指导教师：杨涛　　提交时间：2016/3/18

CVPR2015 Paper Translation

No: 01
姓名：崔宏伟
学号：2013302544
班号：10011304

利用室内场景结构分析的单张图像的深度估计

卓伟，Mathieu Salzmann，何旭明，刘苗苗
澳大利亚国立大学，澳大利亚 NICTA 研究组织

摘要

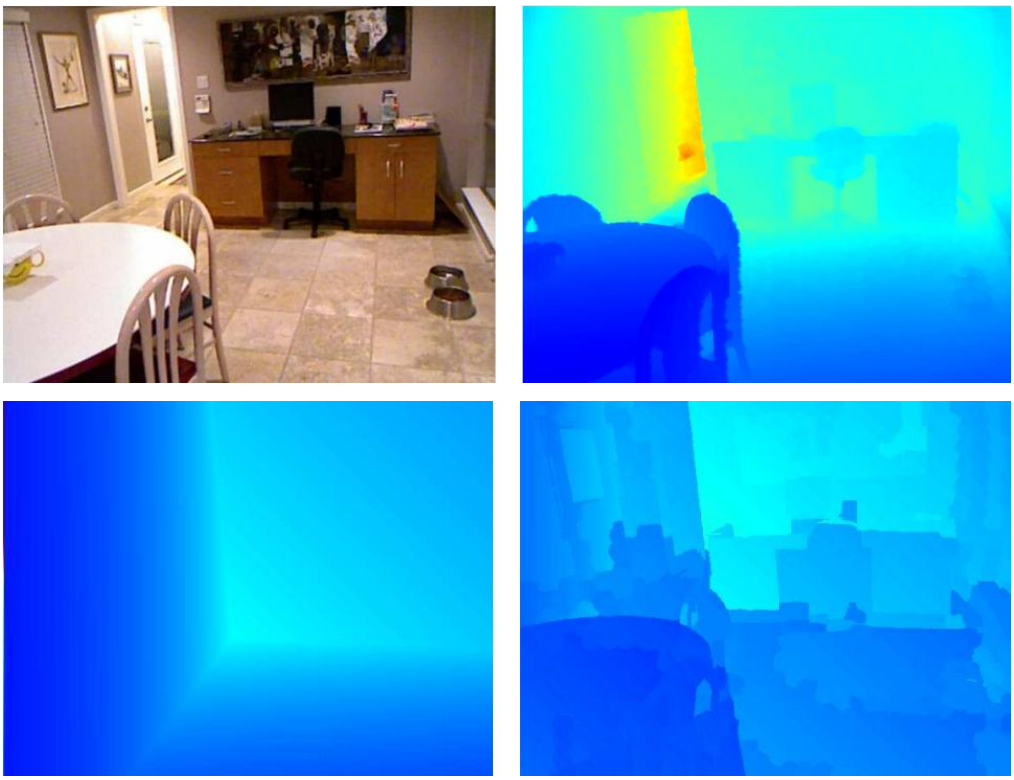
我们在解决单张图片的深度估计问题时由于没有额外的辅助的信息，遇到了很多模棱两可的问题。不像先前局部推理的方法，我们建议利用场景的全局结构来估计它的深度信息。为此，我们引入了一种通过结合局部和全局的场景结构信息来估计局部深度的场景分层表示的方法。我们利用单个图片的深度估计作为一个图形模型的推论公式，同时该模型图像中的边缘信息能够让我们在层次等级制度中来编码各个层次中和不同层次之间的相互作用。因此，我们的方法不仅能够生成详细的深度估计信息，并且还尽可能的得到 更高层次的场景所反映的信息。我们证明了我们的方法相较于局部深度估计的方法在标准室内数据集中的能够发挥出的优势。

1. 介绍

在没有任何先验信息的帮助下，利用单张图片来估计一个场景的深度信息是一个非常模糊困难的问题。然而，由于成年累月积累的数据和知识，人类能够很容易从单只眼睛获取到的图像输入感知到深度信息。很直观地，这个现象建议我们在利用已有的图像深度测试对进行单张图片的深度估计时，应该确定一个现实的能够达到的目标。

这个观察结果已经成为推动一些最近研究单目深度估计方法的动力，例如：单张图像的三维重建，单张图像的三维结构获取，场景的几何模型中的层次，非参数采样的视频深度提取等。但这些方法，通常只在局部范围模拟深度。例如，基于多尺度深网的单图像深度图预测就是单独地预测每个像素的深度信息。相较而下，虽然其他几种方法利用对邻近的超像素之间的关联性建模来编码一些更高层次的信息，但是能够由此产生的方法仍然没有获得场景的全局结构。这违背了我们认为人类利用这样的更高级的场景结构信息来分析所在的环境的直觉。

不过，场景的结构恢复在以前的从图像恢复表面布局方法，利用物体和曲面的体积推理估计房间的空间布



图像 1. 单张图像的深度估计：（顶部）原图和真实场景深度地图。（底部）估计的场景层次和详细的深度地图。颜色代表着深度（红色是远，蓝色是近）

局方法，使用定性几何分析和构成法理解图像的方法，使用外观模型和基于房间的上下文几何分析方法中就已经开始被研究了。而且所得的方法通常是在一个粗尺度表示所关注场景。正因如此，他们没能提供出场景的细节描述。更重要地，尽管这些方法确实推断了场景的结构，但是他们仍然没有得到绝对的深度估计；通常地，只有一些平均值利用这些提供了至少全局模糊的深度范围的技术被预测出来。

在这篇论文中，我们计划利用高层次的场景结构信息来进行细致的单张图像的深度估计。为此，我们介绍了一种依赖于编码局部、较大范围和全局的信息的分层表示景深的方法。当继续从场景的全局结构信息获取好处时，这种方法让我们能够模拟详细的景深。更具体点，我们的分层景深表示方法由三个层次组成：超像素，局部和布局。超像素使得我们能够模拟局部的景深变化。

