Paper6：Holistic Scene Understanding for 3D Object Detection with RGBD cameras

利用RGBD相机进行三维物体检测和整体场景理解

RGBD相机：RGB-D相机是一种同时提供**彩色图像**（RGB）和**深度信息**（D）的相机系统。这种相机的工作原理是通过结合红外（IR）传感器和普通的彩色摄像头，以获取场景中的深度信息。相较于之前的单目相机最大的区别便是能够提供深度信息。

摘要：

提出了一种整体性的方法，利用2D分割、3D几何以及场景和物体之间的上下文关系。

具体而言，将**CPMC框架**扩展到3D，以生成候选立方体，并开发了一个**条件随机场**以整合来自不同来源的信息对立方体进行分类。通过这种公式化，场景分类和3D物体识别被耦合在一起，并且可以通过概率推理对二者进行联合求解。

CPMC框架：一种对RGB图像进行图像分割的框架

条件随机场：一种概率图模型，用于建模一组有依赖关系的随机变量的概率分布，通常有两组随机变量，一组是输入变量，另一组是输出变量。目标是对输出变量建模，考虑给定输入变量的条件下，输出变量的联合概率分布。

在NYU v2数据集上测试效果很好。

介绍：

对于单目相机来说，使用静止的图像来3D Objection Detection较为困难，但如果有额外的信息，如video或depth将会使得该问题变得简化。使用RGBD相机比使用单目相机在3D Objection Detection这项工作上能获得更好的结果。

物体之间的上下文关联性：空间中的物体并不是随意放置的，而是有一定物理和统计特性。例如人大概率是坐在骆驼上而不是以其他方式；床大概率靠在卧室的墙上而不是悬挂在天花板上。利用对象和环境之间的**物理关系**和**上下文关系**是实现分割和检测[9等语义任务的高精度的关键。

利用三维长方体来模拟表示世界上的物体，并模拟物体和环境之间的物理和统计相互作用以及物体之间的相互作用。

扩展CPMC框架，在点云中生成紧密包围底层3D区域的长方体候选框，可包括被遮挡的区域，并按照“物体性”对候选框进行排序。“**物体性**”是用于衡量一个区域在外观上与物体的相似度或可能性，高物体性的区域更有可能被认为含有真实的物体。

利用条件随机场建模3D物体之间的上下文关系，如物体之间的相互作用、特定类型的支撑关系、物体与场景如墙的接近程度等。

使用NYU v2 RGB-D数据集进行评估，与现有的模型相比，上下文模型展现了对物体检测的准确性、场景理解的全局性。

相关工作:

先前的目标识别方法在处理距离数据时通常基于两种假设。要么假设目标对象要么已经从背景中分割出来，要么存在详细的3D模型。但是在某些高度混乱的场景，目标对象可能无法从背景中分离出来，详细3D模型可能因为光照变化、视点变化等因素使用受到限制。本文着重于在高度混乱场景中对场景中的物体进行识别定位。

随着RGB-D数据集的发布，基于上下文的模型变得越来越普遍，这些模型结合有关物理世界的先验知识，专注于图像分割任务。

先前的工作有的基于施加统计物理约束来进行分割，考虑支撑表面和物体对象；有的基于仔细设计的2D物体外观和3D几何特征；有的根据RGB信息和深度信息进行推理；有的根据3D空间之间相互的关系，如在附近、在顶部进行分割；

本文的方法与先前的方法在建模物理关系方面有相似之处，但有一个显著不同点。本文的方法在完整对象的层面上进行推理，这些对象被表示为框选为长方体，使得能够更好地捕捉3D中物体与物体之间的统计相互作用。

方法步骤

- 自底向上产生3D候选框区域：基本计算流程依旧遵循CPMC框架，只是加入了对深度信息的考虑。简单的改进提高了边界检测的准确率2%。

$$gPb_{rgbd} = \alpha \cdot gPb_{rgb} + (1-\alpha) \cdot gPb_{depth}, \quad \alpha = 0.3$$

