

硕 士 研 究 生 读 书 报 告



题目 通过场景地图估计推断场景结构

作者姓名 项叶盛

作者学号 21851407

指导教师 李启雷

学科专业 软件工程

所在学院 软件学院

提交日期 二○一八年十二月

Scene Structure Inference through Scene Map Estimation

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Li Qilei

By

Xiang Yesheng

Zhejiang University, P.R. China

2018

摘要

从一张RGB图像中理解室内场景结构的技术有多种多样的应用，从编辑场景到挖掘空间利用数据都得益于该技术。在场景理解领域中大多数的努力集中在提取例如像素级深度或语义标签一样的稠密信息，或者是例如从目标识别中获取到的边界框。本文提出了“场景地图”概念，一种粗糙的场景表达，它描述物体在场景中的俯视二维位置（即，物体在地板上的位置），就像一个从一张RGB图像提取地图的管道。本文的最后，使用了一个合成管道，这个管道给合适的卷积神经网络提供虚拟的无限的训练数据。本文使用大量数据评估结果，结果清晰地显示出该方案胜于基于稠密信息的方案，并且论证场景地图为抽象的室内场景理解提供了一种有用的表示方式。

关键词：场景结构推断，场景地图

Abstract

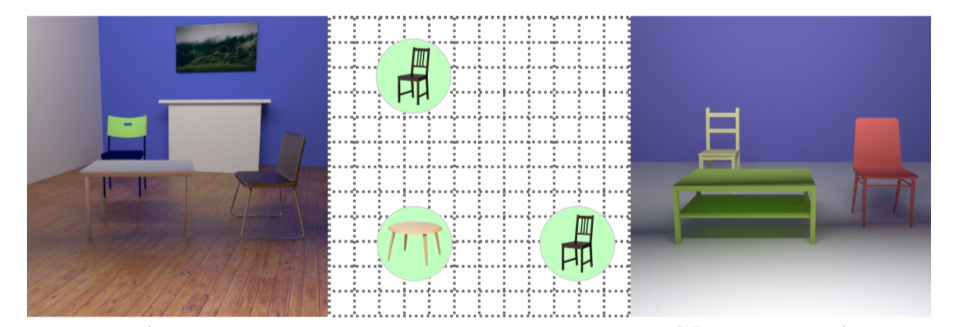
Understanding indoor scene structure from a single RGB image is useful for a wide variety of applications ranging from editing of scenes to the mining of statistics about space utilization. Most efforts in scene understanding focus on extraction of either dense information such as pixel-level depth or semantic labels, or very sparse imformation such as bounding boxes obtained through object detection. In this paper we propose the concept of a scene map, a coarse representation, which describes the locations of the objects present in scene from a top-down view (i.e., as they are positioned on the floor), as well as a pipeline to extract such a map from a single RGB image. To this end, we use a synthetic rendering pipeline, which supplies an adapted CNN with virtually unlimited training data. We quantitatively evaluate our results, showing that we clearly outperform a dense baseline approach, and argue that scene maps provide a useful representation for abstract indoor scene understanding.

**Keywords:** Scene Structure Inference, Scene Map

1 引言

我们对视觉输入进行推理的能力是至关重要的，这对我们不断地分析周边环境，以及根据这种分析形成正确反应都是至关重要的。如果我们要创造出一个像人类一样能够在错综复杂的现实世界中穿梭的智能系统的话，我们就需要找到一种能够重建这令人赞叹的逻辑推理能力。因此，毫不奇怪，室内场景理解在计算机图形和视觉领域获得了重大的研究关注。

室内场景理解的核心任务是场景结构推断。即推断出个体在场景中的存在及位置。对于人类来说，假如提供一张图像，在大多数的情况下，我们能够说出图片场景中的物体类别和物体之间的相对位置关系。当然，通过这些信息可以推理出其他更多的信息，例如通过房间、尺寸和场景类型（包含床和椅子的房间更有可能是卧室，然而包含桌子和办公椅的房间更有可能是办公室）推理出一条通畅的路线。因此，这种平面布局层面的信息提供了一个简洁有效的关于场景的性质和结构的摘要（见图1）。



