指导教师：杨涛　　提交时间：2016/3/17

CVPR2015 Paper Translation

No: 01

姓名：陈烨斐

学号：2013302621

班号：10011306

基于单张图片的特定类别的物体重建

Abhishek Kar*, Shubham Tulsiani*, Joao Carreira and Jitendra Malik
University of California, Berkeley - Berkeley, CA 94720

摘要：

对于单张野外图像的目标重建这个问题，我们现在取得了一些进展并且得到了一些富有意义的数据，这篇论文的主要内容是介绍了一种以现实场景的像素为输入，多种刚性类别的三维表面为输出的自动化流水线。我们方法的核心是以带有噪声的目标自动分割为驱动，从现有的目标检测数据集中获得二维标注，从而生成可变的三维模型，并且我们添加了一个自下而上的模式来还原高频成形的细节。我们利用近期引入的 PASCAL 3D+ 数据集对这种方案进行了全面的定量分析和消融研究，最终在 PASCAL VOC 上呈现出了非常鼓舞人性的自动重建效果。

1. 介绍：

现在考虑图 1. 中的汽车，我们不但可以通过一眼判断出图像中含有一辆车，而且我们可以在脑海中勾勒出它丰富的内部结构，比如说它的位置和三维外形。此外，我们可能从未见过这辆车，但是我们对它的三维外形有一个大概的估计。我们能做到这些是因为我们没有关于这辆车空白状态图像的一些经验，而是在我们

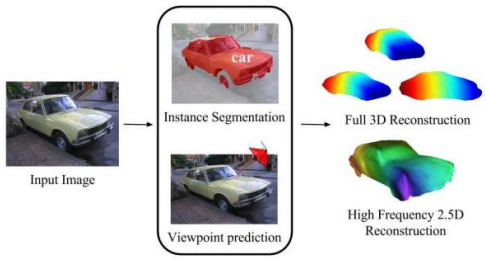

图 1：这就是利用我们的系统进行自动的基于单张图像的目标重建之后获得的结果，我们的方法利用了估计的实例分割，并且预测了视点来产生一个完整的三位网格表面和高频的 2.5 维的深度图。

“对于以前东西的记忆”中搜寻。以前看见的汽车让我们能够形成汽车三维外形的概念，并投射到这个特定的实例中。对于这个特定的实例，我们还可以使它表现得更具体化（例如它可能有的任何自定义的装饰），自上而下和自下而上的线索信号都影响着我们的感觉[26]。

在我们取得的进展中，关键的部分就是一个从以往的视觉体验中建立三维外表模型的机制。我们已经开发了一个算法，利用在现代计算机视觉数据集（例如 PASCAL VOC[15]）中的二维标注（分割编码和一小部分的关键点）的图像，可以建立特定类别的外表模型。然后可以用这些模型来指导对新的二维汽车图像进行自上而下的三维外形重建。我们在自上而

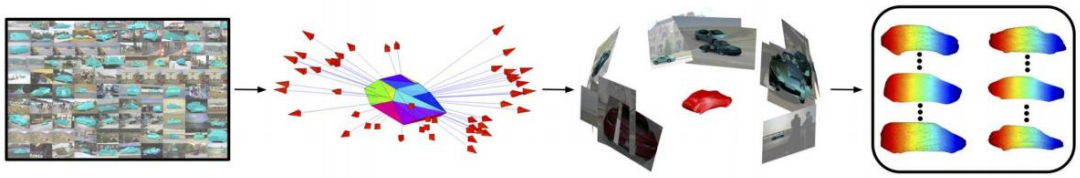

图 2: 训练模型的概览。我们使用了一个标注的图像采集来估计相机的视点，然后利用目标轮廓来训练三位外表模型。如最右边的图所示，我们的外表模型能够产生形变来捕捉类别内部的形状变化。

下的外形推理算法中补充了一个自下而上的模式，这样进一步改进了我们对于一个特定实例的外形估计。最后，在近年来快速发展的识别模块[2,11,17,20,34]（目标检测，分割和位姿估计）中，我们的模型具有较强的鲁棒性，当其应用于野外图像时，在图像输入之后，能够实现完全的自动化重建。

