指导教师：杨涛　　提交时间：2015-3-29

The task of Digital Image Processing

数字图像处理

School of Computer Science

No: 01

姓名：霍高峰

学号：2012302524

班号：10011206

远景取象

摘要

当前最先进的对单视图的深度估计和语义分割的想法和远景的几何属性紧密联系，即可感知的对象规模大小与距离成反比。

在这篇文章里，我们证明了我们可以使用这个属性以减少逐像素深度分类地学习一个像素是在任意固定典型深度的唯一可能性的一个更简单的分类预测。这一对任何其他的深度的可能性可通过施加同一分级后适当的图像的操作而获得。对问题所述规范的改变，可通过诸如删除某些由深度和透视效果引起的训练数据偏差。该方法可以直接推广到多个语义类，通过直接针对独立方法的弱点同时改善深度估计和语义分割性能。调节深度语义标签提供了一种方法使数据对齐到他们的物理尺度，使之成为更具辨识率的分类器。在语义类方面调节深度有助于分类的另有病态问题的模糊度之间进行区分。

我们在 KITTI 道路场景数据集和 NYU2 室内数据集测试了我们的算法，得到的结果是在单视点深度和语义细分领域显著优于目前的先进设备。

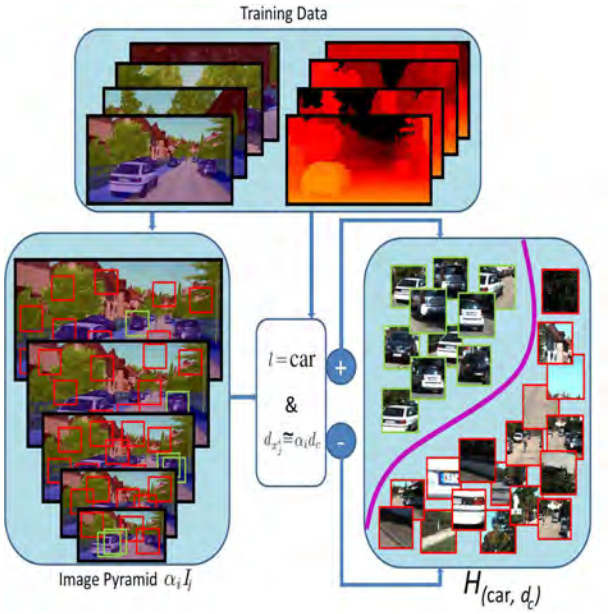

图 1 .深度语义分类训练过程示意图。每个语义类正面训练样本，预计将使用地面实况深度规范深度 dc，针对其他语义类和针对同一类样本，通过训练投影到其他规范的深度。这样的一个分类器能够通过施加适当的图像转换，以预测一个语义类和任何深度。

1.引言

一个单一的 RGB 图像的深度估计一直没有成为计算机视觉的详尽研究的问题，主要是由于它的难度，缺少数据以及一般明显不适定性的问题。然而，人类仍然可以自如执行此任务，这表明，逐像素深度被编码在所观察到的特征，并且可以直接从数据[20，1，10，16]获知。这样的方法典型地预测深度，取向，或适合使用标准对象识别管道的超像素的平面，由密集或稀疏特点的计算，建立丰富的功能表示，如词袋，以及对他们训练分类或回归的程序。一个分类或回归的响应被组合在一个概率框架，并在非常强的几何先验分布最可能的场景的布局进行估计。这个过程是完全数据驱动的，并且不利用立体几何的已知性质，最重要的是，该物体从投影中心的逆距离（深度）的感知大小缩放。这导致训练集中的深度的分布严重偏差;如果一个类似对象尚未见于在训练阶段相同的深度，它是无法估计的对象的深度。这些算法的缺憾可以通过抖动或很简单的加权数据样本被部分解决，但是，经过训练的分类器仍然不会在本质上无偏差的。

针对数据驱动的深度估计一个典型的说法是，要成功执行此任务，我们需要能够认识和了解场景。因此，人们应该等待，直到足够好的识别方法的开发。对于一些识别任务，这已经是如此。计算机视觉和机器学习研究进展导致算法的发展[15，27，29，28]，是能够成功地分类图像转换成数百个[4]甚至数千[3]不同的对象类的。进一步研究显示，这些方法的成功在于问题是如何约束的；所关注的对象通常缩放到图像的大小，并在此设置精心设计的特征的表示变得更加判别。基于这一观察，很明显，这限制了对计算机视觉任务算法性能的魔鬼，在于数据由于立体几何形状的规模不对。对标准的语义分类进行培训，以辨别语义类之间稳健的规模变化。所需性能之间的这种不协调使得学习产生不必要的困难。用于物体检测的问题[2，5，27]的物体的变化的二维尺度典型地通过缩放边界框紧密包围对象的大小相同，并在这之后建立的特征向量为每个单独的边界框的内容处理改造。如果没有这关键的一步，检测方法的性能会急剧下降。假使场景的几何形状是已知的，或者它可以可靠地估计，边界框的位置可以被约束为在特定的位置，如在地平面上[11]。然而，这些方法只能对具有特定空间范围，形状和大小