输入 场景地图输出 三维场景实物模型

图1：给定一张RGB图片（左），本文提出了一个管道，一平面布局的形式生成场景地图，其中包含了被发现对象的网格位置。生成的场景地图可以潜在地用于生成一个三维场景模型（右）。需要注意的是系统还不支持姿势估计；因此在实物模型中所有物体都是朝默认的前方向。

从一张图片推断出这样的一个场景平面布局的一个可行解决方案是融合语义分割和深度估计，以定义场景中所有物体的位置。这种方案有一些缺点。首先，描绘单个对象的边界不是一件容易的事情，现有的语义分割大多不具有实例感知能力。其次，不管是语义分割还是深度估计都是按每个像素建模的。根本上，我们只关注物体间的相对位置，涉及像素标记的中间步骤可能引入一些重要的误差，这些误差影响整体的场景理解。

为了绕开这个问题，本文提出了一种新颖的表现方式，叫做场景地图。场景地图使用一系列网格，这些网格标记不同类别在场景俯视图中的位置（见图2）。重要的是，由于我们针对的是场景全局结构，而不是提取世界坐标中的位置，所以低分辨率的网格就足以满足这项任务。这将训练和评估所需的变量数量限制到最低限度。



图2：根据输入的RGB图像描绘每类别物体在俯视视角下的偏移位置。白色的正方形表明在那个位置存在特定类别的实例。

在本文中，将会提出一种基于单个图像的场景地图评估方法。其核心是基于VGG架构的卷积神经网络。由于训练数据十分稀少，本文创建了一个合成场景的渲染管线，并快速渲染它们，从而为网络提供了一个虚拟的无限数量的训练数据。利用这些合成的训练，对网络进行端到端的训练。该渲染管线的表现是引人注目的，52%的模型位于其地面真实位置的一个网格单元内。

将这种方案与结合语义分割和单帧深度估计基本模型进行比较。最后结果显示出该基于场景地图的评估方案针对该项任务具有更高的准确度，同时只需要处理明显少的变量。最后讨论了该方案的局限性和未来的工作。

因此，本文的主要贡献包括：

* 引进场景地图，用于表示整体场景的理解
* 提出一个利用三维模型集合来合成场景的方法，然后使用产出的场景数据训练卷积神经网络
* 提出了一种利用学习到管线，在给定单帧作为输入的情况下，对场景进行推理的方法

2 新的场景结构推断方法

本文的解决方案是，通过深度神经网络来学习输入到场景地图之间的映射关系。首先讨论这种新的表示方式，然后详细介绍涉及到的核心的网络结构。最后，我们解释用于合成场景的渲染管道，它为网络提供了无限多的训练数据，这有助于弥补缺乏真实训练数据的不足。

2.1 场景地图

将一张RGB图像作为输入，产出的场景俯视图被称为场景地图。直观地说，场景地图以类似地板平面图的方式对场景中存在的物体及其相对位置进行二维度上的概况。场景地图的两个坐标与给定室内场景的地板平面相应的x和y坐标平行。特定坐标系中的值对应于该位置上的物体。重要的是，场景地图完全去掉了第三个坐标(高度)。正如我们在下面演示的那样，这种简化表现方式，极大地促进了推理和学习任务，同时仍然提供了对整个场景结构的非常有用的概述。

更准确地说，假设场景包含N个不同类别的物体，场景地图S由个网格组成，其中，每个类对应一个网格。每个网格用二进制矩阵表示，表示任意类别的所有实例的位置（见图2）。这种表示方法的灵感来源于机器人三维映射应用中常用的占位网格表示，该方法将三维环境建模为一个均匀间隔的二元随机变量，当障碍物出现在相应位置时标记为1。因此，场景图可以看作是一个空间语义占位网格，空间坐标保留了二维，将类标识认作第三维度。

场景地图被设计成有限的分辨率。由于我们关注于场景的总体布局，并且接受在物体间留有一定的间距，例如30厘米。由于这一间距的存在，使得在放置问题中的变量数量（）相比于密集网格显著地减少，或使用的原始像素空间（）。这种简化方法在将回归转化为分类的计算机视觉应用中同样被使用到。例如,在论文[2]中，它的目标是预测物体自我运动的旋转和位移。该问题不是回归到精确的(连续的)角度和平移值，而是通过将每个运动绑定到一定数量的(离散的)运动幅度范围内来将问题转换为分类任务。当面临准确性不是必须的情况时，这种选择在精确性和复杂性之间产生了一种合理的权衡。