更具体点，我们的分层景深表示方法由三个层次组成：超像素，局部和布局。超像素使得我们能够模拟局部的景深变化。相比之下，局部的区域能够让我们解释出中型和大型的场景结构。我们为我们的分层结构的每一层，建立了带有多变量的条件随机马尔可夫模型来解决深度估计的问题。CRF 可以让我们对同层和不同层的相互作用进行编码，从而能够高效的同时得到局部和全局信息。正如图

示 1 所示，利用我们模型的推断，因此产生了深度估计的细节程度由差到好的转变。

通过两个标准的室内图像数据，我们展示了我们方法的高效性。我们的实验证明了利用更高层次的场景结构信息的方法相对于局部深度估计方法的好处和优点。

2. 相关工作

相比于传统的多视角的三维场景重建方法，单张图像的深度估计问题近些年才得到流行。尽管如此，在很短的几年里，这项富有挑战性的工作已经取得了很大的进步。

由于这个问题固有的不确定的性质，现有的方法主要依赖于训练数据（图像深度数据对）。在这些方案中，一种普遍的方法是通过学习，进行数据的的回归分析来预测局部的深度信息。根据预测的语义标签的单张图像的深度估计就是利用了这种方法，在这种情景下，从图像特征到像素深度的一个特定的回归分析被用来训练成为数据中的每一个语义类。利用相似的想法，图像分割方法（18）就是通过训练分级器得到具体的语义标签在某些特定的深度规范情况中。这些分级器就会被用来预测像素的深度信息。

一些方法已经好像超出了纯粹的深度估计问题。例如，可扩展的基于样本的深度转换方法（4）中介绍了一种基于稀疏编码的直接整体场景的深度预测方法。相似的，基于多尺度深网的单图像深度图预测方法（5）是通过训练一个深度神经网络来预测整张图像的像素深度。可是，为了获得较好的精度，这样的全局场景预测方法就需要大量的先验数据的支撑。相比之下，许多技术更喜欢利用相邻像素之间的关系来建立模型以获得图像中的连通性。除了非参数采样的视频深度提取，从互联网中使用 3D 实例自动 2D-3D 图像转换和从 2D 到 3D 图像转换的例子学习深度的方法（15 16 17）中将深度恢复问题制定为连续型优化问题的方法外，通常，在模型中这样的连通性都会被编码。这种带有比较简单的超像素之间关系的方法在单张图像的三维重建和单张图像的三维结构获取研究（25 26）中被引入。一种简单的光滑性规则也有人在根据预测的语义标签的单张图像的深度估计（20）中和局部集合推理以及先前提到的回归方法作为数据项一同使用。在从单个图像离散连续深度估计（21）方法中，增加了一些离散变量用来建立出更加复杂的超像素关系模型，从而产生出高阶离散连续图像模型。不论相邻像素之间的相互作用的原因，所有上述提到的模型都无法考虑到深度估计中最重要的因素，也就是场景的全局性结构。

近些年，估计场景的结构信息已经成为一个活跃的研究领域。例如，从图像结构估计深度以及模式分析与

机器智能的（31）方法中使用一种不精确的方式建立的结构模型，以此作为场景的绝对平均深度。为了建立出更精确的结构模型，在从图像恢复表面布局的研究（13）中大量的工作从几何方面的想法中被引入。从而，依靠曼哈顿猜想，这个想法被延伸用来预测室内场景的布局情况通过一个盒子模型。场景的几何模型中的层次研究方法（22）通过将场景分为 15 类光学类型来代表它的结构，从而代替了盒模型。在利用几何推理的单个图像的结构恢复研究方法（19）中引入了更加精确的表示方法，同时此方法只产生了稀疏曲面法线。相似地，在用于单图像理解的数据驱动的三维图元研究方法（6）中，局部的深度平均值可以被预测出来，但是其中的通过利用局部的利用向量机的方式被认为是有歧义的。近来，展开室内折叠的世界研究方法（7）改善了这种平均值估计问题，它利用的是 CRF 算法和正常的不连续推断。其中一个这些场景结构分析方法的主要问题就在于它们无法真实的估计出深度信息，只能估计出均值信息，于是遗留了至少一个全局性尺度模糊的问题，而且往往更因为不同均值的表面区域的相对顺序可能无法通过结构重建而被确定。

这里，我们提出利用高层次的场景结构信息来进行精确的深度估计。因此，虽然得到了单个图像离散连续深度估计研究方法（21）的工作的灵感，我们的构想建立了更加完整的场景表示模型，而且它包含了局部，大范围的和全局的三个关键层次级别。正如我们结果证实的那样，利用这样的更高层次的推断，单张图像的深度估计能够获益很多。

3. 结构感知的深度估计

我们现在介绍我们的分层模型如何进行单张图像的深度估计。正如先前提到的，深度估计在 CRF 算法中是被表示为一个推理，同时它允许我们能够对同层和不同层的相互关系进行编码。为此，我们设变量 Y,R,L 来分别表示局部深度，中间层次和全局层次的结构。推论将通过极大化 CRF 的联合分布或者等效地，即每一个个体的能量项对应于一个特有的模型中的层次，然后最小化能量分布来达到。

$$E(Y,R,L) = E_l(Y) + E_m(Y,R) + E_g(Y,L)\,, \quad (1)$$

在余下的部分中，我们将详细的介绍这些不同的术语。

3.1. 局部深度估计

我们的模型依赖于图像的超像素来对深度进行详细的估计。每一个超像素都可以看作三维中的一个平面，从而将深度估计问题转化为寻找最适合的对应与每个超像素的平面的参数值。特别是，在这里，我们编码 每个平面中它的质心和法线方向的深度。更具体点，让 Y=$\{y_1, y_{2,\ldots,} y_{Ns}\}$成为代表 Ns 个图像中的超像素的离散变量的集

