# Deeper Depth Prediction with Fully Convolutional Residual Networks

摘要

本文解决了给定单个RGB图像估计场景深度图的问题。我们提出一个全卷积残差结构，以模拟单目图像和深度图之间的模糊映射。为了提高输出分辨率，我们提出了一种新颖的方式来有效地学习网络中的特征图上采样。为了优化，我们引入了反向Huber损失，特别适用于手头的任务，由深度图中存在的分布值驱动。我们的模型由单一结构组成，经过端对端的训练，不依赖后处理技​​术，如CRF或其他细化步骤。因此，它可以在图像或视频上实时运行。在评估中，我们表明所提出的模型包含较少的参数，并且比当前的现有技术要求更少的训练数据，同时优于所有深度估计方法。

1. 简介

从一个角度来看，深度估计是一个像计算机视觉一样古老的学科，并且涵盖了多年来开发的几种技术。这些技术中最成功的一个是Structure-from-Motion（SfM）[34]; 它利用摄像机运动来通过不同的时间间隔估计摄像机姿态，并且通过来自成对的连续视图的三角测量来估计深度。作为运动的替代方案，可以使用其他工作来估计深度，例如照明的变化[39]或焦点[33]。

在没有这样的假设环境情况下，由于将强度或颜色测量映射到深度值的内在模糊性，来自通用场景的单个图像深度估计是不明确的问题。虽然这也是人类的大脑限制，但深度知觉仍然可以从单目视觉出现。因此，开发能够通过利用单目线索来估计深度图的计算机视觉系统不仅是一项具有挑战性的任务，而且在直接深度感测不可用情况下是必需的。此外，众所周知，相当准确的深度信息的必要性，以改善关于只有RGB对应物的许多计算机视觉任务，例如在重建[23]，识别[26]，语义分割[5]或人体姿态估计[35 ]。

有几件作品解决了单眼深度估计的问题。第一种方法是通过Markov Random Fields (MRFs马尔科夫随机字段)将超像素视为平面，推断通过平面系数的深度[30]。在[16,20,37]中还考虑了超像素，其中部署了Conditional Random Fields (CRF条件随机字段）用于深度图的正则化。Data-driven(数据驱动)方法，如[10,13]提出了基于手工特征进行图像匹配来检索给定查询图像的训练集中最相似的候选者。 然后将相应的深度候选者扭曲并合并，以产生最终结果。

近来，已经使用Convolutional Neural Networks（卷积神经网络CNN）来学习色彩像素与深度之间的隐含关系[5,6,11,19,37]。 CNN方法经常与基于CRF的正规化相结合，作为后处理步骤[16,37]或通过结构化深度学习[19]以及随机森林[27]。这些方法包括较高的复杂性，这是由于深层网络中涉及的参数数量很多[5,6,19]，或CNN和CRF联合使用[16,37]。然而，深度学习提高了准确性标准的基准数据集相当多，将这些方法列为最先进的方法。

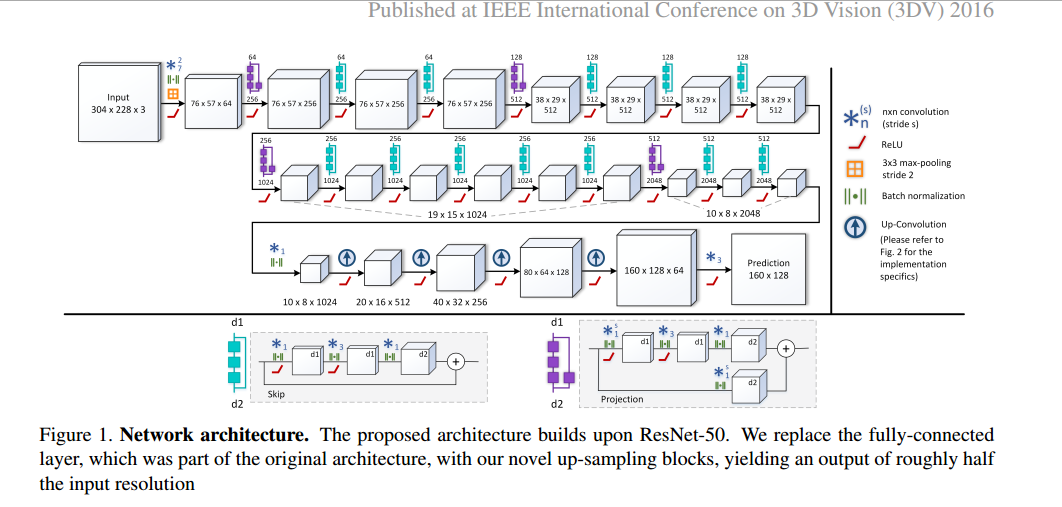
在这项工作中，我们建议使用CNN学习单个RGB图像与其相应的深度图之间的映射。我们工作的贡献如下。首先，我们引入一个全卷积结构进行深度预测，具有新颖的上采样块，允许更高分辨率的密集输出图，同时要求比数据量少一个数量级的参数，训练数量少，在标准基准数据集上优于所有现有方法[23,29]。我们进一步提出了一个更有效的上调方案，并将其与残差学习的概念相结合[7]，为特征图的有效上采样创建投影块。最后，我们通过基于反向Huber函数（berHu）[40]优化损失来训练网络，并在理论和实验两方面证明为什么它有益于更好地适合于当前的任务。我们彻底评估网络深度，损失函数和上采样的具体层次的影响，以分析其优势。最后，为了进一步评估我们的方法的准确性，我们在3D重建场景中采用训练模型，其中我们使用一系列RGB帧及其预测的同时定位和映射（SLAM）的深度图。