最近，Vicente et al.[36]中从相似标注重建三维模型的方法和我们的类似，但是也有一个不同的侧重点：它是以重建一个完全标注的图像集为目的，所以对于模型所适合的分割的品质，它指定了一个很强的假设，因此它不适合在一个不受约束的环境下进行重建。我们的方法可以在这样的环境下工作，部分原因是由于使用了明确的三维外表模型。我们的工作还与 Kemelmacher-Shlizerman et al.[23, 32]有关，这篇论文的旨在从二维图像的表面学习形变模型，而我们关注点在于无约束环境设置下还原出更丰富的外形，和降低重建过程中的分辨率开销。

在计算机视觉的历史中，基于单张图片的目标重建，已经在模型表示上反映出了不同的偏好。广义圆柱[27]导致了对于外形中的某些类别的非常紧凑的描述，并且可以被用于分类级别的描述，但是使用它解决一般外形的拟合问题还是比较有难度的。多面体模型[18, 40]，可以追溯到早期罗伯茨的工作[29]，以及 CAD 模型[25, 31]提供了外形的粗略估计值，并且给出了一组相关的点，这些点对于确定实例的视点是非常有效的。这里我们追求更具有表现力的基础外表模型[1, 7, 42]，模型在两个极端之间建立了一个平衡，因为它们会产生变形但是只是沿着特定类别的模式而变化。和以前的工作[42]相比，我们把它们适用到了自动的数字地面目标的分割中。

我们的论文的组织结构如下：在第 2 章节我们描述了我们的自动学习模型，在这其中我们用外表模型公式(2.2 节)估计了相机对于所有训练目标(2.1 节)的视点来建立三维模型。第 3 章节介绍了我们的管道测试，在

使用学习所得的模型重建新的实例时没有假设任何的标注。我们在第 4 章节评价了在多种设置下的重建效果，并且提供了野外重建的例子。

2. 建立可变的三维模型

我们对于可以稳健地对齐到那些含有噪声的目标分割的三维外表模型很感兴趣，这些模型合并了关于类外形映射到图像的自上而下的特定类别的知识。受助于匹配分割和一些关键点，类似于[36]，我们想从二维训练图像中建立这样的模型。我们的方法首先使用运动结构估计了一个类别中所有目标的视点，然后在一个基于形变的，具有代表性的三维外形上进行优化，它能能最好的解释视点中所有轮廓和条件的三维外形。模型学习的这两个阶段我们将在接下来的小节里进行描述。图 2 说明了我们的这个训练途径。

2.1. 视点估计

我们使用了 NRSfM[10]的框架对每一个类别中的所有训练实例估计了相机的视点（旋转，平移和缩放），最初提出从视频中恢复外形和形变的时候，遇到在较少的对应关系中进行视点估计的问题，NRSfM 是一个很自然的选择，但是如果没有明确的建模，内部类别变异可能成为一个混乱因素。但是这些算法的表现只在简单的类别中进行过探讨和实验，比如多用型汽车[41]或者花瓣和小丑鱼[28]。和我们的工作更接近的是 Hejrati and Ramanan[21]，它在一个很大的类别（汽车）中使用 NRSfM，但是需要一个预测探测器来填充缺失的数据（闭塞关键点），而我们在论文中并没有使用这样的假设。

我们遵循 Torresani et al.[33]中的 EM-PPCA 公式，并且对这个算法提出了一个简单的扩展，就是在建立关键点联系之外合并轮廓信息以便于稳健地恢复相机和基础外形。与我们比较相似的观点现在已经在 shape-from-silhouette literature 和 rigid structur-from-motion[36]中提出。但是据我们所知，和 NRSfM 没有联系。

