指导教师：杨涛　　提交时间：2016-03-19

CVPR2015 Paper Translation

No: 01

姓名：巴鑫

学号：2013302596

班号：10011306

估算曲面法线的深层网络设计

图一：给出一张图，我们的算法可以估计出每个像素的曲面法线。可以看到，我们的算法不只是估计出粗糙的结构，还捕获了良好的局部细节。比如说，在左图中，沙发扶手和旁边桌子腿的法线都精确地估算出来。（详见缩放视图）右图中，椅子的座面和腿，甚至旁边购物袋的顶部都被正确的捕获。坐标系：蓝色->X，绿色->Y，红色->Z

摘要

在过去的几年中，卷积神经网络（CNN）在可视化学习方面显示出了惊人的希望。本文中，我们采用 CNN 来完成从图片中预测曲面法线的任务。但什么样才是正确的架构呢？我们打算在以往几十年 3D 场景解析工作的基础上建立一个新的卷积神经网络结构来估算表面法线。我们发现一体化多个约束（人工制定，曼哈顿标准）和一些有结构性意义的中间代表（房间布局，边框）指引我们作出性能最先进的曲面法线估计。并且，我们的卷积神经网络具有良好的鲁棒性，显示结果在其他数据集也不需任何微调。

1. 简介

过去的两年中，计算机视觉方面有很多令人振奋的发展：深度卷积神经网络已经打破了场景分类检测和细粒度分类方面性能的障碍。例如，标准数据集目标检测的表现在两年中从 33.7 提升到 58.5. 卷积神经网络在语义任务如检测和分类方面取得巨大成功的同时，它在其他视觉任务如 3D 图像解读和建立匹配方面还没有广泛地研究。

我们从预测单一图片的曲面法线开始探索卷积神经网络在这方面的效能。可以直接利用卷积神经网络做逐像素拟合，就像之前做深度预测一样。然而，数十年的研究工作表明，这种方式的输出空间受到强大的物理约束的支配，研究人员计算机视觉一开始的线标签时代就开始利用这些约束，一直到近代。

在本文中,我们演示了如何将 3 D 表示和推理的剖析成果纳入曲面法线预测的深度学习框架。

而深层网络学习图像表示尤为成功使

我们相信它们的设计可以从以往的 3D 影像识别研究中获益。我们通过开发卷积神经网络在窗口上进行局部操作的同时在整张图片上进行全局操作来实现目标。这些预测不仅是曲面法线同时也是边界和立体空间布局。最终的卷积神经网络从消失的点中获得预估和证据进而生成最终预测。我们的方法在曲面法线估计方面获得最先进的性能，它还显示了一个在标准前馈结构上实质性的改进。此外,我们表明,我们的物理约束性能在最严格的评价指标下有 4.6%。更重要的是,我们的网络在不仅在表面法线还有空间布局和边缘标签的关系方面提供了更深度的理解。

2. 相关工作

3D 识别的话题回到了计算机视觉，从第一篇论文，罗伯特的 Blocks World 开始。这个难题的核心是两个相关联的问题：（1）什么才是正确的原始理解？（2）给出一个局部特征，如何获得全局的三维场景理解？

发现原语的问题可以追溯到计算机视觉的早期。吉伦最初提出的原语由线元和体元组成，但这种方法难以检测自然图像。最近的工作都集中在使用边缘,超级像素点或片段作为推理的原语。最近,研究者反而认为数据而不是人类的直觉应该确定原语，并介绍了一种块拼贴结构的原语。类似的逐像素阐述问题，片段只作为数据命令。不幸的是,在百叶窗,橱柜和瓷砖地板所有这些地方基本运行良好的同时，在没有特征结构的地方障碍重重。

为了解决歧义，大多数的工作转到某种形式的推理做自上而下的预测。最近的工作是基于高阶容积表示或着根据体积和边缘推理，通常,这表示通过优化在一个特定领域模型,帮助获得平滑预测和解决模糊的区域,比如空白的墙壁。