在本文的设置中,场景地图被设计成一个正方形区域，该区域是使用一个摄像机拍摄在正面拍摄大小的场景产生的视图。将场景中的地板编码成一个正方形区域，场景地图被编码成一个正方形区域。这个大小足够容纳SUNRGB-D数据集中95%以上的场景，并足够适应英国房间的平均大小。一共使用了个网格，每个网格的实际边长为37.5厘米。

2.2 场景地图推断的综述

本文的主要目标是从一张RGB图像输入中计算出场景地图表现形式。为了达到这个目的，本文采用了数据驱动的方法，数据驱动的方法已经被证明是一种对各种图像处理问题的有效解决方案。换言之，本文训练了一个卷积神经网络，在给定一张图像的情况下，尝试直接输出其场景地图表达形式。然而，采取这种方法的一个挑战是它需要大量的训练数据才能成功，其部分原因是通常需要估计大量的变量。不幸的是，目前还没有足够大的数据集包含真实的地面场景地图标记。（例如，最近的SUNRGB-D数据集包含将近10000张图像）。为了克服这个问题，本文使用多种方法快速合成的训练所需的场景数据，包括：采用现有的三维模型集合，改变场景的中物体的布局和物体的外观，物体的外观来源于大型的纹理数据集。本文创建了一个创建合成管道，它使用使用了一种渲染方式，和随机放置物体的位置，和随机生成模型外观。这个管线有效地给学习框架提供无限多的适应卷积神经网络训练的数据。简而言之，通用的方法包括以下关键步骤：

* 适配一个好训练的能够从单张图像推断场景地图的卷积神经网络架构
* 构建一个随机场景合成的管线，这一管线基于一种基于场景组合模型与外观变化相结合的高效渲染方法。
* 使用合成出来的场景训练网络，从而生成大量由图像和其相关的场景地图组成的正确映射对。
* 使用训练出的网络在新的测试集上估计场景地图。

2.3 神经网络

本文使用一个在VGG11[3]中使用过的深层神经网络，并加了一些修改(见图3)。尤其是，在每个卷积层和全连接层后面添加了批规范化层，从而显著减少训练时间。原始的网络将输入图像映射个不同的特征中，然后使用一个分类器预测每张图片的标签。由于本文的问题要求空间表现，而不是一个类标签，所以移除分类器，添加一层用于将表现信息重塑成大小的场景地图表现信息。注意，大多数为空间任务设计的架构（例如，语义分析）使用对称的镜像编码解码网络， 从而在每一层上加强编码器和解码器所学习到的特征图之间的直接对应关系。但是在本论文中的架构中这种方式是不合理的，因为输入域（图片像素）在分辨率和视角上与输出域（网格单元）不同。本文并没有给出最合适的神经网络结构。整个架构有将近2千万个训练参数。

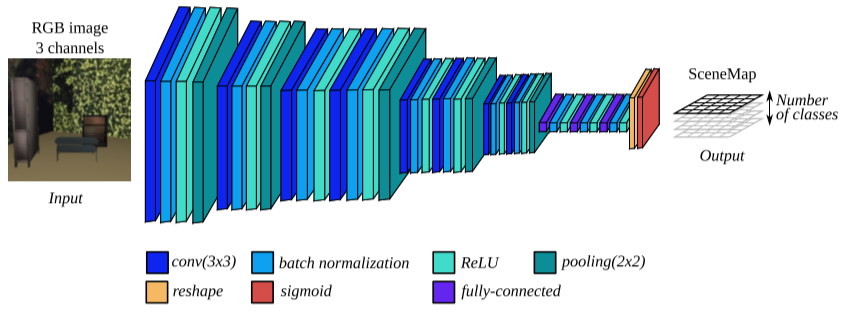


图3 本文使用到的基于VGG11的神经网络结构

2.4 渲染管线

训练一个深度神经网络要求大量的训练数据。SUNRGB-D数据集中有将近10000张图片，包括1000个不同类别物体的60000个边界区域。不过这些数据不足够训练出一个好的网络。为了增加训练数据的数量，设立了一个渲染管线，它动态地呈现训练图像对和相关的场景地图。这种方法为系统提供了无限的、本质上独一无二的训练数据流（虽然在理论上可能会出现两个一样的场景，但是出现的概率是极其低的）。这是一个在线学习的例子，它具有自我调节的能力，因此限制了过拟合的风险。