的前景对象使用，“物品”。对于有背景“东西”，如道路，建筑物，草或树的语义分割的任务，这样的做法是不合适的。如缩放周边草坪，以同样大小的边界框不会使具有代表性特征更容易辨别。但是，仍然存在规模为被绑在真实世界的物理尺寸的东西类的概念。草叶片的物理尺寸，建筑物的一个窗口或树叶的变化相比草坪、建筑物、或是一棵树要小得多。因此，最合适的校准，同时适用于事物和东西，是归一化到相同的物理尺寸。在所述深度是已知的情况下，已经可以被Kinect相机识别。归一相对于所测量深度[22]的分类功能通常执行得显然更好。

语义类视觉外观及其几何深度的相互依存关系表明，语义分割和深度估计的问题，应该共同解决。它已被证明[16]该调节深度语义分割在两阶段算法导致性能显著改进。立体声和多视点图像[14，10]已经表明，联合语义分割及三维重建比单独执行每个任务有一个更好的结果。在这些方法中，利用相互信息量的相当弱的源是高度[14]或表面法线不同的语义类别的[9]的分布。

在本文中，我们表明，使用该透视几何的性质，可以减少一个逐像素深度分类的学习，以更简单的分类预测的像素作为在任意固定的典型深度只的可能性。相似性对任何其他的深度可通过施加同一分级后适当的图像的操作而获得。问题的典型深度的这种转化去除对某些深度训练数据偏差和立体的效果。该方法可以被直向前推广到多个语义类，通过直接定位的独立的方法的缺点同时改善深度估计和语义分割性能。调节的深度语义标签提供给数据对齐到它们的物理尺寸的方式，以及调节深度上的语义类有助于分类的其他不合适问题的模糊度之间进行区分。

我们在非常受限的街道场景 KITTI 数据集[6]和非常具有挑战性的 NYU2 室内集[25]，其中没有关于场景的布局设想可以进行实验。我们的算法显著优于独立的深度估计和语义细分方法，并使用全 RGB-D 的数据与方法语义细分领域获得类似的结果。我们的逐像素分类器可直接放入任何竞争识别或深度估计的框架，以进一步改善结果；无论是作为一元潜力 CRF 的识别方法[13]或作为预测装配到超象素的飞机[20]。

2.无偏深度分类器

首先，我们定义符号。设 $\boldsymbol{I}$ 是一个图像， $\alpha * \boldsymbol{I}$ 是 $\boldsymbol{I}$ 由一个因子 α 几何缩放。设

$W^{w,h}(I,x)$ 是大小为 w*h 倍，中心在 x 点的图像 ***I*** 的子窗口。任何平移不变分级 ***HD（x）*** 中，预测像素 ***x*** 的处于深度 ***d*∈*2 D*** 具有任意大的固定尺寸的函数 ***w*h*** 中心在点 ***x*** 的子窗口：

Hd(x) := Hd(Ww,h(I, x)).

立体几何形状的特征在于：对象有从突出部的观察者的中心反距离缩放从的功能。因此，对于任何无偏深度分类器 ***HD（x）*** 的任何像素 ***x*∈*I*** 的深度 ***d*** 的可能性我应该是相同的深度 ***d/ α*** 的可能性的缩放图像的相应像素 ***α *I***:

$$H_d(W^{w,h}(I,x)) = H_{d/\alpha}(W^{w,h}(\alpha * I, \alpha x)).\quad (2)$$

该属性对保持分类器健壮性至关重要，针对总是波动在训练数据中的存在于小和平均大小的数据集。它似乎简单，但它尚未在任何先前的数据驱动的深度估计方法[20，1，16]中。