合，在集合中，每个 y_i 能够从一个离散状态空间 S 中取值。通过量化有效深度范围来作为超像素的图心的 V 的值，我们定义这个从训练集中获得的有着确定的最大最小范围的状态空间。将超像素的法线方向限定在 3 个可能的占有优势的方向上，而这些方向是利用一种新的方法，在建筑环境消失点检测研究中（24）的消末了的点估计方法来进行定义的。这让我们定义了第一个能量项 在式子 1，就如此：

$$E_l(Y)=\sum_p \Phi_p(y_p)+\sum_{p,q}\Phi_{p,q}(y_p,y_q), \qquad (2)$$

在式中，Φ_p 是一个一元函数用来编码把标记 y_p 分配给超像素 p 的花费，$\Phi_{p,q}(y_p,y_q)$ 是一个双变量函数用于增强超像素之间的一致性。

一元函数是基于在单个图像离散连续深度估计研究方法（21）中的回归项确定的。为此，我们首先通过检索 K 中的候选训练图像来获取输入图像与训练样本的相似度。主要用到的方法是利用基于与 GIST，PHOG 和对象库中的特征之间的距离组合的紧邻搜索。对于输入图像中的每个超像素，我们通过计算在每个候选者中的相应区域的平面参数，而且使用高斯处理来在许多的平面的参数中对感兴趣的超像素的相应平面参数进行估计。高斯处理的回归量取决于一个 RBF 核心，同时它被在训练数据中以一中留一图像输出的方法被训练。根据从回归分析的估计结果中，让 $d^i_{r,p}$ 来代表在超像素 p 中的第 i^{th} 个像素的深度。我们定义一个一元函数：

$$\Phi_p(y_p)=\frac{1}{N_p}\sum_{i=1}^{N_p}(d^i_p(y_p)-d^i_{r,p})^2 \qquad (3)$$

式中 N_p 是超像素 p 中的像素个数，d^i_p 对于一个特定的 y_p 状态来说，它是超像素 p 中的第 i 个像素的深度。

二元函数 $\Phi_{p,q}$ 依赖于从单一图像恢复遮挡边界研究方法中（14）的特征训练的一个闭环训练器。然后给出预测的闭环标签 O_{pq} 作为两个相邻的超像素 p 和 q 之间的边界，这个函数表达式是：

$$\phi_{p,q}(y_p,y_q)=w_l\cdot \begin{cases} 0 & \text{if } o_{pq}=1 \\ g_{pq}\|\mathbf{n}_p(y_p)-\mathbf{n}_q(y_q)\|^2+ & \\ \frac{1}{N_{pq}}\sum_{j=1}^{N_{pq}}(d^j_p(y_p)-d^j_q(y_q))^2 & \text{if } o_{pq}=0 \end{cases} \qquad (4)$$

式中 N_{pq} 代表超像素 p 和 q 共有的像素的个数，$n_p(y_p)$ 是对应于一个特定状态 y_p 的标准式，g_{pq} 是基于图像的超像素边界梯度的权重，$g_{pq}=\exp(-\mu_{pq}/\sigma)$ 是边界上的梯度。

虽然受到的从单个图像离散连续深度估计研究方法（21）的启发，但是上述所描述的能量表达式中至多只

包含二元项，而且还允许我们能够更加高效的进行推理。然而重要的是这个能量表达式依旧在局部层次依旧合理。下一步，我们将展示我们的方法如何通过在式子 1 中增加一些其他项来引入更高层次的场景结构信息。

3.2. 局部深度估计

上述所用到的超像素的相关知识只是很少的一部分，而且知识编码了场景的很少的一部分信息。正因如此，不仅仅是只能编码很少的结构信息，同时也无法可靠的利用它们的编码结构进行深度估计。于是只用它们在利用全局图像描述信息来进行检索到的候选图像的位置是利用到了先前的模型。为了更好的利用结构特征和编码出更多的关于场景的结构信息，这里我们建议充分使用更大的区域。

为此，让 R=$\{r_1, r_2, ..., r_{Nr}\}$作为代表从输入图像中提取到的 Nr 个区域的离散变量的集合，在集合中每一个 r_i 都能够被分配得到一个值从一个个相同的状态空间 S 中，并以此作为超像素变量$\{y_p\}$。我们增加第二个项在式一中：

$$E_m(Y,R)=\sum_{\Upsilon}\Phi_{\Upsilon}(r_{\Upsilon})+\sum_{\Upsilon,p}\Phi_{\Upsilon,p}(r_{\Upsilon},y_p) \quad (5)$$

在式中，Φ_{Υ} 是一个在区域变量中的一元函数，$\Phi_{\Upsilon,p}$ 是一个二元函数，用于表示区域和超像素之间的相互作用关系。

由于我们选择的区域比我们选择的超像素更大，所以它们的外表结构也会更有差别。因此，我们利用了一个从可扩展的非参数图像的超像素解析研究方法（30）中启发得到的无变量的基于特征的方法来定义一元函数 Φ_{Υ}。尤其是，我们首先利用用到的方法是利用基于与 GIST，PHOG 和对象库中的特征之间的距离组合的紧邻搜索进行检索了 Kr 中的候选训练图像。这里，在利用每个独立的特征种类进行过近邻搜索后，基于他们的最优级别，我们选择 Kr 个图像。我们发现这用策略会更加可靠相对于通过结合特征来得到更大的训练集合的方法。对于输入图像的每一个区域，我们计算出区域级别的特征向量，并且在剔除那些和查询区域距离过大或者种类太过相似的后，为每一个特征，在候选图像池中检索 Kc 的近邻区域。在每一个检索到的区域的每一个超像素来分别表决得到质心和法向在 V 维空间和三维空间的直方图。我们用 $P_d(d)$ 和 $P_n(n)$ 来表示上述产生的标准直方图，用 $P_{d,n}(d,n)$ 来表示上面两者的组合的 V 维直方图。我们在式五中来表示这个一元项：

$$\Phi_{\Upsilon}(r_{\Upsilon})=\omega_m*(\max(P_{dn}(d(r_{\Upsilon}),n(r_{\Upsilon})))-P_{dn}(d(r_{\Upsilon}),n(r_{\Upsilon})))$$