在这项工作中,我们解决两个线程。我们使用数据获得的像素表示而不是手动设计原语特性。

受到卷积神经网络在侦测、分割 、深度估计、姿态估计等等方面的成功的激励，我们打算采用卷积神经网络来获取 3D 图像识别的表示和原语。类似地，我们采用卷积神经网络仲裁证据之间的矛盾，而不是手动设计可优化模型来消除歧义。我们并没有放弃过去工作中获得的见解，相反它们被融入到我们的设计中。特别是,我们考虑到全局熔断和局部的重要性。我们建立局部和全局网络,处理这两种形式的迹象。我们把他们的预测通过网络融合，效果大大优于单独预测。我们的融合网络可以看作是以学习推理的形式将前文中的冲突替换为最优化的调和证据。过去的研究表明,室内场景的人为性质提供了强大的约束，例如,场景通常被认为有三个正交方向，同样,通常假设场景是一个由内而外的盒子。受这些方法的启发,我们的全局网络预测在提供粗糙的几何布局基础

上，还以消失的点为参考。这使融合网络应用曼哈顿或箱约束作为数据决定。我们的研究表明这些条件改善了预测结果。

局部结构。过去和最近都出现的另一个主题是用图像中曲面法线和边界来进行推理。受这些局部约束的启发，我们把他们合并在局部网络的学习模式中，并将其作为融合网络的输入。我们证明包含对凸、凹和遮挡边缘的预测提高了简单前馈网络的性能。

这项工作的初步版本发表在 Arxiv，与此同时，引入了堆叠卷积神经网络进行曲面法线估计。我们的贡献是互补的,两者结合可以提供进一步的改进。

3. 综述

本文旨在结合在过去十年中单幅图片 3 d 预测方面卷积神经网络代表学习的知识和力量。我们的总体目标是捕获帧幅图片 3 d 问题的已知框架，这样卷积网络结构可以做他们最擅长的事——从视觉数据标签学习强烈映射。

从我们过去的经验中，我们用以下的架构建立一个网络。我们从两个网络开始:一个全局网络,整个图像作为输入,并作出粗糙的全局解释；一个局部网络，以滑动窗口的方式作用于局部斑块并按局部趋向规划他们。因为全局和局部流程有互补的错误,我们将它们的输出通过融合网络结合,巩固他们的预测。每个输入网络自己可以获得有力表现,但通过结合他们,我们在定量和定性都能得到更好的结果。

为了进行全局和局部的网络融合，我们将人工的全局约束(包括房间布局、消失点)和局部曲面/边缘约束作为额外任务引入到框架中。我们的全局网络预测房间布局,局部网络预测的边缘标签。整合这些额外的任务使最终网络更加健壮。我们在第五节对我们的方法进行评估并且分析了我们哪些方面的设计增进了哪些性能。

4. 函数

现在我们来描述每个组件的函数。对于每一个组件,我们描述他们的输入,输出,中间层次,他们最小化的损失函数。

4.1 输出：分类回归分析

局部和全局网络的输出为：每一像素，空间布局，边缘标签的曲面法线。边缘标签(凸、凹、咬合,圆滑)是一个离散输出空间,可以制定一个分类问题。然而,曲面法线和房间布局都是有序的连续输出空间(标注曲面法线是在单位球面上进行的)。在过去 的工作中，我们把这些问题归纳为分类问题。

曲面法线：我们采用曲面法线三角正交编码技术将法线退化为一个分类问题。具体地说,我们首先开发包含 k - means 德劳内三角覆盖构造单词的电文。基于这种电文和三角测量,法线可以重写为一个码字的加权组合。在训

练时间,我们学习 softmax 码字分类器，在测试时间,我们预测码字分布；这变成了一个通过三角测量正常寻找三角形最大概率的法线，并且使用在三角形的相对概率作为重建法线的权重。

空间布局：空间布局是持续化有组织的输出空间。我们通过以箱式布局开发密码本来将问题重定义为分类问题。码字的开发机遇吧聚类 6000 种空间布局的 k-medoids，每一个码字都是一种类别。