该属性所暗示的，该深度分类可以减少到像素 ***x*∈*I*** 是否处于任何任意固定典型深度 d_c 直流简单得多的预测。可以通过一个因子 d/d_c 作为施加相同分类器 h_{dc} 到适当缩放的图像来获得的任何其他深度 ***d*** 的反应：

$$H_d(W^{w,h}(I,x)) = H_{dc}(W^{w,h}(\frac{d}{d_c} * I, \frac{d}{d_c} x)).$$

因此，深度估计的问题被转换为估计该转换（缩放比例），这将突出象素到规范化深度。分类的特殊形式直接意味着应该如何训练。在训练阶段，分类器应该学会辨别转变，从改造到深处较规范的深度等训练样本的典型深度的训练样本。详情在很大程度上取决于所选择的框架;例如在分类框架的问题将被视为一个标准的 2-lable 正片 VS 底片的问题，在排名框架变换为正则深度应大的训练样本的响应（由足够的余量，如果适当）比对于反应样品转化为任何其它比典型深度。我们的分类器有超过从像素的特征的表示的深度直接学习几个优点。首先，来预测某一物体（例如，汽车），在一定的深度 d 不要求一个类似的对象可见于在训练阶段相同的深度。第二，我们的分类器不具有始终是存在于一个多级分类或回归不平衡的训练数据的问题。直观地说，越接近由多个点和一些在训练数据中可能只是经常出现在一定的深度对象组成的物体。多级分类器或回归这些问题可部分通过使用该训练点的适当采样或重加权抖动中的数据来解决；然而，属性（2）

直接实施注定是一个更好，更原则性的解决方案。

3.语义深度分类

单一视图深度估计是普通的病态问题。如果深度分类的条件为语义标签，几个歧义可以潜在地被解决。训练分类器，应该是一方面区别语义类之间，但在其它健壮规模的变化，是不必要的努力。如果训练样本的规模对齐，这个问题就好办多了。最合适的取向，同时适用于事物和东西，是根据物理尺寸的归一化，而这正是和投影到规范深度（4）一样。深度分类到多个语义类的泛化可以通过学习联合分类直接完成。$H_{(l,d_c)}(W^{w,h}(I,x))$，预测是否一个像素 $\boldsymbol{x}$ 取一个语义标签 $l\in$ 和处于典型深度 $\boldsymbol{dc}$。通过应用（4），所述分类器用于任何其它深度 d 的响应是：

$$H_{(l,d)}(W^{w,h}(I,x))=H_{(l,d_c)}(W^{w,h}(\frac{d}{d_c}*I,\frac{d}{d_c}x))$$

我们的分类器被无偏朝一定深度的优点现在更加明显。$|\boldsymbol{D}||\boldsymbol{L}|$级分类器或学习的另一种方法深度回归量将需要非常大量训练数据，足以代表在语义和深度标签的横产物为每个标签的分布。在培训阶段，我们的分类应该学会区分每个类转化为规范的深度转化到深处较规范的深度等其他类和样本的训练样本的训练样本。转型为规范的深度，不能用于天空类（室外场景），在测试时间的深度将自动分配给 ∞。

4.实施细节

将每个训练样本独立地与特征结果围绕该窗口计算的规范化距离在实际计算上不可行。因此，我们在测试阶段将离散深度估计的问题转化为一组离散的标签 $d_i\in D$。预计一个基于对象的规模预测的误差，以与距离呈线性增长，这表明相邻深度 d_i 和 d_i+1 应该有一个固定的比率 $\frac{d_i+1}{d_i}$，这取决于期望的精度选择。这使我们能够将问题转化为分类过的图片的棱锥 $\alpha_i*I=\frac{d_i}{d_c}*I$ 对每个培训或测试图像 I。

对于深度在 $\alpha_i d_c$ 的像素，通过图像的缩放比例 α_i 对应于变换到规范的深度。因此，在训练阶段，一个点锥形 $x_j^i\in(\alpha_i*I)$ 图像的基于用作正面或反面样品上是

如何接近的相应的像素中的原非缩放图像 $d_{x_j^i} = d_{(x_j/\alpha_i)}$ 地面实况深度与深度 $\alpha_i d_c$。如果它们的比接近 1，例如

$$\max(\frac{d_{x_j^i}}{\alpha_i d_c}, \frac{\alpha_i d_c}{d_{x_j^i}}) < \delta_{POS}$$

像素 x_j 被用作正对相应语义类和负对所有其他类。如果它们是十分不同，例如

$$\max(\frac{d_{x_j^i}}{\alpha_i d_c}, \frac{\alpha_i d_c}{d_{x_j^i}}) > \delta_{NEG}$$

