景解释非常感兴趣。关于图像解析的开创性工作[43]提出了一个通用的贝叶斯框架，以联合建模分割，检测和识别。后来，基于图形模型的方法研究了一致的stuff和thing分割[51、41、42、40]。尽管这些方法具有共同的动机，但并未就任务定义达成共识，并且使用了不同的输出格式和变化的评估指标，包括用于评估stuff和thing类别的结果的独立指标。近年来，也许由于这些原因，该方向变得不太流行。

在我们的工作中，我们致力于恢复这个通用的方向，但与早期的工作相反，我们专注于任务本身。具体来说，如前所述，PS：（1）处理stuff和thing类别，（2）使用简单格式，（3）引入stuff和thing的统一指标。先前关于联合分割的工作使用不同的格式和不相交的度量来评估stuff和things。生成非重叠实例分割的方法[18、3、28、2]使用与PS相同的格式，但是这些方法通常仅处理thing类。通过使用简单的格式处理stuff和things，并引入统一的度量标准，我们希望鼓励更广泛地采用联合任务。

**无模式分割任务**。在[56]中，物体被无模式注释：标记了每个区域的整个范围，而不仅仅是可见区域。我们的工作重点是对所有可见区域的分割，但是将全景分割扩展到无模态环境是未来工作的有趣方向。

3 全景分割格式

**任务格式**。全景分割的格式很容易定义。给定一个由编码的L个语义类预定义的集合，这项任务需要一个全景分割算法来将图像的每个像素映射到一个对，表示像素的语义类，表示它的实例id。将同一类的像素分成不同的分割。真实值注释的编码方式是相同的。可以为模糊或不合格的像素分配特殊的空白标签，即不是所有的像素都必须有一个语义标签。

**Stuff和thing标签。**语义标签集由子集和组成，因此，。这些子集分别对应于stuff标签和thing标签。当一个像素被标记为，这与它对应的实例id是不相关的。也就是说，对于stuff类，所有的像素都属于相同的实例（例如：相同的天空）。否则，所有由相同的赋值的像素，这里的，属于同一个实例（例如：同一辆汽车），相反，属于单个实例的所有像素必须具有相同的。像以前的数据集中一样，选择哪个类是stuff还是thing是由数据集创建者设计的选择。

**与语义分割的关系**。PS任务格式是语义分割格式的严格概括。实际上，两项任务都要求为图像中的每个像素分配一个语义标签。如果基本事实未指定实例，或者所有类都是stuff类，则任务格式相同（尽管任务指标不同）。此外，包含thing类（每个图像可能具有多个实例）可以区分任务。

**与实例分割的关系**。实例分割任务需要一种方法来分割图像中的每个对象实例。但是，它允许重叠的分割，而全景分割任务只允许将一个语义标签和一个实例ID分配给每个像素。因此，对于PS，通过构造不可能有任何重叠。在下一节中，我们将说明这种差异在绩效评估中起着重要作用。

**置信分数**。与语义分割类似，但是与实例分割不同，我们不需要与PS的每个分割相关联的置信度得分。这使全景任务相对于人和机器是对称的：两者都必须生成相同类型的图像注释。这也使评估PS的人类一致性变得简单。这与实例分割相反，实例分割不容易接受此类研究，因为人类注释者没有提供明确的置信度得分（尽管可以测量单个精度/召回点）。我们注意到，置信度得分为下游系统提供了更多信息，这可能是有用的，因此，仍可能需要PS算法在某些设置下生成置信度得分。

4 全景分割指标

在本节中，我们介绍了一种用于全景分割的新指标。首先，我们注意到现有度量专用于语义或实例分割，不能用于评估涉及stuff和thing类的联合任务。之前有关联合分割的工作通过使用独立的指标（例如[51、41、42、40]）评估stuff和thing的性能，从而解决了这一问题。但是，这给算法开发带来了挑战，使比较更加困难，并阻碍了交流。我们希望为stuff和things引入统一的度量标准将鼓励对统一任务的研究。

在进行进一步的详细介绍之前，我们首先要确定以下适合于PS的度量标准的度量标准：

**完整性**。度量标准应以统一的方式处理stuff和thing类，捕获任务的所有方面。

**可解释性**。我们寻求一种具有可识别意义的度量标准，以促进沟通和理解。

**简单性**。此外，该指标应易于定义和实施。这样可以提高透明度，并易于重新实现。与此相关的是，该度量应该有效地计算以实现快速评估。

在这些原则的指导下，我们提出了一种新的全景质量（PQ）指标。 PQ衡量相对于真实值的预测全景分割的质量。它涉及两个步骤：（1）分割匹配和（2）给定匹配项的PQ计算。我们接下来描述每个步骤，然后返回与现有指标的比较。

4.1 分割匹配

我们指定，只有当预测分割和真实值分割在并集上的交集（IoU）严格大于0.5时，预测分割和真实值分割才能匹配。这一要求，加上全景分割的不重叠特性，给出了一个独特的匹配：每个真实值分割最多只能有一个预测分割匹配。

定理1。给定图像的预测和真实值全景分割，每个真实值分割最多可以有一个对应的IoU严格大于0.5的预测分割，反之亦然。

证明。设为真实值分割，和为预测分割。根据定义，（它们不重叠）。由于，我们得到以下结果：

对求和，由于，，我们得到：

图2 图像的真实值和预测全景分割的玩具插图。相同颜色的分割对的IoU大于0.5，因此是匹配的。我们展示了如何将person类的分割划分为真阳性TP、假阴性FN和假阳性FP。