1. 相关工作

图像数据的深度估计最初依赖于立体视觉[22,32]，使用相同场景的图像对来重建3D形状。在单视图情况下，大多数方法依赖于运动（Structure-from-Motion [34]）或不同的拍摄条件（Shape-from-Shading [39]，Def-from Defocus [33]）。 尽管由于缺乏这种信息而引起的歧义，但是通过与单眼线索的人类深度感知类比，灵感来自单一RGB图像的深度图预测也已被研究。下面我们来关注单RGB输入的相关工作，类似于我们的方法。

单眼深度估计的经典方法主要依靠手工制作的特征，并使用概率图形模型来解决问题[8,17,29,30]，通常对场景几何作出假设。 Saxena等人的第一部作品之一 [29]使用MRF从图像中提取的局部和全局特征推断深度，而在MRF公式中引入超像素[1]以实现邻近约束。他们的工作后来延伸到3D场景重建[30]。受这项工作的启发，Liu et al. [17]将语义分割的任务与深度估计相结合，其中使用预测标签作为附加约束来促进优化任务。Ladicky等人 [15]而是在分类方法中共同预测标签和深度。

第二组相关工作包括深度传递的非参数方法[10,13,18,20]，其通常执行给定RGB图像与图像之间的基于特征的匹配（例如GIST [24]，HOG [3]）的RGB-D存储库，以便找到最近的相邻; 然后将检索到的深度对应物扭曲并组合以产生最终深度图。Karsch等人 [10]使用SIFT Flow [18]执行扭曲，其次是全局优化方案，而Konrad等人 [13]计算所检索的深度图的中值，然后进行双边滤波以进行平滑处理。 Liu等人[20]不是对候选人进行扭曲，而是将优化问题作为具有连续和离散的可变电位的条件随机场（CRF）来制定。不可否认，这些方法依赖于RGB图像中区域之间的相似性意味着 也有类似的深度线索。



最近，深度学习领域的显着进步推动了研究使用CNN进行深度估计。由于任务与语义标签密切相关，大多数作品都基于ImageNet Large Scale Visual Recognition Challenge (ImageNet大型视觉识别挑战ILSVRC）[28]最成功的结构，通常使用AlexNet [14]或更深层次的VGG [31]初始化其网络，Eigen et al.[6]已经是第一个使用CNN从双尺度结构中的单个图像中回归密集深度图，其中基于AlexNet的第一阶段产生粗略输出，第二阶段优化原始预测。他们的工作后来被扩展到另外预测了基于VGG的更深入和更歧视的模型的正态和标签，以及一个三维结构进一步完善[5]。与[5,6]的深层结构不同，Roy和Todorovic [27]提出将CNN与回归森林相结合，在每个树节点使用非常浅的结构，从而限制了对大数据的需求。

提高预测深度图质量的另一个方向是综合使用CNN和图形模型[16,19,37]。Liu et al. [19]提出在CNN训练中以CRF损失的形式学习一元和双向的潜力（原文为：Liu et al. [19] propose to learn the unary and pairwise potentials during CNN training in the form of a CRF loss），而Li[16]和[37]使用分级CRF来优化从超级像素向下到像素级别的补片CNN预测。

我们的方法使用CNN进行深度估计，与以前的工作不同之处在于，通过采用完整的卷积模型并入有效的残差向上采样块，可以改进典型的完全连接层，这些层对于参数数量是冗余的，我们称之为高估，在解决高维回归问题时更为适用。

1. 方法

在本节中，我们从单个RGB图像描述我们的深度预测模型。 我们首先介绍所采用的结构，然后分析这项工作中提出的新组件。 随后，我们提出一个适用于优化给定任务的损失函数。

### 3.1 CNN结构

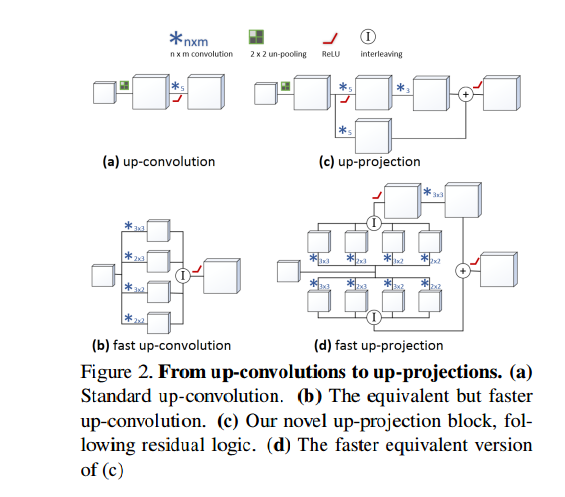
现在几乎所有的CNN结构都包含一个收缩部分，通过一系列的卷积和汇集操作逐渐降低输入图像分辨率，从而使更高级别的神经元具有较大的接收场，从而捕获更多的全局信息。 在所需输出为高分辨率图像的回归问题中，为获得较大的输出图需要某种形式的上采样。 Eigen et al. [5，6]，使用全连接层，如在典型的分类网络，产生一个完整的接受领域，然后将结果重新整理为输出分辨率。