缩放由 α_i 未样本变换到规范深度 *dc* 并因此被用作负对所有语义类。

在培训阶段，同真实世界的大小和形状的每个对象无论多远，他们都应该对所学分级有相同的影响。因此，将样品从取样 $\alpha_i * I$ 用相同的子采样为所有 α_i 和是如果满足（5）或（6）分别对应的限制它们用作正或负。

转化问题到规范化深度与基于其真实世界的物理尺寸的数据保持一致，这可能是对于不同的语义类的显著不同。因此，最适合的分类器是基于上下文的具有自动学习上下文大小，如[24，23,13]。继多特征[13]扩展的 TextonBoost 方法[24]，该密集特征，即纹理基元[18]，SIFT[17]，局部量化三元的格局[12]和自相似特征[21]，是提取在每个图像中的棱锥的每个像素 $\alpha * I$。每个特征被集群化到使用 k 均值聚类 512 视觉词并且每个像素的软加权为 8 个最近邻使用基于距离的加权[7]具有指数内核计算的。为一个窗口 $W^{w,h}(\alpha_i * I, x_j^i)$ 中的特征表示由词袋软加权表示在其固定的随机组 200 的子窗口的串联的，如[24]。多类增强分类器通过学习可组成一个决策树，比拟 $\theta \in T$ 阈内一维特征点。不像在[24]中，该组的阈值 T 被独立地由在训练集中均匀地分割其范围发现针对每一特定尺寸。长的特征向量（512×200 针对每个特征）针对各像素不能保持在内存中，但还可以计算使用积分图像[24]在每个棱锥图像和特征的每一个视觉词 fly。我们实施的几个启发式，以减少所产生的内存需求。从（0,1）的软权重是使用 1 个字节的近似。基于该要求的范围为每个单独的视觉词的每个图像 （1-8 字节）的积分图像仅用于一个子窗口的图像的，其覆盖给定视觉词的所有功能，使用整数型建造。分类器多重功能，在[13]中，融合使用后期的融合，例如不同该分类的每个功能都是独

立的培训和最终的平均值。由于这些启发式，对内存的需求下降约 40 倍的 NYU2 数据集[25]下面 32GB。一个大约 10×下降是由于积分图像的更新和 4×由于后期融合。

5.实验

我们对 KITTI[6]和 NYU2[25]的数据集测试了我们的算法。该 KITTI 数据集被选择来证明我们的分类器的学习深度为语义类具有相对小的数目的训练样本的能力。该 NYU2 数据集用于显示深度可以预测为一种无约束问题，没有关于场景的布局的假设。此外，我们表明，对于两个数据集学习问题的共同导致性能的改善。

5.1 KITTI 数据集

该 KITTI 数据集[6]是由大量的分辨率室外街景图片 1241×376，其中有 194 图像包含稀疏的视差图由 Velodyne 激光扫描仪获得。我们标记语义分割地面实况的 60 张图片地面真理的深度和他们分成 30 个培训和 30 个测试图像。标签集包括 12 类语义标签（参照表 3）。对三个语义类有高的变化（自行车，人及符号）的评价，由于训练数据不足（在训练集中只有 2 实例）被忽略了。我们的目的是识别深度在 2-50 米的范围内具有最大相对误差 $\delta = \max(\frac{d_{gt}}{d_{res}}, \frac{d_{res}}{d_{gt}}) < 1.25$，其中 d_{res} 是估计的深度和 d_{gt} 的地面实况深度。因此，我们设置 $\delta_{POS} = 1.25$。对人视觉识别深度较高的精度也是非常困难的。规范深度被设定为 20 米。训练样本被视为负的，如果他们的误差超过 $\delta_{NEG} = 2.5$。训练样本 δ_{NEG} 与和 δ_{POS} 之间错误被忽略了。与最先进的设备进行定量比较，一维分类在表 1-3 中给出。所述 Make3D[20]的定量比较，培养具有相同的数据，用我们的深度只和联合深度语义分类器在表 2 给出。我们共同的分类显著优于竞争算法在这两个领域。定性结果列于图 3。只考虑深度和联合分类定性比较给出在图 4 中。估计的深度的相对误差的分布在图 6 中给出。

5.2 NYU2 数据集

该 NYU2 数据集[25]组成的分辨率为 640×480 的 1449 室内图像，包含地面实况语义分割与由 Kinect 的传感器获得 894 语义类和逐像素的深度。对于该实验，我们使用了 40 语义类的子集，如在[19，25，8]。图像被分成 725 训练和 724