- 从候选区域生成候选立方体：从候选区域中选择排名前K的候选区域，这些区域是通过其“物体性”进行排名的，通过执行非极大值抑制来确保它们不过于重叠（使用0.5作为最大重叠阈值）。对于每个选定的候选区域，通过在该区域周围拟合一个3D立方体来生成候选立方体，常规方法

是将给定区域内的像素映射到2D坐标，找到包围这些坐标的最小边界立方体，为了提高鲁棒性，只找到包含95%3D点的最小立方体。大多数感兴趣对象与地面平行，引入约束条件，将问题简化为只有一个变量，沿X-Z平面移动的方向。非极大抑制：（Non-Maximum Suppression，NMS）是一种用于排除冗余目标框的技术。在目标检测中，算法通常会生成多个候选目标框，而非极大值抑制的目的是从这些候选框中选出最可能是真实目标的一个，以消除冗余的检测结果。

- 接下里需要对候选立方体进行贴标签分类，传统的依据物体特征进行分类可能会受到姿态变化、物体遮挡、光照变化等因素。本文采取条件随机场（CRF）模型，即整合了外观、集合、上下文形象来提高识别的准确性。

- 三维对象检测的上下文模型有四种潜在的函数，两种一元潜在函数，表示给定场景或对象标签的观测数据的可能性；两种二元潜在函数，表示场景和对象之间以及对象与对象之间的上下文关系。通过训练数据学习每个潜在函数的权重，这也是概率图模型的参数学习过程。 □□□