**数据。**渲染管线将一系列标有类别的模型O和纹理T作为输入。模型O取自IKEA数据集的子集。这个数据集包括四类物体：椅子、架子、桌子和沙发，每种类别中包含16个样式。这些模型需要手工整理，使得他们都具有精确的相对比例，并都以世界原点为中心，朝向同一方向。纹理集合T取自VisTex数据集136个纹理中的一部分，他们同样需要手工整理成合适的纹理。最后将模型O和纹理T按照75%/25%的比例分为训练集和测试集（见图4）。

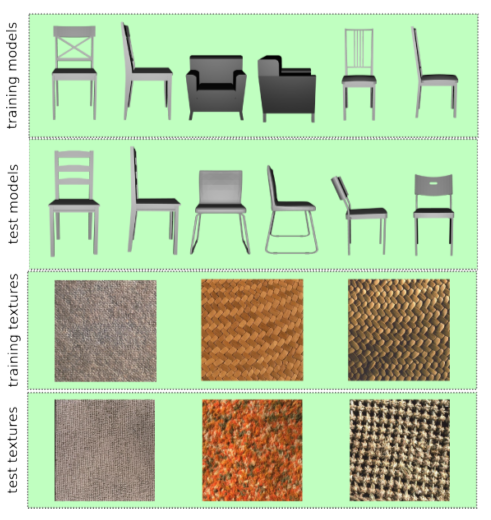


图4 按照75%/25%的比例将数据分为训练集和测试机。

**场景生成。**当管线被询问新的训练示例时，就会生成一个随机场景（见图5）。首先从模型集合O中随机取出2到6个模型，并随机放在一个6x6平方米的中心周围的某个位置上。之后再以上矢量为方向随机旋转90的倍数度。在对100张公开的室内照片的检查显示，一面墙几乎总是可见的，而另一面垂直的墙大约有75%的概率是可见的。因此，两个垂直的墙壁被随机地放置在场景的周围，但是具有连贯的方向。最后模型和墙面分别从纹理集合T中随机采样，然后进行缩放。如果包含了较小的纹理，就是用重复的模式模拟纹理。

相机被随机放置在高度在I米和1.8米之间的地方，以模拟手持相机拍摄出来的照片。

渲染。生成的场景是由装有简单手机渲染模型的OpenGL程序渲染而出。对于每一个场景图，论文都生成场景地图、语义标签和深度图（见图5）。

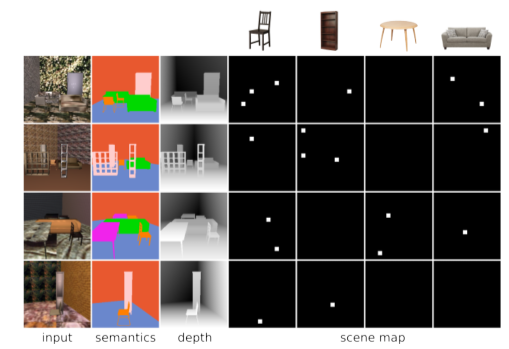


图5 数据生成管道中的4个示例说明。分别呈现了RGB输入、语义分割、深度和场景地图，从而产生一系列具有完整定义的训练数据流。

3 结果评估

本文使用场景地图作为场景理解的表现形式，其与密集的像素方法相比，其低维特性在处理的过程中删去的不必要的变量。本节内容将比较场景地图方法和结合了语义分析和深度估计的密集方法。他们的输入都是一张RGB图像。

3.1 语义分割和深度估计

**语义分割。**语义分割使用文章[4]的版本，使用VGG11[5]作为基础的编码器而不是VGG16，同样地除去了全连接层。最后输出层是6张映射，每一张映射代表一类别（例如：椅子、柜子、桌子和沙发），墙面和地板对应两张映射。使用了空间交叉熵损失，并对每个像素进行单独分类。训练数据与场景地图管线使用的一样（见2.4节）。