4.2 全局网络

这个网络的目标是捕捉粗糙结构，解释出不能仅靠局部证据解码的模糊图像部分。输入：整个图片转制为 55 ↠ 55 ↠ 3。输出：以整张图片为输入，我们产生两种互补的全局解释：(i) 一个对整张图片曲面法线的结构预测。(ii)图像的立方近似。对于曲面法线预测,输出层次为 M t↠ Mt ↠ Kt ，Mt↠ Mt 是输出图像曲面法线的尺寸，Kt 是密码本中空间结构的使用级别。我们采用了超过 300 例的简单分类。我们规定 Mt = 20, Kt = 20。

架构：球形全局网络包含四个卷积层，这些层由两个任务共享(表面法线和空间间布局估算)。

第四卷积层的输出神经元完全连接到这两个标签。为了简化描述,我们用 C(k,s)表示卷积层，

这表明有 K 个内核，每个的型号为 s ↠ s。在卷积中,我们设置所有的步幅为 1。我们也表示局部标准化反应层为 LRN，最大级联层为 MP。级联的步幅是 2 且级联操作符的大小设置为 3 ↠ 3。那么卷积层的网络体系结构就可以表示为：C(64, 5) ！M P ！LRN ！C(192, 3) !M P ！LRN ！C(384, 3) ！C(256, 3)。对于曲面法线估计，第四卷积层的神经元完全连接到输出空间 Mt ↠ Mt ↠ Kt，空间尺寸为 20 ↠ 20 ↠ 20 = 8000。对于空间结构估算，我们将同样设定的神经元连接到 Kl = 300。损失函数：我们把这些任务都当作分类问题。对于空间布局分类，我们简单地应用最软回归来定义损失。

对于曲面法线估算,我们用 Fi(I)表示曲面法线输出示意图上第 i 像素的 Kt 类输出。我们还应用 softmax 回归优化 Fi(I)函数。然后曲面法线结构输出的损失就可以表示为：

$$L(I, Y) = - \sum_{i=1}^{M \twoheadrightarrow M} \sum_{k=1}^{K^X} ((y_i = k) \log F_{i,k}(I)), \quad (1)$$

Fi,k(I)代表这第 i 像素的法线应该被第 k 码字定义的可能性。𝟙(yi = k)是指示函数，Y = {yi}

是曲面法线的地面真值标签，M = Mt，K= Kt。

训练学习过程中，我们发现网络同时存在两种损失。由于我们有结构化的曲面法线输出却只有一种空间布局预测，我们需要平衡这两种损失的训练

比率。如果σ表示曲面法线估算的训练几率，那么我们定义空间架构估算的训练比率为50σ。

4.3 局部网络

这个网络的目的是捕捉那些在更高分辨率时可能被全局网络遗漏的局部数据。我们采用滑动窗口的方法，在一个窗口提取特征，在窗口中心预测图像属性。这种类型的模型已成功应用于生成局部图像的曲面法线和语义边缘的解释。输入：给出一张尺寸为195 ↠ 260的图片，我们在其上运行滑动窗口模型，窗口尺寸为55 ↠ 55，步幅为13.输出：局部网络提供两种类型的输出：(i)曲面法线(ii)边缘标签。每一个局部滑动窗口提供中心窗口的Mb ↠ Mb像素的曲面法线预测。我们的网络以较小的一部分图像作为输入，并预测中等大小的曲面法线补丁。因此从局部纹理和它的环境预测曲面法线。我们使用Kb = 40码字来定义输出空间。我们期望局部网络能捕获更好的细节，所以采用的更大范围的码字。对于边缘，我们使用凸、凹、闭塞或非边缘的经典分类方式。请注意我们预测一个标签为13↠13像素的边缘。出于可视化的目的,我们在输出的结构化边缘的基础上设计这些边缘标签。