的测试图像。我们的目的是识别深度范围为 1.2 - 14 米的最大相对误差 $\delta = \max(\frac{d_{gt}}{d_{res}}, \frac{d_{res}}{d_{gt}}) < 1.25$。规范深度被设定到 6.9 米。训练样本被视为正，如果他们的误差小于 $\delta_{POS} = 1.25$，如果他们的误差超过 $\delta_{NEG} = 2.5$，则为负。定量比较的类仅一元的分类[13]和国家的最先进的在测试时间同时使用 RGB 和深度算法给出在表 1。我们的分类器的性能是可比于这些方法，即使在这些不公平的条件。在深度域中的定量结果可以在表 2 中找到。还有我们可以比任何其他的分类;所有其他的方法只限制为非常特定的场景具有较强布局约束，如道路场景。主要是由于显著更窄的范围内的深度，NYU2 数据集性能比 KITTI 数据集的高。

6.讨论和结论

在本文中，我们提出了一个新的逐像素的分类，可以共同从单个图像预测语义类和深度标签。在一般情况下，所获得的结果看起来非常有前途的。在深度域中的问题分解成一个深度分类的减少仅翻起是非常强大的，由于其固有的处理的数据集的偏差。在语义分割域，我们发现正确对准的重要性，导致定量更好的结果。该方法的主要缺点是无法处理低分辨率图像，以硬件和显然无法更精确地定位在对象的语义类具有高方差方面非常大的需求。得到的结果表明在未来三个方向的工作：使用法线估计，估计深度为潜变量的过程中与少量的图像（或无）与地面实况深度的数据集培训的方向进一步调整，以及不同形式的正则化发展适合的问题。单纯使用标准 Potts 模型成对电位没有导致一种改进，因为这种形式的正规化通常被去除所有的小远处的物体，因此，我们在文件内省略了这些结果。然而，我们的联合深度/语义分类器有可能使代表不同类的对象之间的预期空间关系更富有成对的电位。

RGB methods		RGB-D methods		
Class-only [13]	Joint classifier	[19]	[25]	[8]
34.85	37.11	38.23	37.64	45.29

表 1 在频率加权路口 VS 联合的措施在 40 级 NYU2 数据集的定量比较。我们的逐像素的方法优于使用相同的功能基线非规模自适应方法，并获得可比较的结果到使用过程中的测试时间全 RGB-D 数据的方法。

	KITTI			NYU2	
	Make3D	Depth-only	Joint	Depth-only	Joint
$\delta < 1.25$	26.21%	43.83%	47.00%	44.35%	54.22%
$\delta < 1.25^2$	48.24%	68.01%	72.09%	70.82%	82.90%
$\delta < 1.25^3$	64.20%	82.19%	85.35%	85.90%	94.09%

表 2Make3D 的定量比较[20]，我们的深度仅分类器和分类器联合上 KITTI 和 NYU2 数据集中的像素的比率正确标记，这取决于允许相对误差的最大值 $\delta = \max(\frac{d_{gt}}{d_{res}}, \frac{d_{res}}{d_{gt}})$，其中 d_{res} 是估计的深度和 d_{gt} 的地面实况深度。