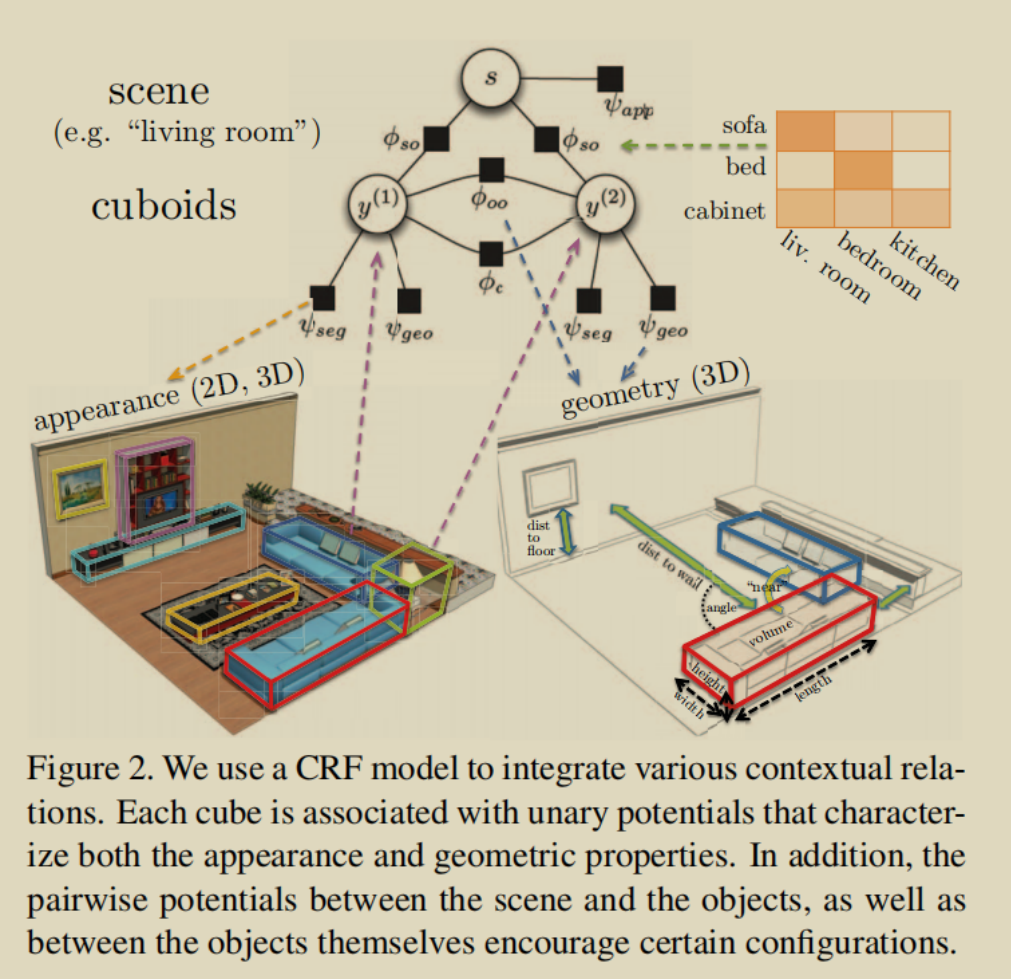

Figure 2. We use a CRF model to integrate various contextual relations. Each cube is associated with unary potentials that characterize both the appearance and geometric properties. In addition, the pairwise potentials between the scene and the objects, as well as between the objects themselves encourage certain configurations.

考虑物体的几何属性：用一个向量来表示其十个几何属性。这些属性不

仅能够了解到物体的内在属性，也能了解到物体相对于场景布局的位置。 □□□

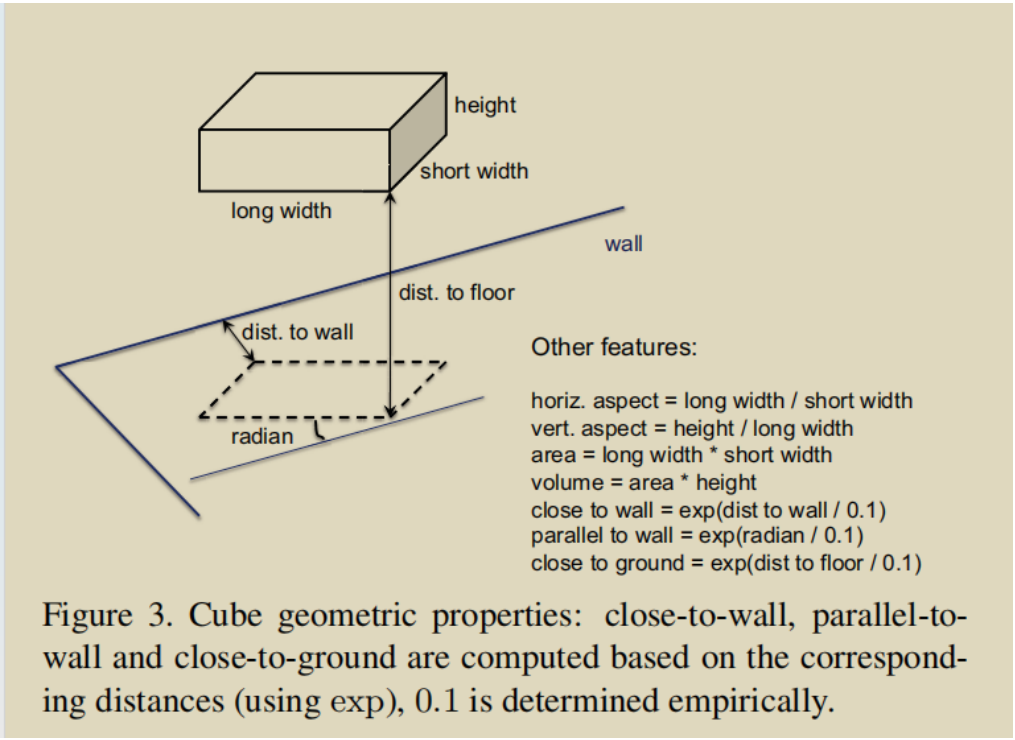

Figure 3. Cube geometric properties: close-to-wall, parallel-to-wall and close-to-ground are computed based on the corresponding distances (using exp), 0.1 is determined empirically.