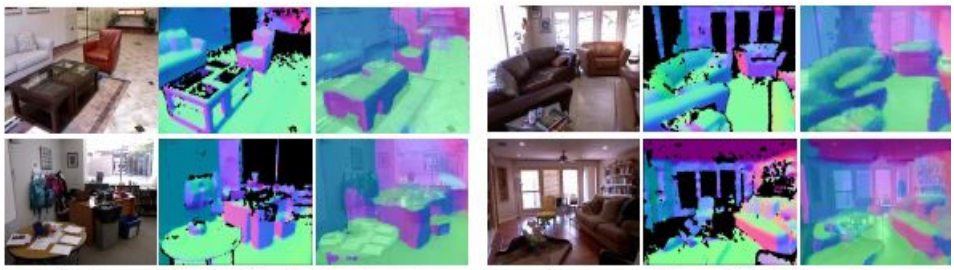
结构：本地网络的体系结构包括4层卷积层和2套完全连接层。卷积层由两个任务共享，我们使用全局网络提到的相同的设置参数。同时在第二层卷积网络中，我们通过4096个神经元把两层全联通网络堆叠起来，每一次对应一个任务。局部曲面法线估算的输出是Mb ↠ Mb ↠ Kb = 13 ↠ 13 ↠ 40 = 6760；边缘标签有四项输出。损失函数：这两个任务都被定义为分类。我们用最软递推来定义局部曲面法线的损失并设定M = Mb ，K = Kb。和粗糙网络相似，我们在训练中共同优化这两个任务。局部的曲面法线和边缘标签估计的学习速率分别是σ和50σ。

4.4 可视化

我们现在尝试分析全局和局部网络的学习。请注意,这两个网络共享相同的卷积层结构，可容纳单位的字段大小是31↠31。全局网络的单位捕捉高标准的建筑，比如床的一边，走廊和墙上的画。局部网络单位应对对局部纹理和边缘。

4.5 融合网络

这个网络的目标是融合并完善两个早期网络的结果。每个方法都有互补的失效模式，我们表明,通过融合两个网络工程,可以获得更好的结果。此外，局部网络和粗糙网络都独立的处理每一像素；我们的融合网络也可以将输出用一种学到的推理形式输出。作为融合网络的输入，我们把局部和全局网络从输入图像得到的输出连接起



来。

串联过程如下:

- 全局粗糙输出:全局网络的输出是 20 个等级的 20 ↠ 20。我们将输出解码为三维的连续曲面法线图,并升级为 195 ↠ 260 ↠ 3。
- 布局:我们选择对应概率最高的空间布局标签。布局是一个代表结构表曲面法线的三路特性图。我们把它的大小调整为 195 ↠ 260 ↠ 3。
- 局部曲面法线:局部网络滑动窗口的输出形式是 195 ↠ 260 ↠ 3。
- 边缘标签:我们每个窗口获得 4 种概率的边缘标签。由于概率和为 1,我们不把无边缘的输出传递给融合网络。我们把每个窗口的三维向量补强到 13↠ 13↠ 3,从而获得 195 ↠ 260 ↠ 3 的输入。
- 消失对齐点粗略输出:我们调整我们的全局输出的解释来匹配消失点,产生的另一个特征表述具有相同的大小。

除了上面描述的 15 个渠道,我们还链接原始图像,因此,最后深层网络的输入为 195 ↠ 260 ↠ 18。

输出:融合网络也在 195 ↠ 260 图像上应用了滑动窗口机制。通过捕捉 55 ↠ 55 的输入,我们通过融合网络估计 Mb↠ Mb 中心补丁的曲面法线。注意,我们在局部网络使用相同的输出窗口大小 Mb = 13,并且输出空间通过 K_b = 40 码字来定义。