我们引入了一个完全卷积网络进行深度预测。在这里，接受领域是建筑设计的一个重要方面，因为没有明确的完整连接。具体来说，假设我们设置了一个304×228像素的输入（如[6]所示），并且预测输出图将大约是输入分辨率的一半。我们调查流行的结构（AlexNet [14]，VGG-16 [31]）作为收缩部分，因为它们的预训练权重有助于收敛。 AlexNet的最后卷积层的接收场是151×151像素，当真实的全局信息（例如单眼线索）应该被网络捕获而没有完全连接的层时，只允许非常低分辨率的输入图像。通过VGG-16实现了276×276的较大的接收场，但仍然对输入分辨率设置了限制。Eigen和Fergus [5]在从AlexNet转换到VGG时显示出实质性的进步，但是由于两者的模型都使用完全连接的层，这是由于VGG的更高的辨别力。

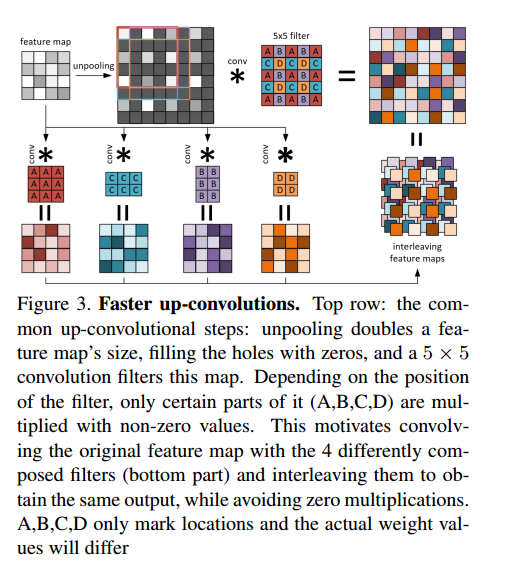
最近，ResNet [7]引入了跳过层次，它们绕过两个或更多个卷积，并将其与输出相加，包括每次卷积之后的批量归一化[9]（见图1）。按照这种设计，可以创建更深层次的网络，而不会面临弥散或消失的梯度。这些非常深刻的结构的另一个有益的特性就是他们的大型接受领域; ResNet-50捕获483×483的输入大小，足够大，即使在较高分辨率下也能完全捕获输入图像。考虑到我们的输入大小和这种结构，最后的卷积层导致了2048个空间分辨率10×8像素的特征图，当去掉最后一个层叠层时。如我们稍后所示，所提出的使用残余上卷积的模型产生160×128像素的输出。如果我们替代地添加了相同大小的完全连接的层，那么它将引入3.30亿的参数，值为12.6GB的内存，这使得这种方法在当前的硬件上是不可能的。这进一步推动了我们提出了一个完全卷积结构，其中包含较少权重的上采样块，同时提高了预测深度图的准确性。

我们提出的结构可以在图中看到。特征图尺寸对应于训练用于输入尺寸304×228的网络，在NYU深度v2数据集的情况下[23]。网络的第一部分基于ResNet-50，并使用预先训练的权重进行初始化。我们的结构的第二部分指导网络通过一系列的未加工和卷积层来学习其升级。在这些上采样块的集合之后，通过产生预测的最终卷积层来应用和成功丢弃。

Up-Projection Blocks(up-projection 块) 反池化层[4,21,38]执行池的反向操作，增加了特征映射的空间分辨率。我们调整[4]中描述的方法来执行解压缩层，以便通过将每个条目映射到2×2（zero）内核的左上角来将尺寸加倍。每个这样的层之后是5×5卷积，使得它被应用于每个位置处的多于一个非零元素，并且通过ReLU激活连续地应用。我们将此块称为上卷积。在经验上，我们堆叠四个这样的上卷积块，即最小特征图的16倍放大，导致存储器消耗和分辨率之间的最佳权衡。 我们发现添加第五个块时性能没有增加。



我们使用类似但是相反的概念进一步扩展简单的up-convol [7]来创建上采样重构块。这个想法是在上卷积之后引入简单的3×3卷积，并从下分辨率特征图添加到结果的投影连接，如图2（c）所示。由于尺寸不同，小尺寸地图需要在投影分支中使用另一个上卷积进行上采样，但是由于两个分支只需要应用一次解压缩，因此我们只需将5×5卷积分别应用于两个分支。我们称之为新的上采样块up-projection ，因为它将投影连接[7]的概念扩展到上卷积。链接up-projection 块允许在网络中更高效地传递高级信息，同时逐渐增加特征图大小。这使得我们能够构建一致的完全卷积网络进行深度预测。图2显示了上卷积块与up-projection 块之间的差异。它还显示了相应的快速版本，将在下面的部分中介绍。