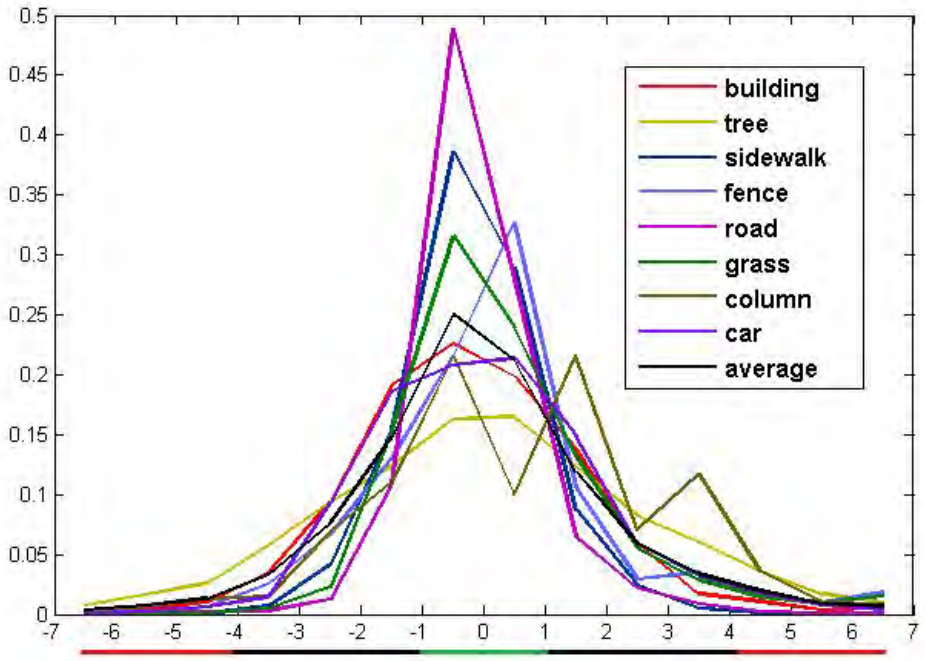

图 2 估计的深度的相对误差的分布给出一个语义标签，KITTI 数据集的日志空间与基 1.25 为每个语义类。平均分配的计算不包括天空标签。下面的 x 轴的绿色和红色线表示，在其中的时间间隔分别训练样本被用作正和负。

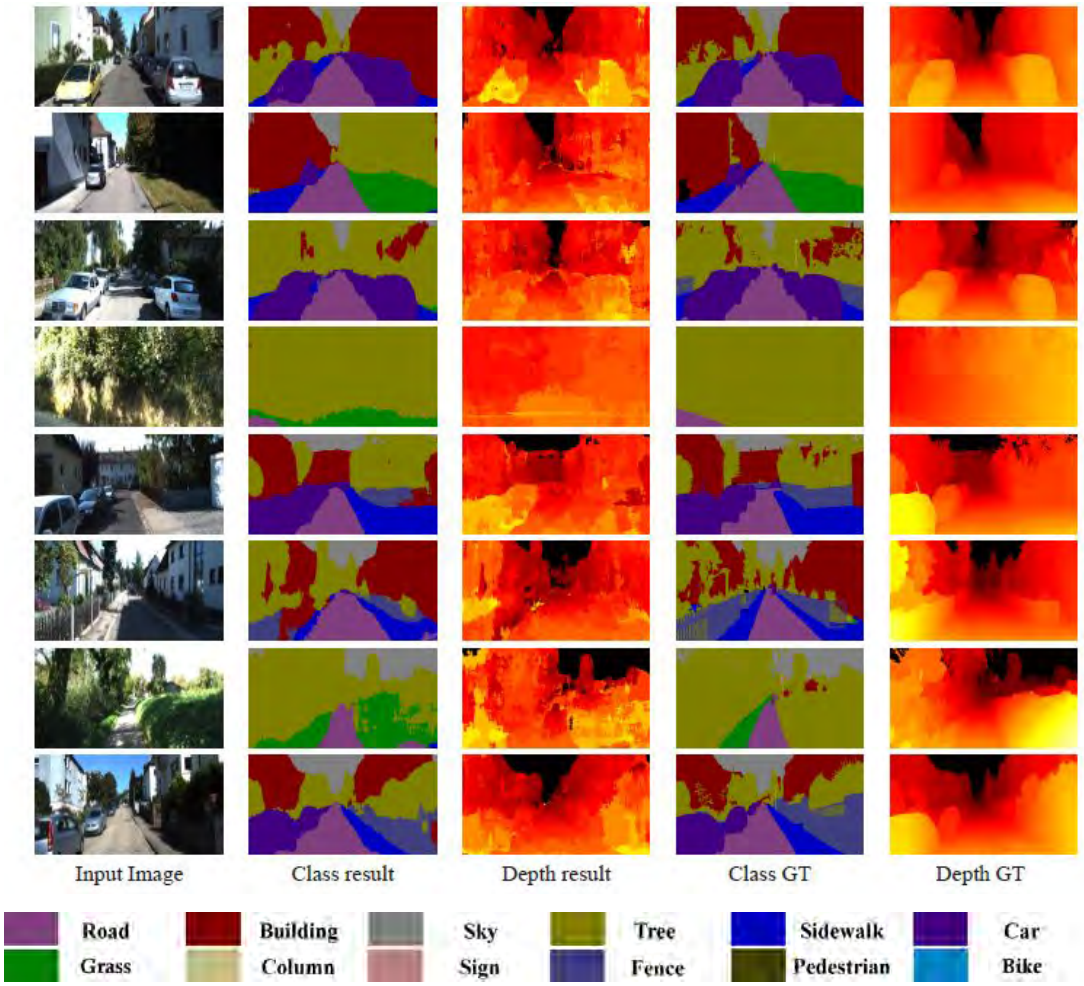

图 3 我们的分类器在测试集的逐像素的结果。注意，关于场景的结构，它不需要额外的假设。每个像素单独地分类基于周围不使用 2D 位置的像素的窗口中。最 mislabellings 是含糊的双类，例如草 VS 树（图片 7），人行道路 VS（6），建筑 VS 栅栏（3）。注意，分类器的噪取决于深度和远处的物体被以更高的精度确认，不像一类分类器。