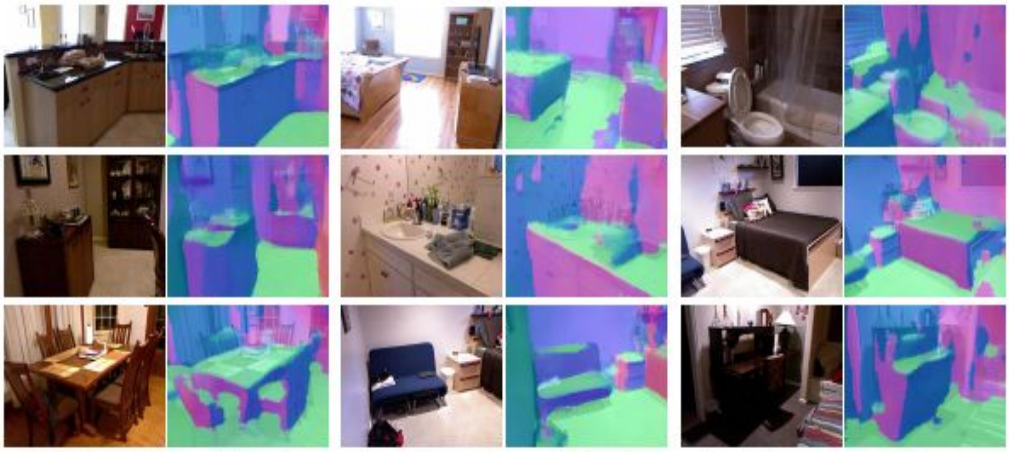
架构:这个网络的架构是四层卷积层和两层完全连通层。全局和局部网络卷积层共享相同的参数设置。第二层卷积网络中,我们通过 4096 个神经元把两层全联通网络堆叠起来。在测试时间,我们设步幅=Mb 在特征图谱上应用融合网络。

损失函数:在训练时,我们修正全局和局部网络的参数并获取训练数据的特征图谱。为了训练网络,我们应用随机梯度来下降学习速率。

五. 实验

我们现在来介绍我们的实验。

数据集和设置:我们用纽约大学 v2 深度数据集评估我们的方法。我们使用对应的原始视频数据作为训练图像。我们使用提供的开发工具处理视频数据,但通过 TV 去燥改善法线。我们使用官方的分裂:249 场景培训和 215 场景测试。我们从 249 中提取 200 k 帧场景进行训练和用标准测试集的 654 张照片进行测试。我们通过拟合由内向外的框体来提取空间布局并估算曲



面法线。边缘标签估计采用真实的深度数据。

训练过程中，我们使用学习几率 σ =1 .0×10^-6 衰减随机梯度来很好的调整网络。注意,在联合优化布局和边缘时，我们设定这些损失的学习几率为 50σ 。为了训练粗糙网络，我们通过翻转、颜色变化和随机裁剪增大了数据。为了训练局部和融合网络，我们将训练图片调节为 195×260，并随机 300K 帧截取了尺寸为 55×55 的样本。

评估标准：我们在整个数据集评价每一像素的错误，忽视那些由于丢失的深度数据而无法获得的线索。我们通过统计均值和中位数来总结全体逐像素错误，就像 PGP 指标。出于简单比较,我们提供实际结果[22],相对性能类似的实例[9]和我们的去噪法线。

基准：我们的主要基准是最新发布的曲面法线预估标准。我们评估都至少有一个最先进的指标。由于没有卷积神经网络已发表的结果，我们采用 Eigen et al 的粗糙网络特征。曲面法线通过使用消极的点积计算损失。这种粗网络在深度性能上几乎全系统匹配，并且，作为一个没有中介表示或结构设计的前馈卷积神经网络,这是一个很好的基准。

5.1 实验结果

定性：首先,我们展示我们的定性结果。图 2 和图 3 展示了我们的完整结果。注意我们的结果完美捕获了输入图像的细节。与许多过去的方法不同,我们的算法可以估计桌腿等，它甚至可以估计整个沙发曲面法线的变化。

定量：表一比较了我们的算法与几个标准，结果表明,我们的方法明显好于所有的基线的所有指标。在许多情况下,我们的结果比先前的结果改善高达 15%。

为了完整性,我们也报告了当代 Arxiv 论文中使用了堆叠卷积神经网络的结果。虽然我们未使用任何图像网数据做预训练，性能指标仍具有可比性。在模糊像素方面,他们倾向于产生平均的方法和平滑输出而我们的方法选择其中一个解释。因此,他们的平均误差较低而我们的中值误差较低。

消融分析：接下来,我们进行一个全面的消融分析探讨网络的哪个组件有助于提高性能。这种消融分析有助于评估基于有意义的中间表示形式和可以帮助提高性能的约束设计网络的假设。

首先,我们讨论一些表格 1 中展示的定性结果。

Table 1: Results on NYU v2 for per-pixel surface normal estimation, evaluated over valid pixels.