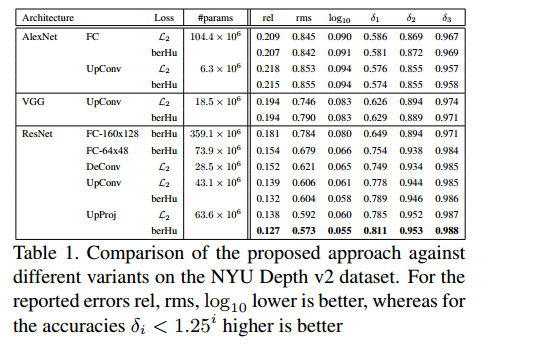
Fast Up-Convolutions(快速上卷积) 这项工作的另一个重要贡献就是重新整合上卷积操作，使其更有效率，导致整个网络的训练时间减少了15％左右。这也适用于新推出的up-projection 操作。主要的直觉如下：在unpooling后，75％的结果特征图包含零，因此以下5×5卷积主要以零进行操作，这可以在我们修改后的配方中避免。这可以在图3中观察到。在左上角，原始特征图未被打开（顶部中间），然后由5×5滤波器进行卷积。我们观察到，在未打开的特征图中，根据5×5滤波器的位置（红色，蓝色，紫色，橙色边界框），只有某些权重与潜在的非零值相乘。这些重量落入四个不重叠的组，由不同的颜色表示，图中的A，B，C，D。基于滤波器组，我们将原始的5×5滤波器布置成尺寸为（A）3×3，（B）3×2，（C）2×3和（D）2×2的四个新滤波器。现在可以通过交织如图3所示的四个所得到的特征图的元素来实现与原始操作（解压缩和卷积）相同的输出。从简单的上卷积块到所提出的up-projection 的相应变化如图2（d）所示。

### 3.2 损失函数

回归问题优化的标准损失函数是损失，最小化预测和真值之间平方的欧几里德范数，虽然这在我们的测试用例中产生了良好的结果，但我们发现使用反向Huber（berHu）[25,40]作为损失函数产生比更好的最终误差。

当并等于范围之外Berhu损失等于。这里使用的版本是在从到的切换发生的点c处的连续和一阶可微分。在每个梯度下降步骤中，当我们计算时，我们设置，其中i对当前批次中每个图像的所有像素进行索引，即每批次错误的最大值的20%。在实践中，BerHu在给定问题的两个规范之间表现出良好的平衡; 由于L2项，它对具有高残差的样本/像素具有较高的权重，例如与强大的损失相反，忽略具有高残差的样本的Tukey的重量函数[2]。与此同时，L1对比L2的较小残差的影响更大。

我们提供两个关于L2和berHu损失之间的区别的直觉（原文：We provide two further intuitions with respect to the difference between L2 and berHu loss）。在我们实验的两个数据集中，我们观察到深度值的重尾（heavy-tailed）分布，也在[27]中报道，Zwald和LambertLacroix [40]表明，berHu损失函数更合适。这也可以解释为什么[5,6]在预测深度值的对数时会有更好的收敛，有效地将对数正态分布拟合到高斯。 其次，由于L1衍生物大于L2，我们看到berHu在训练中的小残差中获得了更大的收益。这表现在相关的错误措施。和δ1（第4节），对小误差更敏感。



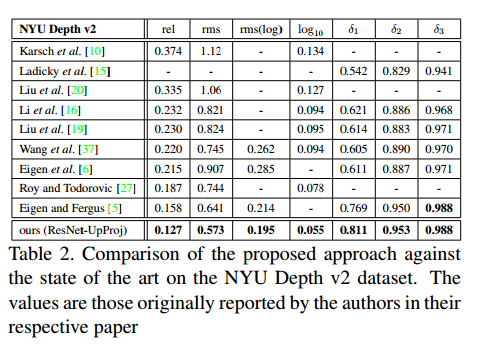
1. 实验结果

在本节中，我们将全面分析我们的方法，评估构成CNN结构的下采样和上采样部分的不同组件。我们还报告了我们的模型获得的定量和定性结果，并与两个用于深度预测的标准基准数据集（即NYU Depth v2 [23]（室内场景））和Make3D [30]（室外场景）进行了比较。

### 4.1 实验装置

为了实现我们的网络，我们使用MatConvNet [36]，并在单个具有12GB GPU内存的NVIDIA GeForce GTX TITAN上进行训练。 结构的下采样部分的权重层由对图像分类的ILSVRC [28]数据预先训练的相应模型（AlexNet，VGG，ResNet）进行初始化。 上升采样部分的新添加层被初始化为从正态分布采样的随机滤波器，其平均值为零，方差为0.01。