语意上下文关联：物体和场景的关联以及物体与物体的关联，例如床更可能在卧室里而不是在厨房里，桌子的旁边通常有椅子。几何上下文关联：考虑两个立方体之间的空间关系，主要定义两种，两物体相互邻近（对称关系）和某物体在某物体之上（非对称关系）。用训练集中两者之间的关系的频率作为潜在函数的值。

实验结果评估

采用NYUv2数据集，该数据集包含1449个场景，每个场景都包含一个RGB图像和一个depth地图，共有894类物体。本文将一些相近的物体类合并（如table和desk）并且忽略了一些很少出现的类，最终整合了21个物品类。

将数据集分为两个不相交的集合，一个是训练集、一个是测试集，其中训练集包含795个场景（3630个物体），测试集包含645个场景（3050个物体）。通过立方体检测 和 场景&物体分类 两项指标进行评估。

对立方体检测性能的评估:

采用base recall基本召回率进行评估，基本召回率是指在固定数量的候选立方体（K_c）情况下，成功检测到的地面真实对象的比例。对于一个真实对象，如果存在一个候选立方体与其重叠度超过0.5 IOU（交并比），则认为该对象被成功检测到。

随着K_C值的增加，base recall显著增加，当K_c=50时，基于深度扩展+非极大值抑制的方法base recall达到75%。

□□□

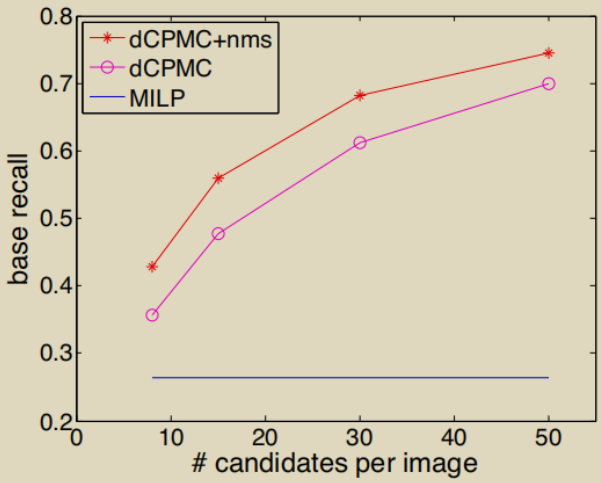

Figure 4. Comparison of different methods in cuboid detection. Performance measured in recall, *i.e.* the fraction of objects which are covered by at least one candidate cube (with overlap > 0.5.)

对分类结果的评估:

使用分类精度来衡量性能，即正确分类的物体个数比上总的物体个数。

不同条件随机场CRF配置产生不同的结果。

仅利用场景外观特征，分类精度为55.02%

仅利用分割特征和几何特征精度分别为54.46%和42.85%，两者相结合精度可达到59.02%有显著提高。

使用上下文潜在函数可以提高性能，例如，使用场景-对象潜在函数将目标标签的准确度从59.02%提高到60.00%，将场景分类准确度从55.20%提高到57.65%。

随着向框架添加更多潜在函数，准确度增加。当使用完整的CRF模型时，达到了最高的准确度，为60.49%，场景分类性能也有所提高（从55.20%提高到58.72%）。

□□□

configuration	object	scene
scene appearance only	-	55.20
segmentation only	54.46	-
geometry only	42.85	-
seg. + geo.	59.02	-
app. + scene-obj	55.87	57.49
app. + obj-obj	54.49	55.20
app. + obj-spa	55.61	55.20
unaries + scene-obj	60.00	57.65
unaries + obj-obj	58.92	55.20
unaries + obj-spa	59.41	55.20
unaries + scene-obj + obj-obj	60.13	58.56
unaries + scene-obj + obj-spa	60.33	58.10
unaries + obj-obj + obj-spa	59.28	55.20
all combined	**60.49**	**58.72**

Table 1. Performance of scene & object classification on ground-truth cuboids in terms of the percentage of correct labels. Here, "scene-obj", "obj-obj", and "obj-spa" respectively refers to scene-object co-occurances, object-object co-occurances, and spatial relations between objects. In addition, "app." refers to appearance-based potentials, including segmentation and scene appearance, and "unaries" refers to all unary potentials.