因此，如果，那么一定小于0.5。逆转和的作用可以证明，只有一个真实值分割可以具有预测分割严格大于0.5的 IoU。

匹配的要求必须具有大于0.5的IoU，这反过来产生唯一的匹配定理，实现了我们所期望的两个特性。首先，它是简单而有效的，因为对应关系是唯一的，而且很容易获得。其次，它是可解释且易于理解的（并且不需要像这些类型的度量通常的情况那样解决复杂的匹配问题[13，50]）。

注意，由于唯一性属性，对于IoU>0.5，任何合理的匹配策略（包括贪婪和最优）都将产生相同的匹配。对于较小的IoU，需要其他匹配技术；然而，在实验中，我们将表明，由于IoU≤0.5的匹配在实践中是罕见的，因此不需要较低的阈值。

4.2 PQ计算

我们独立计算每一个类的PQ，并对类进行平均。这使得PQ对类不平衡。对于每一类，唯一匹配将预测和真实值分割分为三组：真阳性（TP）、假阳性（FP）和假阴性（FN），分别表示匹配的分割对、不匹配的预测分割和不匹配的真实值分割。一个例子如图2所示。给定这三组，PQ定义为：

检查后PQ很直观：是匹配的分割的平均IoU，而被添加到分母来惩罚没有匹配项的分割。注意无论面积大小，所有分割都有同等的重要性。此外，如果我们用PQ乘以并除以TP集的大小，那么PQ可以看作是分割质量（SQ）项和识别质量（RQ）项的乘积：

这样写，RQ是熟悉的在检测设置中广泛用于质量评估[33]的F1分数[45]。SQ是简单的匹配的分割的平均IoU。我们发现PQ=SQ×RQ的分解为分析提供了依据。但是，我们注意到，这两个值并不独立，因为SQ只在匹配的段上测量。

我们对PQ的定义实现了我们的设计。它使用简单且可解释的公式以统一的方式衡量所有类的性能。最后，我们讨论如何处理无效区域和实例组[25]。

空标签。真实值中的无效标签有两个来源：（a）超出类别的像素和（b）模棱两可或未知的像素。通常我们无法区分这两种情况，因此我们不会评估空白像素的预测。具体来说：（1）匹配期间，在预测分割的真实值中被标记为空的所有像素从预测中移除，并且不影响IoU计算；（2）匹配后，包含小部分超过匹配阈值的空相素的不匹配的预测分割将被删除且不算作假阳性。最后，输出也可能包含空白像素，这些不会影响评估。

组标签。常见的注释实践[6，25]是，如果很难准确描述每个实例，则对同一语义类的相邻实例使用组标签代替实例ID。为了计算PQ：（1）在匹配期间不使用组区域，并且（2）在匹配之后，包含小部分超过匹配阈值的来自一组相同类别的的像素的不匹配的预测分割将会被移除，并且不计为误报。

4.3 与现有指标的比较

我们通过将PQ与现有的语义和实例分割度量进行比较来得出结论。

**语义分割指标**。语义分割的常用指标包括像素精度，均值精度和IoU [30]。这些指标仅基于像素输出/标签计算，而完全忽略对象级标签。例如，IoU是正确预测的像素与每个类别的预测或真实值中像素总数之间的比率。由于这些指标会忽略实例标签，因此它们不适合评估thing类。最后，请注意用于语义分割的IoU与我们的分割质量（SQ）不同，后者是根据匹配分割上的平均IoU计算得出的。

**实例分割指标**。实例分割的标准指标是平均精度（AP）[25，13]。AP要求每个对象段都有一个置信度分数来估计精度/召回曲线。请注意，尽管置信度得分对于对象检测非常自然，但它们并未用于语义分割。因此，AP不能用于测量语义分割或PS的输出（另请参见§3中对置信度的讨论）。

**全景质量**。PQ以统一的方式对待所有类（stuff和things）。我们注意到，尽管将PQ分解为SQ和RQ有助于解释结果，但PQ并不是语义和实例细分指标的组合。而是针对每个类（stuff和thing）计算SQ和RQ，并分别测量分割和识别质量。因此，PQ统一了所有类别的评估。我们通过对§7中的PQ进行严格的实验评估来支持这一主张，包括分别与IoU和AP进行语义和实例分割的比较。

5 全景分割数据集

据我们所知，只有三个公共数据集具有密集的语义和实例分割注释：Cityscapes [6]，ADE20k [55]和Mapillary Vistas [35]。我们使用所有三个数据集进行全景分割。此外，将来我们将把分析范围扩展到最近已注解[4] 1的COCO [25]。

Cityscapes[6]在城市环境中具有以自我为中心的驾驶场景的5000张图像（2975训练，500 验证和1525个测试）。它具有19个类的密集像素注释（覆盖率达97％），其中8个具有实例级分割。

ADE20k [55]具有超过25k图像（20k训练，2k 验证，3k测试），并使用开放词典标签集进行了密集注释。对于2017年Places Challenge2，选择了覆盖89％像素的100种事物和50种事物。我们在研究中使用这个封闭的词汇。

Mapillary Vistas [35]具有25k的高分辨率的街景图像（18k训练，2k验证，5k测试）。数据集的“研究版”带有28种stuff和37种thing的密集注释（98％像素覆盖率）。