在式中 $d(r_{\Upsilon})$ 是相对于状态 r_{Υ} 的质心而且和法向方向在某种程度上很相似。

在式 5 中的二元项代表在进行它所覆盖的区域深度预测和超像素深度

预测的不一致性所进行的乘法函数。对于区域中的每一个超像素，这个式子定义为：

$$\Phi_{\Upsilon,p}(r_\Upsilon, y_p) = \frac{\omega_{m,l}}{N_p}\sum_{i=1}^{N_p}(d_p^i(y_p) - d_\Upsilon^i(r_\Upsilon))^2 \quad (6)$$

在式中，N_p 是在超像素中的像素的数量，带有下标 i 的 $d_p^i(y_p)$ 和 $d_\Upsilon^i(r_\Upsilon)$ 分别代表着第 i^{th} 个超像素中的像素的深度和其相对应的 ϒ 区域中的像素的深度。

注意，在该层我们的模型的能量可认为是被编码的超像素之间的范围更远的连接。然而，重要的是由此产生的模型保持着二元函数的性质。

3.2.1 提取区域

在这里，我们大致的描述了我们的策略来提取在上述的函数中作为中级结构的区域。我们的目标是为了包含一些区域，这些区域更优选于（接近于）平面且在相对均匀的外观条件下尽可能的大。为此，我们依赖于轮廓检测与分层图像分割研究（3）中的 gPb 分割架构。

由于我们的测试数据是有 RGB-D 的图片组成的，所以我们能够直接地利用 RGB-D 分割架构的延伸，该架构最近的介绍在从 RGB-D 图像中学习丰富的特征的目标检测和分割和 RGB-D 现场标签的功能和算法研究中（10,23）。然而在测试时间，我们仅能使用 RGB 图像。一种能够解决这个问题的简单方法是直接利用轮廓检测与分层图像分割研究（3）的院士理论。不幸的是，所得到的结果区域不是高度非平面的，就是太小，而这两种都不能正常的适应我们的目的。

为了解决这个问题，我们通过结合不同的信息来源提出了一个边界的概率计算。首先，我们需要把标准 gPb 算法应用到我们的输入图像中。作为第二信息来源，我们在形成利用几何推理的单个图像的结构恢复研究中（19）的方向图时使用了预估场景几何。方向图为图像中的像素指定一个主要的法线方向。不幸的是，这些地图是稀疏的（即，不是所有的像素被分配的方向）。此外，为了我们的目的，我们不希望拥有同一个方向的所有像素属于同一地区，因为它们可能属于不同的表面。因此我们计算了方向图的连接部分，并为每个像素分配了一个标签，用来说明他们是属于哪一个部分。之后，我们只为了为所得到的标签图像的 gPb 算法提供了亮度条件。

让我们分别用 gPb_{rgb} 和 gPb_g 来表示从 RGB 图像和几何图像中得到的边界概率。一个位置在(u, v)，边界角度为 θ 的合并边界概率如下：

$$gPb_c(u,v,\theta) = (1-\alpha)gPb_{rgb}(u,v,\theta) + \alpha gPb_g(u,v,\theta),$$

在实践中我们使用 $\alpha = 0.5$。为了包含最后的区域，我们之后提供了轮廓检测与分层图像分割研究（3）中的 OWT-UCM 理论，并这个组合边界图上的阈值设置为 0.1。我们发现 RGB 和几何线索对于大，平坦并且均匀的区域的

结合非常合适我们的方法。

3.3 与全局结构的结合作用

在我们表述的最终层中，我们的目的是了解场景的全局结构，而他并不是超像素，也不是区域能够模型化的。因此，我们运用了恢复杂乱房间的空间布局研究（11）中的布局估计方法。这种方法将一个室内场景模型化为一个像是盒子做的五面的几何体。(换言之，左/中/右面，屋顶和地面)，并且每一个像素的概率的额外语塞属于杂波。但是，请注意，该方法的输出并不是真正的三维表示，因为在这个意义上，该框的全球规模是不确定的。

为了能够使用这样的全局结构，让我们用 L 来代表能够编码预测布局规模的离散变量，它可以量化一个代表量化尺度的状态空间 L。我们模型的最后一层的能量可以被写作：

$$E_g(Y,L)=\sum_p \Phi_{L,p}(L,y_p) \qquad (7)$$

它由一个双变量函数组成，可以加强超像素和布局之间的连贯性。特别的，我们定义这个函数为：

$$\Phi_{L,p}(L,y_p)=\frac{\omega_g}{N_p}\sum_{i=1}^{N_p}(1-P_c^i)\cdot(d_p^i(y_p)-d_L^i(l))^2, \quad (8)$$

其中 P_c^i 代表着像素 i 属于杂波的概率。一个之前的相似的被提到并使用的标记 w.r.t.指数 i。$d_p^i(y_p)$ 和 $d_L^i(L)$ 代表着在超像素 P 中第个 i^{th} 像素的深度和在布局中它与像素的一致性。重要的是，杂波概率的使用防止我们将过度平滑了超像素预测的深度。

因为这个能量项是成对的，所以我们的整个模型都是成对的。在我们的实验中，我们利用分布式凸置信传播（DCBP）理论来在我们的 CRF 中执行推理。注意推理的结果区域不仅仅是一个由超像素得到的细节深度估计，同样也是该地区的深度估计，就好像是一个真正的 3D 场景布置。