该网络对RGB输入进行了训练，以预测相应的深度图。 我们使用数据增加来增加训练样本的数量。 输入图像和对应的地面实况使用小的旋转，缩放，颜色变换和翻转0.5次机会进行变换，Eigen [6]。 最后，我们通过增加图像的随机作物将小型翻译建模到网络的所选输入大小。



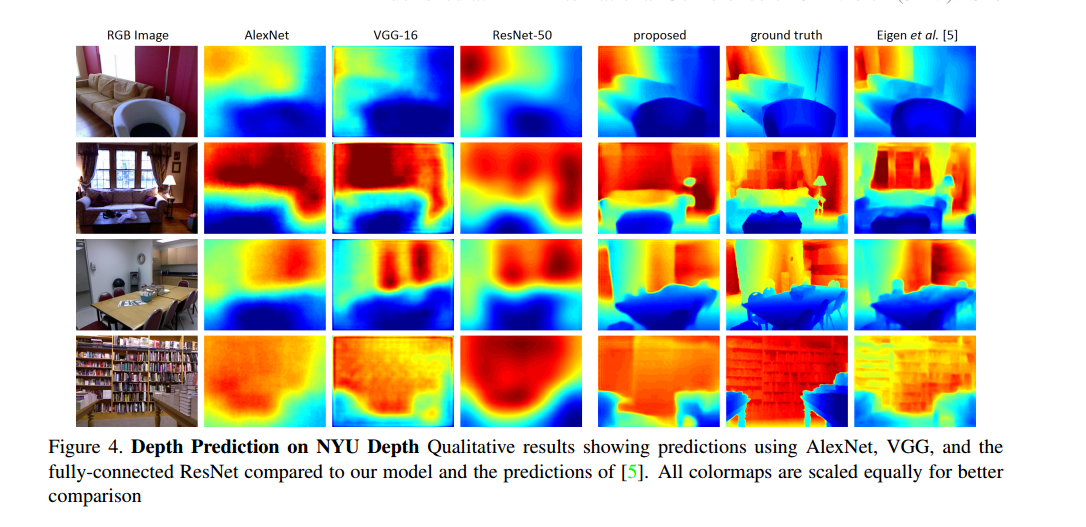
对于下面的定量评估，在我们的实验结果中计算了在以前的工作中使用的相同的误差度量[5,6,15,16,19]。

### 4.2 NYU深度数据集

首先，我们评估了用于室内场景重建的最大的RGBD数据集之一，NYU Depth v2 [23]。原始数据集由464个场景组成，由Microsoft Kinect捕获，官方分组包括249个训练和215个测试场景。然而，对于训练，我们的方法只需要原始分布的一小部分。我们从每个训练序列中采样等距离的帧，导致约12k个独特的图像。在提取的帧离线增强之后，我们的数据集包括大约95k对RGB-D图像。我们指出，我们的数据集大小小于在[5,6]中训练模型所需的数据集，其中包括120k个独特的图像，以及[16]的补丁方法中提取的800k个样本。以下[6]，尺寸为640×480像素的原始帧被下采样为1/2分辨率，并且中心裁剪为304×228像素作为网络的输入。最后，我们以大约20个回合的batch size为16训练我们的模型。所有层次的起始学习率为，当我们观察到极值时，我们逐渐减少每6-8个回合，动量为0.9。

对于我们的方法的定量评估和与该数据集的现有技术的比较，我们计算了654个图像的常用测试子集中的各种误差测量。预测的大小取决于具体的模型; 在我们的配置中，由四个上采样阶段组成，AlexNet的相应输出分辨率为128×96，VGG为144×112，基于ResNet的模型为160×128。 然后使用双线性插值将预测上采样回到原始大小（640×480），并将其与提供的真实值与无效像素的填充深度值进行比较。

结构评估(**Architecture Evaluation**) 在表1中，我们比较了所提出的结构的不同CNN变体，以便研究每个组件的效果。首先，我们使用AlexNet，VGG-16和ResNet-50的卷积块来评估结构深度的影响。显而易见的是，AlexNet上的完全卷积结构（UpConv）的表现优于具有全连接（FC）的典型网络。详见Sec.3.1，这样做的一个原因是AlexNet中的相对较小的视野，这不足以捕获删除全连接的层所需的全局信息。相反，使用VGG作为核心结构，提高了深度估计的准确性。作为高维度回归的完全连接的VGG变体将包含大量参数，我们仅在这里对全卷积（UpConv）模型执行测试。然而，具有完全连接层的基于VGG的模型确实被采用[5]（其结果见表2），其性能优于我们的完全卷积VGG变体，主要是由于其多尺度结构，包括细化尺度。