**深度估计。**深度估计与文章[6]一致，不过最后输出层是单个映射，而不是像之前一样是类别。在最后的输出层使用对数均分误差损失。

**联合输出。**将上述两个方法的结果合并到一张场景地图中。首先，从语义分割中提取链接的组件。然后，对每个组件，使用深度估计映射计算平均深度。最后使用已知的照相机焦距计算3维物体的中心坐标，之后映射到场景地图中。

**网络性能。**这些网络本身在各自的任务中就表现出非常高的性能。经过测试，语义分析网络达到96.5%的准确度，如此高的准确度（在原始文章[4]中，其准确度只有72.5%）的一部分原因是无限多的由场景合成管线随机产生的训练数据。甚至，原始文章[4]中使用更多的变量，同时也有更多的噪音，因为他们都是来自真实的照片，而不是渲染管线合成的场景，因此更具有挑战性。

3.2 比较

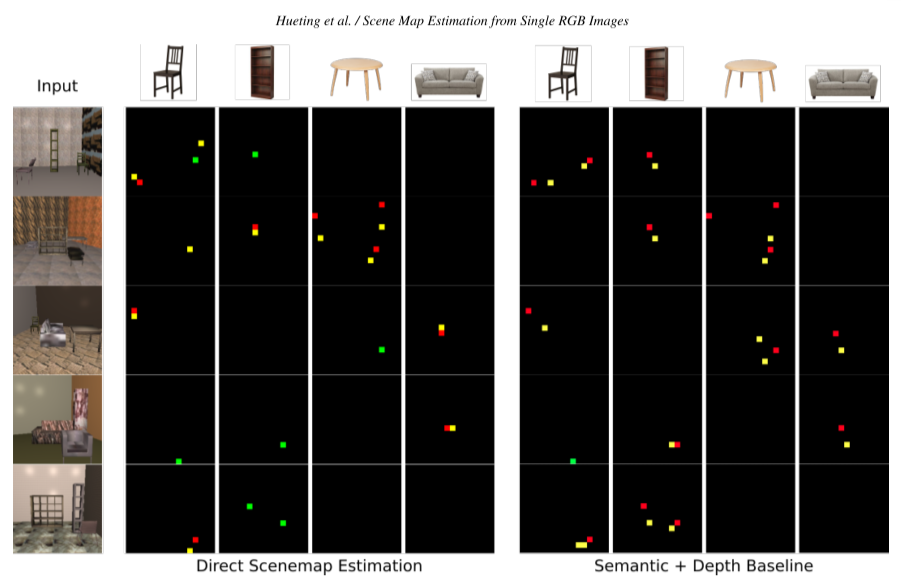


图6 左：由渲染管线生成的场景。每列代表一类别，每行代表一个样本。放置的正确性用颜色表示：绿色为正确的位置；黄色为错误的位置；红色为遗漏的。右：使用语义分割和深度估计的联合输出的场景地图。

在图6中，展示了三种方法的部分输出。可以看出本方法的清晰优势。可以发现，虽然本方法并非每次都能找到准确的位置，但是实际上出现的物体都被正确的检测到了。语义分割和深度估计的方法会遗漏一些对象，而且多次将一个对象检测成两个对象。

在表1中，对比了两者之间的识别质量。并且同时检测训练模型和测试模型的性能。除了正确率外，还测试了允许1格误差和2格误差的准确率（相应物体在正确位置的1格或2格内被检测到）。观察结果明显发现本方法各项都优于语义分割和深度检测联合方法。不过从训练集到测试集时，其准确度都下降了。

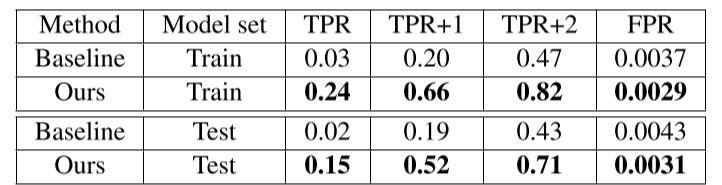


表1 语义分割和深度检测联合方法与本方法之间的比较。TPR是完全正确率，TPR+1是允许1格错误的正确率，TPR+2是允许2格错误的正确率。FPR是完全错误率，由于大多数网格单元是空的，所以错误率十分低。

3.3 密集对象的效果

为了测试本方法的较高密度的检测能力，本论文做了两组物体数量分别时6到9个和10到15个的实验。场景生成管线生成的物体会十分靠近，这使得分辨物体变得更加困难。样本采样的结果见图7。很明显，神经网络更难地检测出物体的正确位置。由于只有一个单元格能够被激活，所以它经常将两个物体融合在一起。表2展示的结果表明在物体数量增多时，检测性能会下降。