4. 实验评估

我们对我们的方法在两个普通的可获取到的数据上进行评估：室内分割和 RGBD 图像的支持推论研究方法（29）中的 NYUv2 数据集和挑战 RMRC 比赛（1）中的 RMRC 室内数据集。这两个数据集中包含了重很多的室内场景中收集到的图像。对于 NYUv2 数据集，我们把我们的结果和艺术状态单张图像的深度估计方法进行了比较。尤其是，我们更考虑到了下面三个基本准则：

1. 深度转移（15）。这个方法预测深度是通过转移从训练的数据集中找到的相似的图像的深度地图。然后这些深度地图通过尽可能使得图像光滑连续的最优策略进行合并。
2. 离散连续深度（21）。这个技术是充分利用一个高阶离散连续 CRF 算法来估计深度，这个算

法中复杂的相邻超像素之间的关系可以通过离散变量来编码出来。

3. 语义深度（18）。这个方法在规范的深度数据集中的每个语义类中学习得到一个逐像素的分类器。注意，因此这个方法充分使用了一个在语义像素标签中的额外资源。同时注意，它利用一个数据集提供的一个不同的测试或者训练划分中进行训练。因此对于他们产生的结果也只是暂时性的作为参考。

为了完整期间，我们也会报告出基于多尺度深网的单图像深度图预测研究方法（5）中的超深度方法的结果。但是要注意的是这个方法依赖于一个更大的训练集合，这个集合有超过120000 张 NYUv2 中的原始图像,所以它不应该被看作为一个好的方法。

除了和这些方法进行比较，我们也对结果进行了消融研究，这个结果是分别从只有局部结构的模型，只有中层结构的模型和只有全局结构的模型中得到的。我们称这些模型分别为局部模型，中层模型，单全局模型。我们完整的模型将被成为总体模型。

对于我们的定量评价，我们提出以下三个标准的指标：平均相对误差（相对），平均 log10 的误差，以及均方根误差（RMS）。我们也是用了在(18)中提到的误差度量方法，定义为：

$$\%correct:(\frac{1}{N}\sum_{u=1}^{N}[\![\max(\frac{d_u}{g_u},\frac{g_u}{d_u})=\delta<t]\!])\bullet 100$$

式中 t 取 1.25 1.25^2 1.25^3，g_u 代表真实的在像素 u 的图像深度，d_u 是相应地方的估计深度值，N 是总计的图像中的像素个数，[[.]]表示指标函数。而且，尽管法线角度的估计不是我们的方法的主要研究方向，但是我们得到了五个法线的误差度量方法，这些都曾经在用于单图像理解的数据驱动的三维图元研究的方法（6）中被使用过：估计的法线角度和真实的角度差的平均值和中值，以及和真实值之间的角度差小于某个阈值的像素所占的比例，阈值可以去 11.25,22.5 和 30 度左右。为了评价这些度量方法，我们利用用于单图像理解的数据驱动的三维图元研究的方法（6）来从预测的深度地图中估计出场景的法线。

在我们的试验中，超像素是通过 SLIC 算法进行实现的。对于每一个测试图像，我们从训练图像中检索出 7 个候选图像，从而得到输入到超像素的回归模型。对于区域而言，我们就需要检索出 250 的候选图像。对于每一个查询的区域和每个局部特征而言，在对所有的候选区域中的质心和要查询区域的质心的距离超过了 100 并且区域和查询区域的大小比值小于 0.2 的进行删减，然后获得 30 个候选区域。当建立法线取向的直方图时，我们只考虑其角度小于 45 度并且三维空间中的三个角度中至少有一个和查询图像中的法线的三个角度中的一个相同。这样就允许我们能够舍弃一些

在查询图像中和场景中的法线方向差异较大的候选者。

我们设定的超像素和区域变量的状态值是通过量化 0.5 到 10 以内的步长为 0.5 的数来代替的。三个法线方向相结合，这就会产生有 60 个状态的变量。在具体的实现过程中，为了加快推理的速度，我们将P_{dn}的值限制在 20 个，P_{dn}是指在为区域一元函数建立的三维直方图中出现频率最高的变量。值得注意的是，我们这样的行为只会影响结果的降低很少的精确度。通过这样的设定，我们大致只需要两分钟左右就可以得到单张图像的深度估计所需要的大致 650 个超像素和结果相应的 gPb。

我们的 CRF 算法参数是通过从训练数据中抽取 69 张图像进行验证得到的最后结果。为此，我们一直遵循着一个策略，就是在先前的权值被确定后，函数表达式会逐渐增加到能量表达式中。需要注意的是我们没有很好的调整权重，但是却发现了大多数的每个隐藏的幅度的正确的排列顺序。

NYUv2:

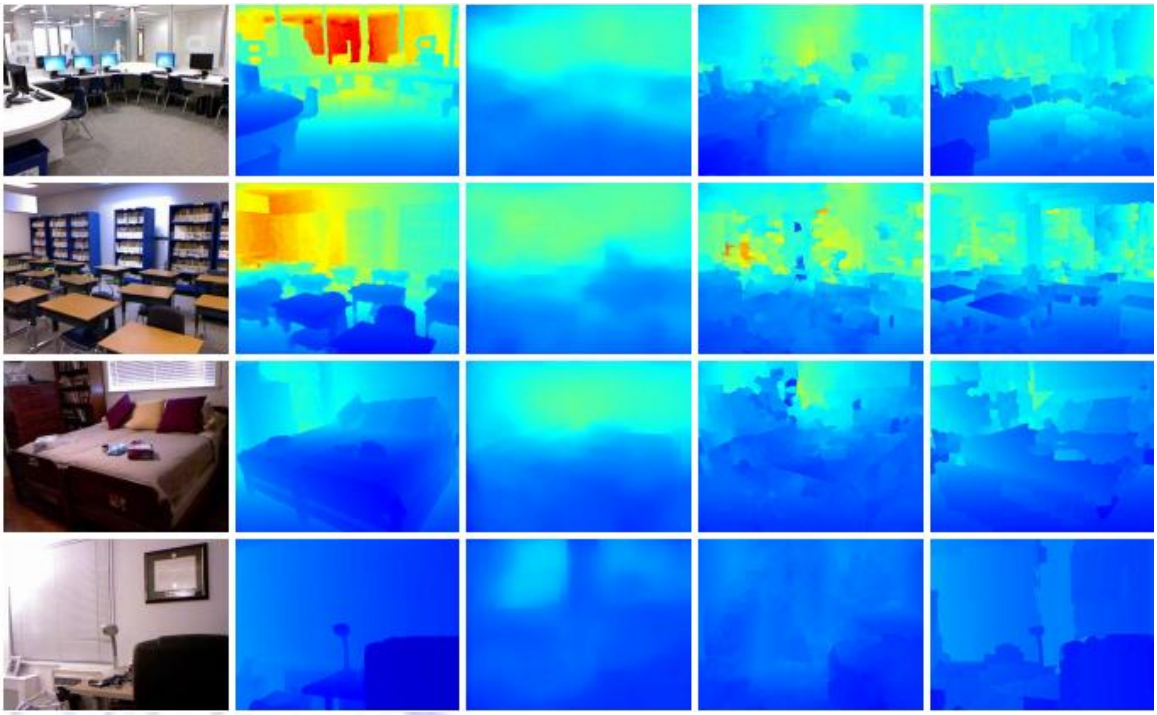
对于 NYUv2 的深度数据集中包含了 1449 对 RGB 和深度均衡分布的图像，然后将它们划分为 795 张训练使用的图像和 654 张测试图像。这些图像被要求能够满足各种各样的室内的场景图像。每张图片都被裁剪为大小为 427*561 像素的统一大小。在我们的测试中，我们充分利用了每张图像中的仅考虑非零深度的真实中的像素的掩膜。

在表 1 中，显示的数据分别是我们的方法测试得到的数据和基准方法测试的数据结构。就深度估计的精确性来说，我们在所有结果的错误率上的度量标准比深度转化和离散连续深度估计方法做得更好，同时也达到了超过三分之二的语义深度估计中的阈值预测，尽管事实上，语义深度估计方法在训练过程中利用了更多的额外知识。深度神经网络的深度估计方法的结果的深度指标如下。虽然他们更加精确了，但是，为了获得神经网络的所有参数，就需要更加庞大的数据集的支持。就普通的精度而言，我们的的方法的精度比基准测试方法的精度的三分之五更高。图 2 提供了利用不同方法在一些图像上的深度地图的恢复的定义比较示意图。总的来说，这些结果证明了使用中层和全局结构的方法更加有利。

在表 2 中，我们提供了我们模型的不同部分的分析。分析的结果表明一个事实，就是模型中的每一层都会对结果的精确性造成影响。它同时也显示出中层结构的对结果的影响占了很大一部分。在图 3 中，我们提供了一个我们模型的中各个不同部分单独作用的结果的定性比较。尽管在这个尺度上看起来不明显，但是我们通过观察发现，他们仍然保留着图像中的不连续性。而且，全局结构的结果拥有更加精确的深度顺序在整幅图中。

Method	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	mean	median	$\theta < 11.25$	$\theta < 22.5$	$\theta < 30$
DepthTransfer	0.374	0.134	1.12	49.81%	79.46%	93.75%	**43.0**	**40.5**	6.9%	23.2%	34.9%
DC-Depth	0.335	0.127	1.06	51.55%	82.32%	95.00%	45.7	42.2	19.7%	25.7%	35.4%
SemanticDepth	-	-	-	**54.22%**	82.90%	94.09%	-	-	-	-	-
Ours	**0.305**	**0.122**	**1.04**	52.50%	**83.77%**	**96.16%**	46.7	41.9	**21.1%**	**35.2%**	**41.7%**

表 1. NYUv2：我们的方法和几种基准方法的结果比较。就深度估计的精度而言，我们比其中两个基准测试方法在同样的输入集合下结果更好。而且，相对于语义深度方法，我们能够达到其三分之二的阈值要求，这还是在不管我们没有使用任何像素标签信息的基础上的结果。话说回来，语义深度方法利用了不同的训练和测试部分。就法向精度而言，我们比接近五分之三的基准测试方法要好。



原图　　真实深度图　　深度转换方法　　DC 深度方法　　我们方法

图 2. NYUv2：大致比较。通过各种不同的基准测试方法和我们方法的深度估计结果的比较。注意，我们的方法典型的避免了深度转换方法的过度模糊的问题，同时还能过比 DC 深度方法获取更好的场景结构。

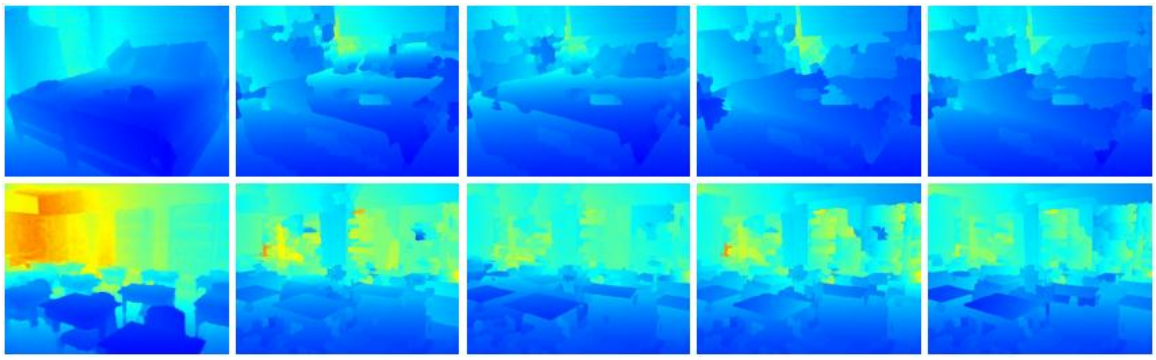
除了估计得到超像素的深度外，我们的模型也可以预测出局部和全局层次的深度。一些结果深度图在图 4 中进行了展示。