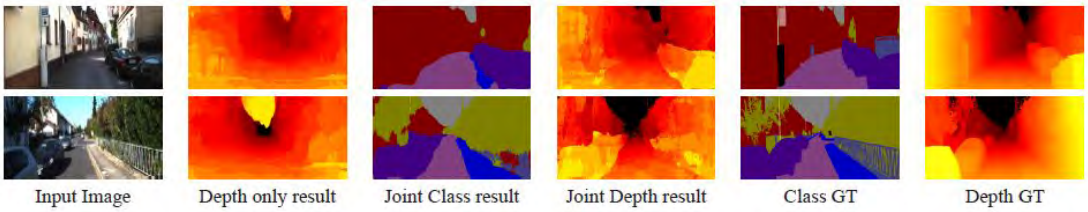

图 4 单一视图视差估计比较不具有和具有语义类。联想上图的语义类中。深度只分级趋于正确估计只为统治阶级（建筑，道路，树木）的深度，而忽略较小的类，如汽车或列/极。深度分类器只有不能处理的天空类。

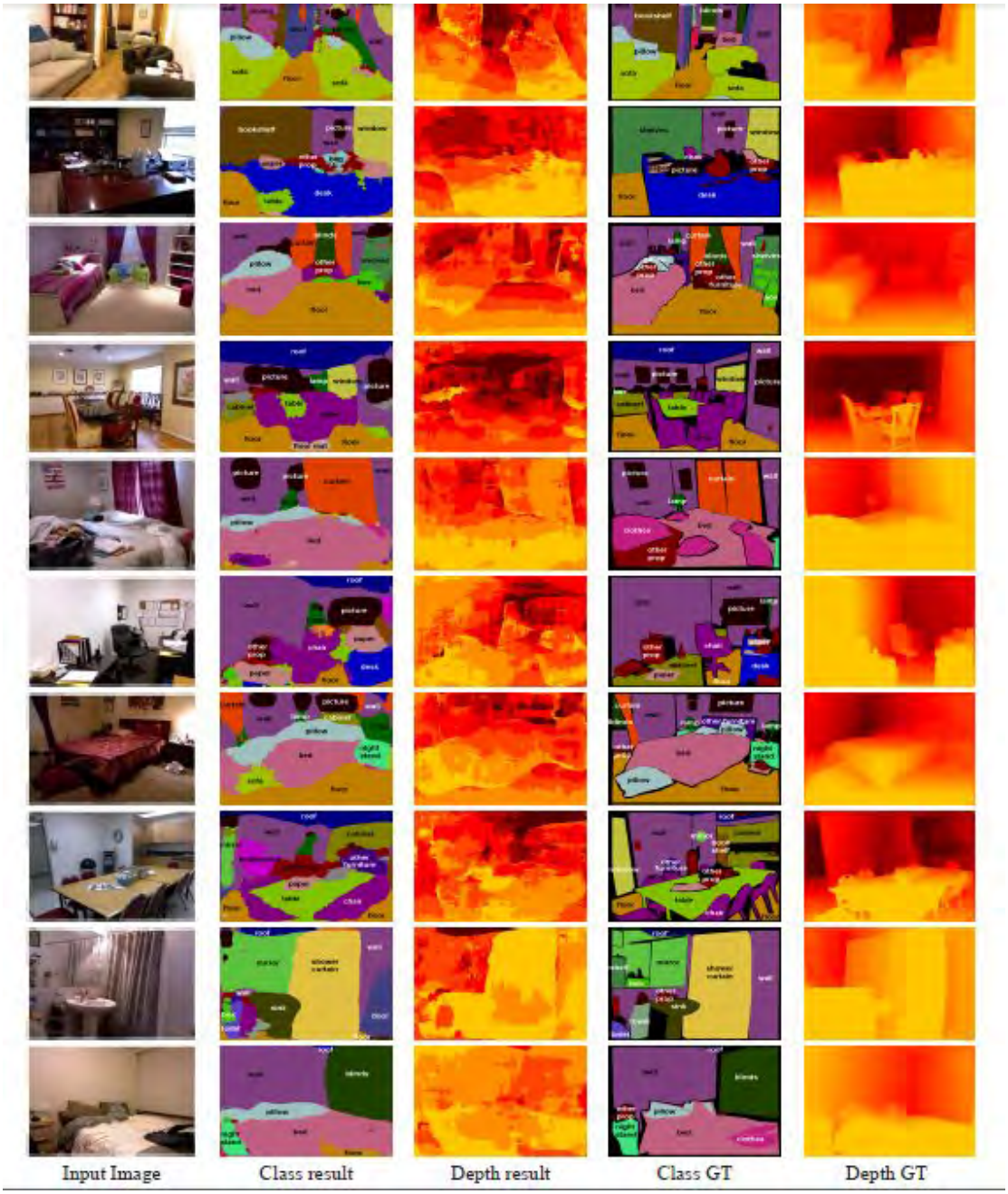

图 5 在 NYU2 数据集的一元分类的定性比较。我们的分类通常识别场景的两个语义细分和深度域的要点。然而，分类是基于上下文的，往往对象不局限准确，边界不匹配的地面实况。典型错误分类为视觉上相似的类，例如书桌 VS 表（图像 2），地板 VS 地板垫（4），窗口 VS 镜（8）之间;由于数据集的不一致，例如书架打成货架（2）或书籍（1）;或类没有严格限定，例如其他支柱（5）或其它家具（7）。

	Global	Average	Building	Tree	Sky	Sidewalk	Fence	Road	Grass	Column	Car
Class-only classifier [13]	80.2	66.2	87.0	82.8	89.7	68.4	31.6	**84.8**	61.2	7.3	83.2
Joint classifier	**82.4**	**72.2**	**87.2**	**84.6**	**91.6**	**76.5**	**39.4**	83.2	**69.9**	**28.5**	**88.9**

表 3 定量结果的召回措施为吉滴的数据集。使用相同的功能，我们的规模调整分类跑赢类唯一的基线分类[13]。得到的稀有类和来自照相机，属于改进远的对象象素。

参考文献

[1] O. Barinova, V. Konushin, A. Yakubenko, K. Lee, H. Lim, and A. Konushin. Fast automatic single-view 3-d reconstruction of urban scenes. In *European Conference on Computer Vision*, 2008. 1, 3

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Conference on Computer Vision and Pattern Recognition*, 2005. 2

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2011. 2

[4] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples an incremental bayesian approach tested on 101 object categories. In *Workshop on GMBS*, 2004. 2

[5] P. F. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Conference on Computer Vision and Pattern Recognition*, 2008. 2

[6] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomousdriving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition*, 2012.2, 4

[7] J. C. V. Gemert, J. Geusebroek, C. J. Veenman, and A.W.M. Smeulders. Kernel codebooks for scene categorization. In *European Conference on Computer Vision*, 2008. 4

[8] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from RGB-D images. *Conference on Computer Vision and Pattern*

Recognition, 2013. 4, 5

[9] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D scene reconstruction and class segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2013. 2

[10] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In *International Conference on Computer Vision*, 2005. 1

[11] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *Conference on Computer Vision and Pattern Recognition*, 2006. 2