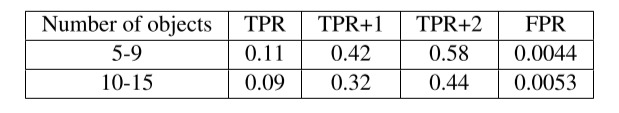


表2 本方法对较密集对象的检测结果。性能随着物体数量的增多而减低，但是没有缩小到不能使用的地步。

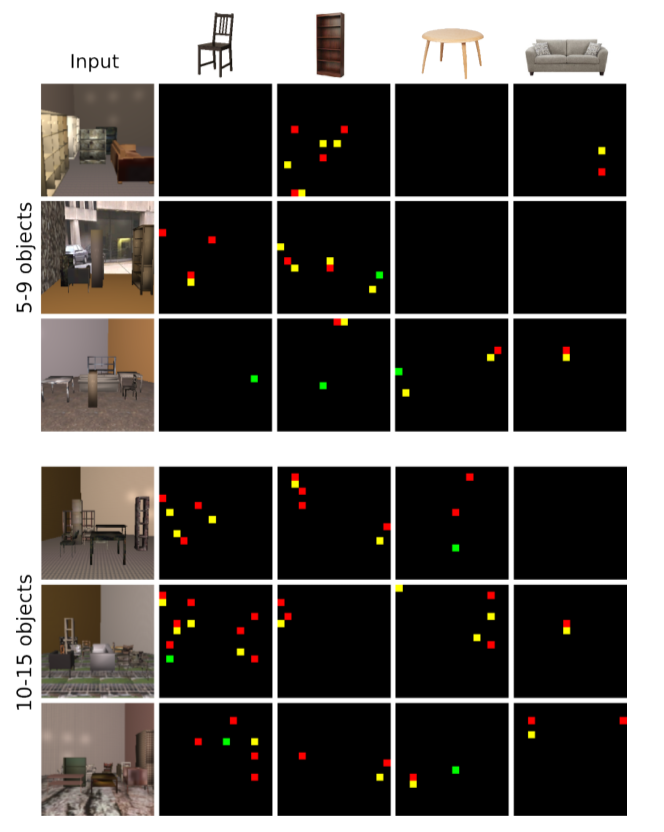


图7 增加场景物体密度之后的测试样本。放置的正确性用颜色表示：绿色为正确的位置；黄色为错误的位置；红色为遗漏的。右：使用语义分割和深度估计的联合输出的场景地图。物体的正确定位变得更加困难，但是大多物体的存在都能被检测到。

4 小结

本论文展现了一种新的场景结构表现方式，称之为场景地图。这将表现场景结构所必须的参数数量降到最低，因此减少了神经网络优化过程中必要的变量数量。虽然该方法的准确度收网格单元大小的限制，但是任然可以应用于多数任务，下文将罗列一些。该与像素化方法相反，像素化方法被设计成输出正确的预测位置，但是它的结果需要复杂的后处理工作才能在实践中使用，正如评估结果分析的那样。

不足之处。作者从单RGB图像中提取出场景地图的管线，还需要进一步的改进。目前而言，没有足够的真实数据去训练网络，为了扩展将该方法延伸到真实图像，更好的解决方案是将合成数据与真实图像结合使用。与该方法对比的其他方法还不够多，例如语义分割可以使用物体检测管线代替[7]。

参考文献

[1] Moos Hueting, Viorica Patraucean, Maks Ovsjanikov, et al: Scene Structure Inference through Scene Map Estimation. In Vison, Modeling, and Visualization(2016).

[2] AGRAWAL P., CARREIRA J., MALIK J.: Learning to see by moving. In ICCV(2015), pp. 37-45.

[3] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recongnition. ICLR(2014).

[4] NOH H., HONG S., HAN B.: Learning deconvolution network for semantic segmentation. In ICCV(2015), pp. 1520-1528.

[5] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for larg-scale image recongnition. ICLR(2014).

[6] EIGEN D., FERGUS R.:Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In ICCV(2015),pp. 2650-2658.

[7] REN S., HE K., GIRSHICK R. SUN J.: Faster r-cnn:Towards real-time object detection with region proposal networks. In NIPS(2015), pp. 91-99.