总体效率：利用CRF模型，几何特征、场景-对象关系和对象-对象关系的使用进一步提高了整体性能，相较于只关注物体本身的特征。使用F1分数（综合考量召回率和精度的一个指标）进行评估，出现频率越高的物体的准确度越大。

□□□

	dpm			seg.			seg.+geo.			seg.+geo.+rank			+scene-object			all		
	recall	prec.	F1	recall	prec.	F1	recall	prec.	F1	recall	prec.	F1	recall	prec.	F1	recall	prec.	F1
MILP	-	-	-	31.41	6.75	11.11	14.25	40.90	21.13	15.09	39.90	21.90	15.21	41.00	22.19	15.18	41.92	**22.3**
K = 8	38.22	4.47	8.01	25.76	33.13	28.98	25.53	37.02	30.22	32.14	38.82	35.17	32.79	37.96	35.18	31.67	39.68	**35.23**
K = 15	40.47	3.56	6.54	29.11	27.59	28.33	30.69	28.30	29.44	35.57	34.29	34.92	36.03	33.93	34.95	34.19	37.04	**35.56**
K = 30	44.52	2.62	4.96	32.02	20.25	24.81	33.48	20.70	25.58	38.29	28.30	32.54	39.56	27.67	32.57	30.19	36.68	**33.10**
K = 50	48.71	1.75	3.37	27.59	16.82	20.90	28.74	17.52	21.77	33.83	27.41	**30.28**	35.47	26.33	30.22	32.16	27.80	29.82

Table 2. Performances of the integrated framework. Each row corresponds to a specific detection setting, and each column corresponds to a model configuration. In particular, the first row shows the results obtained using the MILP detector [13], while the other four rows corresponds to the setting where a tunable detector is used to generate $K = 8, 15, 30, 50$ candidates per images. The first column shows the results by DPM, and the other five columns show the results obtained by our framework with different combinations of potentials. The performance are measured in terms of recall, precision, and F1-score. Note that these numbers are percentages.

	mantel	counter	toilet	sink	bathtub	bed	headboar	table	shelf	cabinet	sofa	chair	chest	refriger	oven	microwav	blinds	curtain	board	monitor	printer	overall
# samples	12	137	30	55	24	153	32	341	237	566	213	519	135	42	31	39	154	139	53	113	25	3050
seg.	0.0	6.7	1.7	2.7	0.2	10.6	0.9	14.6	12.3	27.8	11.8	27.9	4.1	0.4	0.6	0.9	8.5	7.4	1.6	4.5	0.5	28.22
seg.+geo.	0.0	7.4	1.5	2.9	0.2	11.6	1.0	14.7	13.0	30.7	12.5	29.1	4.0	0.6	0.5	0.7	8.9	8.0	1.8	4.8	0.5	29.25
seg.+geo.+rank.	0.0	11.2	2.6	3.4	0.3	16.6	1.3	19.1	18.2	44.3	18.4	39.4	7.7	1.2	0.6	2.1	14.0	11.7	2.4	8.8	0.0	34.78
+scene-obj.	0.0	9.7	2.4	3.8	0.3	15.6	1.3	19.7	17.6	41.9	17.9	37.2	8.1	1.5	0.9	1.4	13.5	11.0	2.6	8.2	0.7	34.94
all	0.0	10.9	2.6	4.0	0.3	16.7	1.3	20.8	18.7	44.5	18.6	40.3	8.0	1.4	0.9	1.8	13.9	11.5	2.8	9.0	1.0	35.04

Table 3. Class-specific performances obtained using CPMC-nms detector (with $K = 15$) + our CRF model (with four different configs). This is combination yields the best overall performance. Note that the numbers of testing samples in different classes are unbalanced. We show the reweighted F1 scores, which are defined as $F1 \cdot m/m_{max}$, thus emphasizing frequent classes.

总结：

NYU v2数据集上的实验证明，这种方法通过有效地结合分割特征、几何属性以及物体之间的上下文关系，优于现有的先进检测器。该工作中开发的框架非常灵活，本文作者认为可以扩展到包含来自其他信息源（例如视频）的信息，从而进一步提高性能。