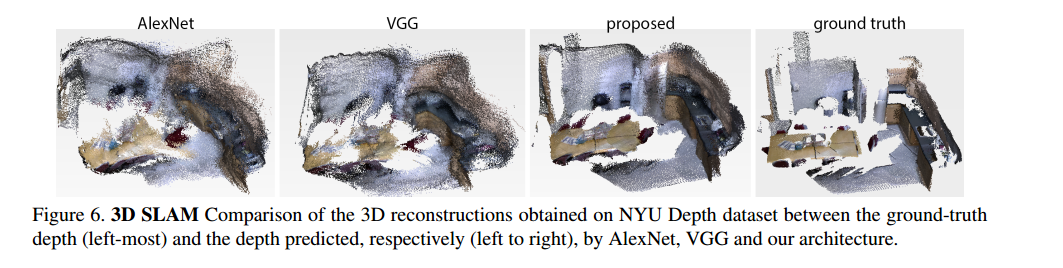


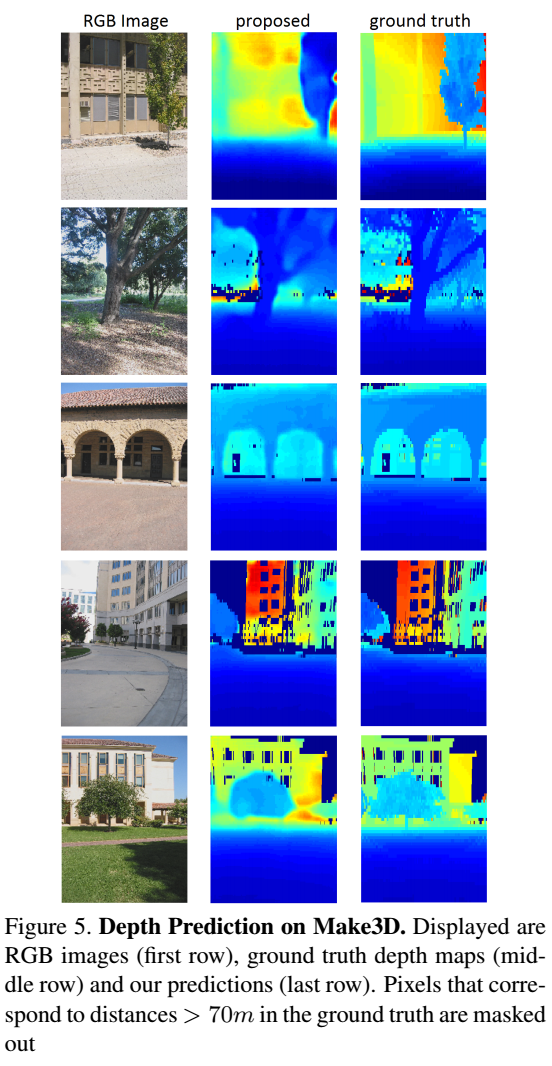
最后，使用全连接的层（ResNet-FC）切换到ResNet，而不去除最终的池化层 - 对于低分辨率输出（64×48），使用10倍的数据实现与[5]类似的性能;然而，增加输出分辨率（160×128）会导致如此大量的参数使收敛变得更加困难。这进一步激发了在处理高维问题时更换全连接层的理由和更有效的上采样技术的需要。我们使用简单的up-convolutions（ResNet-UpConv）的完全卷积变体提高了精度，最后，通过up-projection 块增强的建议结构（ResNetUpProj）给出了最好的结果。就参数的数量而言，当从全连接的层切换到全卷积网络时，我们看到大幅减少。我们研究的另一种常见的上采样技术是连续2×2内核的去卷积，但是上升显着超过了它。在本质上，由于我们的方法包括四个连续的上采样步骤（每块2x分辨率），当与FC变体相比时，它可以在输出中保留更多的结构（见图4）

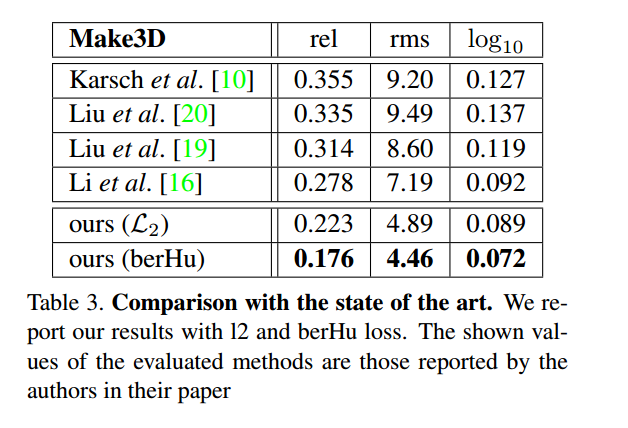
在所有显示的实验中，berHu损失优于L2。 相对误差的差异较大，可以通过L1（berHu）对L2小的残差较大的梯度来解释; 对相对误差的影响较大，因为距离较小的像素对较小的误差较敏感。这种效果也很好地看出，在具有挑战性的δ1测量中有更强的增益。

最后，我们测量单个图像的单个上卷积块（1.5 ms）的时序，并与我们的up-projection （0.14 ms）进行比较。 这超过了理论上的4速度，这是因为较小的滤波器尺寸从cuDNN内的线性池中获益更多。 此外，我们的模型的优点之一是总体计算时间。在我们的设置中，使用建议的上采样（上卷积78ms）预测单个图像的深度图仅为55ms。 这使得能够实时地处理图像，例如来自网络摄像机。当批量处理几张图像时，可以进一步加快速度。批量大小为16，每个图像的14ms up-projection 和28ms上升卷积。