RMRC Indoor：

然后我们利用 RMRC 室内数据集对我们的方法进行了评估。由于这个数据集并没有提供真实的测试图像的深度信息，而且也是由于我们要对模型中的各个不同的组成部分进行评估，所以我们只是利用了 4105 训练图像，并且从这些图像中我们随机的取出 114 张图像组成了一个测试数据集。在这个实验中，我们使用了和 NYUv2 测试中相同的参数。在表 3 中，我们提供了很多种的错误评判标准分别用于各

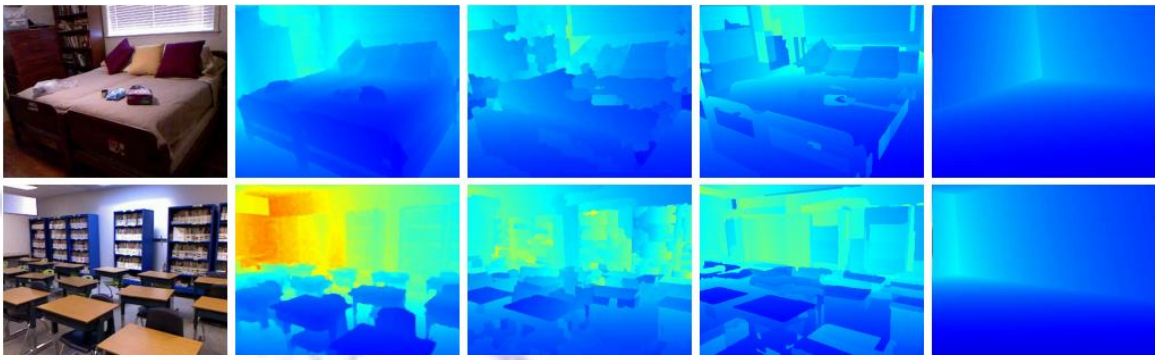
Method	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Ours-local	0.334	0.128	1.05	50.35%	82.31%	95.44%
Ours-mid	0.312	0.123	**1.03**	52.08%	**83.92%**	96.13%
Ours-global-only	0.325	0.128	1.07	50.38%	82.06%	95.35%
Ours	**0.305**	**0.122**	1.04	**52.50%**	83.77%	**96.16%**

表 2. NYUv2：消融研究。我们评估的我们的模型中的各个组成部分对实验最后的结果的影响程度。结果证明模型中的每个部分对于结果都有着贡献，但是相对而言，中层结构层次的作用更大。



真实深度　局部测深度　中层测深度　单全局测深度　整体测深度

图 3. NYUv2：消融研究。不同模型中的组成部分得到的深度地图。



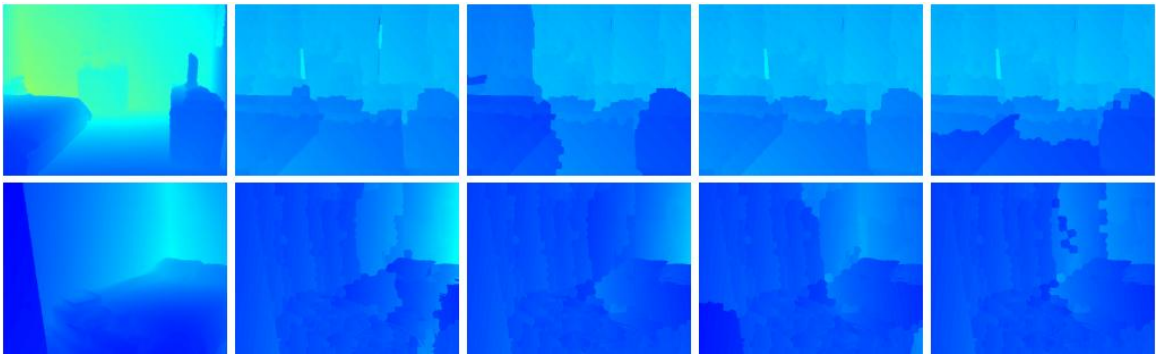
原图像　真实深度图　超像素　区域　层次

图 4. NYUv2：模型中不同层次阶段的深度图像。我们展示了最终结果的图像和相应的模型中的各个层次中间产生的深度图的联系。

个不同模型中的部分。对于 NYUv2 的测试，我们可以看出每个模型中的部分对于最后的结果的贡献。然而，这个测试集中，中层结构对于结果的影响看起来比 NYUv2 中的更大。为了给读者提供一个大致的关于我们和其他的方法的结果比较的概况，注意在 RMRC 中的测试数据中，最好的相关估计深度的错误率是 0.33 对于基于多尺度深网的单图像深度图预测研究的方法（5）来说，第二则是 Baig 和 Torresani 的方法，错误率为 0.39。在图像 5 和 6 中，我们分别展示了我们方法中不同部分所得到的深度地图信息和最终模型中的不同层次的变量所预测得到的深度地图。

Method	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Ours-local	0.440	0.167	1.24	39.38%	72.41%	89.83%
Ours-mid	0.395	**0.159**	**1.22**	**41.25%**	**74.29%**	**90.75%**
Ours-global-only	0.423	0.167	1.26	38.64%	71.09%	88.76%
Ours	**0.379**	**0.159**	**1.22**	40.67%	73.67%	90.01%