	(Lower Better)		(Higher Better)		
	Mean	Median	11.25°	22.5°	30°
Our Network	26.9	**14.8**	**42.0**	61.2	68.2
Stacked CNN [7]	**23.7**	15.5	39.2	**62.0**	**71.1**
UNFOLD [10]	35.2	17.9	40.5	54.1	58.9
Discr. [22]	33.5	23.1	27.7	49.0	58.7
3DP (MW) [9]	36.3	19.2	39.2	52.9	57.8
3DP [9]	35.3	31.2	16.4	36.6	48.2

正如图中所看到的,全局网络只能捕获房间的粗糙结构。例如,在前面的图,它忽略了沙发内一侧的垂直面或者错过垂直方向由于沙发之间的书架所发生的改变。另一方面，局部网络很好的捕捉到了这些细节。然而,由于它只观察局部的补丁,它完全分类错了相框下面的墙上的补丁。尽管融合两个网络保存的细节(沙发的内部一侧和墙垂直方向的变化)，一墙上相框附近的个大补丁依然分类错误。然而,一旦网络使用边缘标签(比如架子的凸边和墙上消失的边缘)就会改善边界。

定量的看,我们在表 2 中一个接一个的比较所有组件。

Table 2: Ablative Analysis

	Mean	Median	11.25°	22.5°	30°
Full	**26.9**	**14.8**	**42.0**	**61.2**	**68.2**
Full w/o Global	28.8	17.7	34.6	57.8	66.0
Fusion (+VP)	27.3	15.6	40.2	60.1	67.5
Fusion (+Edge)	27.8	16.4	37.5	59.4	67.4
Fusion (+Layout)	27.7	16.0	38.8	59.9	67.4
Fusion	27.9	16.6	37.4	59.2	67.1
Local	34.0	25.1	25.6	46.4	56.2
Global	30.9	20.8	31.4	52.3	60.5
Coarse CNN [8]	30.1	24.7	24.1	46.4	57.9

融合网络结合了原始图像和从局部和全局网络曲获得的面法线预测，在性能方面具有显著提高。此外,添加布局、边缘和消失点独立地提高了网络的性能。通过把他们结合在一个完全融合网络,我们在所有指标上都获得了更好的结果,在最严格的标准下也获得 4.6%的增长。

我们注意到,添加组件仅占基本系统中每个组件的最大收益。虽然很容易改进劣质系统,很难改进一个融合网络这么强大的系统:本身,它在大多数指标将是最先进的。我们因此报告只融合本地网络的约束,能使所有 PGP 指标上涨 9%。这突显出我们约束的有效性和价值。最后,我们注意到,我们的性能明显优于我们在图 3 的实现的粗网络,一个前馈 CNN。

5.2 伯克利 B3DO 数据集

为了证明我们的模型可以推广，我们直接在伯克利在 B3DO 数据集应用它。两数据集之间有明显不匹配的偏差:纽约大学包含几乎完全完整的场景而 B3DO 包含很多特写镜头。也,因为 B3DO

包含了许多场景视图，不像纽约大学那样，我们改正结果来发现消失点并进行补偿。我们在表 3 中报告完全融合网络的结果和如图 3 所示的一些定性结果。我们的方法在所有指标优于基线一大截。

6. 结论

我们提出了一种新颖的以卷积神经网络为基础的曲面法线估计方法。通过在 3 d 表示中注入深刻见解，我们的模型实现了最先进的性能。定性的来说，我们的模型是有效的，不仅抓住了粗场景结构，也捕捉到了细节，如桌腿和沙发的曲面。

引用

[1] I. Biederman. Recognition-by-components: A theory of human image understanding. Psychological Review, 94:115 – 147, 1987. 2