相关方法(**Comparison with related methods**) 在表2中，我们将建议的结构与相关工作报告的结果进行比较。另外，如图4我们使用提出的方法（ResNet-UpProj）与不同变体（AlexNet，VGG，ResNet-FC-64x48）的估计深度图的准确性进行定性比较，以及公众对Eigen和Fergus的预测[ 5]。可以清楚地看到从AlexNet到ResNet的质量有所改善，尽管ResNet的完全连接变体的准确性提高仍然限于粗略的预测。提出的全卷积模型大大提高了预测深度图中的边缘质量和结构定义。







有趣的是，我们的深度预测表现出值得注意的视觉质量，即使它们是由单一模型导出的，经过端对端的训练，无需任何额外的后处理步骤，例如CRF推断[16,37]。另一方面，[5]通过结合RGB图像和原始预测的多尺度结构来改进他们的预测，以创建视觉吸引人的结果。然而，即使在真实值（最后一行）中没有实际的深度边界，它们有时也会误认为全局尺度（第二和第三行）或在原始图像中高纹理区域的情况下引入噪声。此外，我们与[5]中的参数数量进行比较，我们计算出的三个尺度为2.18亿，比我们的模型大约3：5倍。相反，这里提出的CNN结构是考虑到可行性的;参数的数量不应在高维问题中不可控制地增加。这进一步意味着减少所需的梯度步骤数量以及训练所需的数据样本。我们的单一网络更好地概括了，成功地解决了以前的CNN深度估计方法遇到的粗糙度问题。

### 4.3 Make3D数据集

此外，我们评估了我们的Make3D数据集[30]的户外场景模型。它由400个训练和134个测试图像组成，使用自定义3D扫描仪进行收集。由于数据集的获取可追溯到数年前，地面真相深度图分辨率限制在305×55，与原始的1704×2272像素的RGB图像不同。随着[20]，我们将所有图像的大小调整为345×460，由于结构和硬件的局限性，导致RGB输入到网络的分辨率进一步降低了一半。我们使用最佳性能模型（ResNet-UpProj），对30个时期的批量大小为16张图像，对大约15k个样本的增强数据集进行训练。当使用berHu损失时，开始学习率为0.01，但是在使用L2优化时需要从0.005开始更仔细的调整。动量为0.9。请注意，由于数据集的限制，考虑到低分辨率地面实况和长距离不准确（例如，映射到80米处的天空像素），我们通过掩蔽距离超过70米的像素来训练地面真相深度图。

为了将我们的结果与状态进行比较，我们使用双线性插值将预测的深度图提取到345×460。 表3根据（C1）标准报告了与以前工作相比的错误，在[20]所建议的深度小于70m的区域计算，并且由我们的训练所暗示。 除此之外，[20]使用每像素天空分类预处理图像，也将其排除在训练之外。 我们的方法在用L2或berHu损失功能训练时，显着优于以前的所有作品。 在这个具有挑战性的数据集中，berHu损失的优势更为显着。 同样类似于纽约大学，由于靠近深度值的权重，berHu改善了相对误差，而不是均方根。 该数据集的定性结果如图5所示。

### 4.4 应用于SLAM

为了补充以前的结果，我们展示了SLAM应用程序中深度预测的有用性，目的是重建3D环境的几何。特别地，我们部署一个SLAM框架，其中通过高斯 - 牛顿优化获得帧间跟踪，如[12]中提出的在连续帧对上计算的像素强度差异，同时融合当前帧与全局模型之间的深度测量通过点融合进行[11]。我们希望指出，据我们所知，这是基于单一图像深度预测的SLAM重建的第一个示范。

使用提出的ResNet-UpProj结构估计的深度值获得的SLAM重建之间的定性比较与使用纽约大学深度数据集的一部分序列上的地面真值深度值获得的SLAM重建之间的定性比较，如图6所示。该图还包括与使用AlexNet和VGG结构获得的深度预测进行比较。可以看出，深度预测的精度提高以及我们的上采样方法良好的保护质量不仅在图4的定性结果中显而易见，而且还产生了比其更准确的SLAM重建其他结构。我们希望指出，尽管我们不认为它的准确性与通过利用SfM和单眼SLAM等深度估计的时间一致性的方法相比，我们的方法并不明确依赖于视觉特征来估计深度，因此具有将其应用于通常存在于室内环境中的诸如墙壁，地板和其它结构的低纹理表面的场景上。虽然本文的范围显然不在本文的范围内，但我们认为这些方面足够有用，值得今后的分析。