表 3. RMRC Indoor：消融研究。我们评估的我们的模型中的各个组成部分对实验最后的结果的影响程度。就像 NYUv2，结果证明模型中的每个部分对于结果都有着贡献，但是相对而言，中层结构层次的作用更大。



真实深度　　局部测深度　　中层测深度　　单全局测深度　　整体测深度

图 5. RMRC Indoor：消融研究。不同模型中的组成部分得到的深度地图。

原图像　　真实深度图　　超像素　　区域　　层次

图 6. RMRC Indoor：模型中不同层次阶段的深度图像。我们展示了最终结果的图像和相应的模型中的各个层次中间产生的深度图的联系。

5. 实验总结

我们已经介绍了一种利用场景在不同层次的细节结构的来进行的单一图像深度估计方法。我们的实验说明了这样的一种感知结构方法相较于局部深度预测理论的好处。尤其是我们的评估证实了这样一个事实：中级结构，例如区域，在模型最终的精确性中做出了最巨大的贡献。在未来，如果这个现象能够被利用并在我们的模型中发挥更多的潜力，那么我们将会去调查和学习关于他的更多细节。除此之外，我们计划将语义标签的使用吸收进我们的深度预测框架。

6. 特此鸣谢

第一位作者有中国奖学金委员会支持。NICTA 是由宽带，通信和数字经济部门和 ARC 通过卓越计划 ICT 中心而代表的澳大利亚政府赞助。

引用

[1] RMRC challenge 2014. http://cs.nyu.edu/silberman/rmrc2014/, June 2014. 5, 7

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Suesstrunk. Slic superpixels compared to state-of-the-art superpixel methods. PAMI, 2012. 5

[3] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on,33(5):898－916, 2011. 4

[4] M. H. Baig, V. Jagadeesh, R. Piramuthu, A. Bhardwaj,W. Di,and N. Sundaresan. Im2depth: Scalable exemplar based depth transfer. In Applications of Computer Vision (WACV),2014 IEEE Winter Conference on, pages 145－152. IEEE,2014. 2

[5] D. Eigen, C. Puhrsch, and R. Fergus. Depth map predictionfrom a single image using a multi-scale deep network. arXiv preprint arXiv:1406.2283, 2014. 1, 2, 5, 6,7

[6] D. F. Fouhey, A. Gupta, and M. Hebert. Data-driven 3d primitives for single image understanding. In Computer Vi-sion (ICCV), 2013 IEEE International Conference on, pages3392－3399. IEEE, 2013. 2, 5

[7] D. F. Fouhey, A. Gupta, and M. HebertUnfolding an indoororigami world. In Computer Vision－ECCV 2014, pages 687－702. Springer, 2014. 1, 2

[8] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In Computer Vision－ECCV 2010, pages 482－496. Springer, 2010. 1

[9] A. Gupta, M. Hebert, T. Kanade, and D. M. Blei. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In Advances in Neural Information Processing Systems, pages 1288－1296, 2010. 1

[10] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In European Conference on Computer Vision, 2014. 4

[11] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In Computer vision, 2009 IEEE 12th international conference on, pages 1849－1856. IEEE,2009. 4

[12] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In Computer Vision－ECCV 2010, pages 224－237.Springer, 2010. 1, 2

[13] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surfacelayout from an image. International Journal of Computer Vision, 75(1):151－172, 2007. 1, 2

[14] D. Hoiem, A. N. Stein, A. A. Efros, andM. Hebert. Recovering occlusion boundaries from a single image. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1－8. IEEE, 2007. 3

[15] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In Computer Vision ECCV 2012, pages 775－788. Springer, 2012. 1, 2, 5

[16] J. Konrad, G.Brown, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee. Automatic 2d-to-3d image conversion using 3d examples from the internet. In SPIE Stereoscopic Displays and Applications, 2012. 2

[17] J. Konrad, M. Wang, and P. Ishwar. 2d-to-3d image conversion by learning depth from examples. In 3DCINE, 2012.2

[18] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 89－96. IEEE,2014. 1, 2, 5

[19] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conferenceon, pages 2136－2143. IEEE, 2009. 1, 2, 4

[20] B. Liu, S. Gould, and D. Koller. Single image depth esti-mation from predicted semantic labels. In Computer Visionand Pattern Recognition (CVPR), 2010 IEEE Conference on,pages 1253－1260. IEEE, 2010. 1, 2

[21] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 716－723. IEEE, 2014. 1, 2, 3, 5

[22] V. Nedovic, A. W. Smeulders, A. Redert, and J.-M. Geusebroek. Stages as models of scene geometry. Pattern Analysis and Machine Intelligence, IEEE Transactions on,32(9):1673－1687, 2010. 2

[23] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2759－2766.IEEE, 2012. 4

[24] C. Rother. A new approach to vanishing point detection in architectural environments. Image

and Vision Computing, 20(9):647－655, 2002.3

[25] A. Saxena, S. H. Chung, and A. Y. Ng. 3-d depth reconstruction from a single still image. IJCV, 2007. 1, 2

[26] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. PAMI, 2009. 1, 2

[27] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed message passing for large scale graphical models. In CVPR, 2011. 4

[28] A. G. Schwing, S. Fidler, M. Pollefeys, and R. Urtasun. Box in the box: Joint 3d layout and object reasoning from single images. In Computer Vision (ICCV), 2013 IEEE International Conference on, pages 353－360. IEEE, 2013. 1, 2

[29] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In ECCV, 2012. 5

[30] J. Tighe and S. Lazebnik. Superparsing: scalable nonpara-metric image parsing with superpixels. In Computer Vision ECCV 2010, pages 352－365. Springer, 2010. 3

[31] A. Torralba and A. Oliva. Depth estimation from image structure. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(9):1226－1238, 2002. 2