[2] M. Clowes. On seeing things. Artificial Intelligence, 2:79 – 116, 1971. 1, 2

[3] J. Coughlan and A. Yuille. The Manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In NIPS, 2000. 2

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005. 2

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR, 2009. 7

[6] P. Doll´ar and C. L. Zitnick. Structured forests for fast edge detection. In ICCV, 2013. 4

[7] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multiscale convolutional architecture. CoRR, abs/1411.4734, 2014. 3, 6, 7

[8] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. CoRR, abs/1406.2283, 2014. 1, 2, 6, 8

[9] D. F. Fouhey, A. Gupta, and M. Hebert. Data-driven 3D primitives for single image understanding. In ICCV, 2013.2, 4, 6, 7, 8

[10] D. F. Fouhey, A. Gupta, and M. Hebert. Unfolding an indoor origami world. In ECCV, 2014. 1, 2, 6, 7

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014. 2

[12] A. Gupta, A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In ECCV, 2010. 1, 2

[13] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from RGB-D images. In CVPR, 2013. 6

[14] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In ICCV, 2009. 1, 2, 4, 5, 6, 8

[15] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In ICCV, 2005. 2

[16] D. Hoiem, A. Stein, A. Efros, and M. Hebert. Recovering occlusion boundaries from a single image. In ICCV, 2007. 2

[17] D. Huffman. Impossible objects as nonsense sentences. Machine Intelligence, 8:475 – 492, 1971. 1, 2

[18] A. Janoch, S. Karayev, Y. Jia, J. Barron, M. Fritz, K. Saenko,and T. Darrell. A category-level 3-d object dataset: Putting the kinect to work. In Workshop on Consumer Depth Cameras in Computer Vision (with ICCV), 2011. 8

[19] T. Kanade. A theory of origami world. Artificial Intelligence,13(3), 1980. 1, 2

[20] K. Karsch, Z. Liao, J. Rock, J. T. Barron, and D. Hoiem.Boundary cues for 3D object shape recovery. In CVPR, 2013.2

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012. 2

[22] L. Ladick´ y, B. Zeisl, and M. Pollefeys. Discriminatively trained dense surface normal estimation. In ECCV, 2014.2, 3, 4, 6, 7

[23] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. E. Howard,W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In NIPS, 1990. 2

[24] D. C. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In NIPS, 2010. 1, 2, 4

[25] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In CVPR, 2009. 2, 8

[26] D. Munoz, J. A. Bagnell, and M. Hebert. Stacked hierarchical labeling. In ECCV, 2010. 5

[27] L. Roberts. Machine perception of 3D solids. In PhD Thesis,1965. 1, 2

[28] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In NIPS, 2005. 2

[29] A. G. Schwing, S. Fidler, M. Pollefeys, and R. Urtasun. Box In the Box: Joint 3D Layout and Object Reasoning from Single Images. In ICCV, 2013. 2

[30] A. G. Schwing and R. Urtasun. Efficient Exact Inference for 3D Indoor Scene Understanding. In ECCV, 2012. 1, 2, 4

[31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In ICLR, 2014.2

[32] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In ECCV, 2012. 6

[33] K. Sugihara. Machine Interpretation of Line Drawings. MIT Press, 1986. 2

[34] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In CVPR, 2014. 2

[35] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. TPAMI, 32(10):1744 – 1757, 2010. 5

[36] D. Waltz. Understanding line drawings of scenes with shadows. In The Psychology of Computer Vision. McGraw-Hill,1975. 2

[37] X. Wang, D. F. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation. CoRR, abs/1411.4958,2014. 3

[38] X. Wang, L. Zhang, L. Lin, Z. Liang, and W. Zuo. Joint task learning via deep neural networks with application to generic object extraction. In NIPS, 2014. 2

[39] Y. Zhang, S. Song, P. Tan, and J. Xiao. Panocontext: A whole-room 3d context model for panoramic scene under-standing. In ECCV, 2014. 2

[40] Y. Zhao and S. Zhu. Scene parsing by integrating function,geometry and appearance models. In CVPR, 2013. 1