五、 总结

在这项工作中，我们提出了一种从单一图像中深入估计问题的新方法。与典型的CNN方法不同，需要一个多步骤过程才能完善其原始的粗略深度预测，我们的方法是一个功能强大的单规模CNN结构，可以进行残差学习。所提出的网络是全卷积的，包括允许训练更深层配置的up-projection 层，同时大大减少要学习的参数数量和所需的训练样本数量。此外，我们说明了一个更快更有效的上卷积层的方法。不仅通过优化典型的l2损失，而且对berHu损失函数进行优化，对不同结构部件进行了全面的评估，表明它更适合于真实值深度图的底层值分布。总而言之，来自我们贡献的模型不仅比现有的方法更简单，可以在较少的时间内用较少的数据进行训练，而且还实现了更高质量的结果，使我们的方法在两个基准数据集中达到最先进用于深度估计。

参考文献

1. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua,and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274–2282, 2012. 2
2. V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab. Robust optimization for deep regression. *Proceedings of the International Conference on Computer Vision*, 2015. 6
3. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. Conf.Computer Vision and Pattern Recognition (CVPR)*,pages 886–893, 2005. 2
4. A. Dosovitskiy, J. Tobias Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1538–1546, 2015. 4
5. D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multiscale convolutional architecture. In *Proc. Int. Conf.Computer Vision (ICCV)*, 2015. 1, 2, 3, 4, 6, 7, 8, 9
6. D. Eigen, C. Puhrsch, and R. Fergus. Prediction from a single image using a multi-scale deep network. In *Proc. Conf. Neural Information Processing Systems (NIPS)*, 2014. 2, 3, 4, 6, 7
7. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprintarXiv:1512.03385*, 2015. 2, 4, 5
8. D. Hoiem, A. Efros, M. Hebert, et al. Geometric context from a single image. In *Computer Vision,2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 654–661. IEEE, 2005.2
9. S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of The 32nd International Conference on Machine Learning*,pages 448–456, 2015. 4
10. K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In *Proc. Europ. Conf. Computer Vision (ECCV)*, pages 775–788, 2012. 2, 3, 7, 9
11. M. Keller, D. Lefloch, M. Lambers, S. Izadi,T. Weyrich, and A. Kolb. Real-time 3d reconstruction in dynamic scenes using point-based fusion. In *Proc. Int. Conf. 3D Vision(3DV)*, pages 1–8, 2013.10
12. C. Kerl, J. Sturm, and D. Cremers. Robust odometry estimation for RGB-D cameras. In *Proc. Int.Conf. on Robotics and Automation (ICRA)*, 2013.10
13. J. Konrad, M. Wang, and P. Ishwar. 2d-to-3d image conversion by learning depth from examples. In *Proc. Conf. Computer Vision and Pattern Recognition Workshops*, pages 16–22, 2012. 2, 3
14. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3, 4
15. L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 89–96, 2014. 2, 7
16. B. Li, C. Shen, Y. Dai, A. V. den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*,pages 1119–1127, 2015. 2, 3, 7, 9
17. B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1253–  
    1260. IEEE, 2010. 2
18. C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):978–994, 2011. 2, 3
19. F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 5162–5170, 2015. 2, 3, 7, 9
20. M. Liu, M. Salzmann, and X. He. Discretecontinuous depth estimation from a single image.In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 716–723, 2014. 2, 3, 7, 9, 10
21. J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440,2015. 4
22. R. Memisevic and C. Conrad. Stereopsis via deep learning. In *NIPS Workshop on Deep Learning*,volume 1, 2011. 2
23. P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 1, 2, 4, 6,7
24. A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. Journal of Computer Vision (IJCV)*,pages 145–175, 2014. 2
25. A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007. 6
26. 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. 1
27. A. Roy and S. Todorovic. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR (CVPR)*, 2016. 2, 3, 6,7
28. O. Russakovsky, J. Deng, H. Su, J. Krause,S. Satheesh, S. Ma, Z. Huang, A. Karpathy,A. Khosla, M. Bernstein, A. C. Berg, and L. FeiFei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 3, 6
29. A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems*, pages1161–1168, 2005. 2
30. A. Saxena, M. Sun, and A. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Analysis and Machine Intelligence(PAMI)*, 12(5):824–840, 2009. 2, 6, 10
31. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 4
32. F. H. Sinz, J. Q. Candela, G. H. Bakır, C. E. Rasmussen, and M. O. Franz. Learning depth from stereo. In *Pattern Recognition*, pages 245–252. Springer, 2004. 2
33. S. Suwajanakorn and C. Hernandez. Depth from focus with your mobile phone. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1, 2
34. R. Szeliski. Structure from motion. In *Computer Vision*, Texts in Computer Science, pages 303–334. Springer London, 2011. 1, 2
35. J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Computer Vision and Pattern Recognition (CVPR),  
    2012 IEEE Conference on*, pages 103–110. IEEE, 2012. 1
36. A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. 2015. 6
37. P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2809, 2015. 2, 3, 7, 9
38. M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2018– 2025. IEEE, 2011. 4
39. R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape-from-shading: a survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(8):690–706, 1999. 1, 2
40. L. Zwald and S. Lambert-Lacroix. The berhu penalty and the grouped effect. *arXiv preprint arXiv:1207.6868*, 2012. 2, 6