[12] S. u. Hussain and B. Triggs. Visual recognition using local quantized patterns. In *European Conference on Computer Vision*, 2012. 4

[13] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. In *International Conference on Computer Vision*, 2009. 2, 4, 5, 8

[14] L. Ladicky, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. H. S. Torr. Joint Optimization for Object Class Segmentation and Dense Stereo Reconstruction. *International Journal of Computer Vision*, 2012. 2

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Conference on Computer Vision and Pattern Recognition*, 2006. 2

[16] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *Conference on Computer Vision and Pattern Recognition*, 2010. 1, 2, 3

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. 4

[18] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 2001. 4

[19] X. Ren, L. Bo, and D. Fox. RGB-(D) scene labeling: Features and algorithms. In *Conference on Computer Vision and Pattern Recognition*, 2012. 4, 5

[20] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems*, 2005. 1, 2, 3, 4, 5

[21] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Conference on Computer Vision and Pattern Recognition*, 2007. 4

[22] J. Shotton, A. Fitzgibbon, M. Cook, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Conference on Computer Vision and Pattern Recognition*, 2011. 2

[23] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2008. 4

[24] J. Shotton, J. Winn, C. Rother, and A. Criminisi. *Texton- Boost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European Conference on Computer Vision*, 2006. 4

[25] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision*, 2012. 2, 4, 5

[26] A. Torralba, K. Murphy, and W. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Conference on Computer Vision and Pattern Recognition*, 2004. 4

[27] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *International Conference on Computer Vision*, 2009. 2

[28] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Conference on Computer Vision and Pattern Recognition*, 2010. 2

[29] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *Advances in Neural Information Processing Systems*, 2009. 2