NRSfM 模型. 对已每个实例 $n \in \{1, \cdots, N\}$ 给定 k 个关键点联系，与 NRSfM 算法[33]相比，我们的改变在于最大化以下模型的可能性：

$$\begin{aligned} &P_n = (I_k \otimes c_n R_n) S_n + T_n + N_n \\ &S_n = \bar{S} + V z_n \\ &z_n \sim \mathrm{N}(0, I), \quad N_n \sim \mathrm{N}(0, \sigma^2 I) \\ &\textit{subject to}: \quad R_n R_n^T = I_2 \end{aligned} \tag{1}$$

$$\sum_{k=1}^{k} C_n^{mask}(p_k, n) = 0, \quad \forall n \in \{1, \cdots, N\} \tag{2}$$

这里，P_n 是三维外形 S_n 的投影，其中含有白噪声 N_n，刚性变换由正交投影矩阵 R_n，尺度 c_n 和二维变换 T_n

给出。外形被参数化为一个高斯模型，其中参数有平均外形 $\bar{S}$ ， m 个基向量 $[V_1, V_2, \cdots, V_m] = V$ 和潜在的形变参数 z_n 。我们关键的修改在于约束（2），C_n^{mask} 表示第 n 个实例的二值模板的倒角距离区域，并且说明第 n 个实例的所有的关键点 $p_{k,n}$ 在其二值模板中。我们发现这可以从数据中学习得到更加精确的视点，以及更有意义的外形。

图三：NRSfM 视点估计：使用三位汽车线条框架进行可视化的视点估计

学习. 我们使用 EM 算法来使上述模型的相似性达到最大。在 E 步之后，缺失的数据（闭塞关键点）可以利用之前的等式来填入。对于每一个训练实例 n ，这个算法计算外形参数 $\{\bar{S}, V\}$ ，刚体变换 $\{c_n, R_n, T_n\}$ ，以及形变参数 $\{z_n\}$ 。在实践中，我们使用水平镜像图像增加了数据，以便于在所考虑的目标分类中利用双边对称。我们还预先计算了整个集合中的倒角距离场来加速计算。如图 3 所示，NRSfM 让我们可以可靠地预测视点，同时在遇到类内的变化时依然稳健。

2.2. 三维基础外表模型学习

在具备了整个训练集合中的投影镜头参数和关键点对应关系（通过 NRSfM 提升到三维）之后，我们开始在一个类别中依据目标轮廓建立形变三维外表模型。在校准设置下，利用从单个目标预测所得的多个轮廓，进行三维外形重建已经在学术界开张了广泛的研究。两个较为突出的方法是可视外壳[24]和从蛇派生出的变分法[14, 30]，它们迭代的改变表面匹配直到收敛。近期一些有趣的论文已经对类别处理[12, 13]扩展出了变分方法，但是一般需要一些三维标注格式来引导模型。最近提出的可视外壳方法[36]和我们一样，只需要二维标注来进行基于类别的重建，并且它已经 PASCAL VOC 上成功的演示，但是这并没有达到我们的目标，因为它对于分割的精确性做了很强的假设，并且实际上用三维像素层完全填满了每一个分割。

形状模型公式. 我们将类别外形建模成可变的点云 —— 每个点代表种类里的一个子类别。有一种潜在的直觉是一些外形类型的变化可能

会被某个参数模型很好的解释，如丰田和雷克萨斯等轿车。但是期望它能够在帆船和游轮之间建模是不合理的。这样的模型一般需要目标的组成部分和它们的空间布置[22]等知识，并且涉及难以优化的复杂公式。我们所做的并不是给类型中不同的子类型训练单独的现象外表模型，和 NRSfM 模型类似，我们使用了一个基础的线性组合来对这些形变建模。需要注意的是我们从轮廓中学习了这样的模型，并且这使我们能够学习可变的模型，而不依赖于扫描三维模板所得的点之间的对应关系[8]。

我们的外表模型 $M=(\bar{S},V)$ 包括了一个平均外形 $\bar{S}$ 和可变基向量 $V=\{V_1,\cdots,V_k\}$，这些是从一个训练集合 $T:\{(O_i,P_i)\}_1^N$ 中学习所得的，其中 O_i 是实例的轮廓，P_i 是从世界到图像坐标的映射函数。需要注意的是我们利用 NRSfM 得到的 P_i 与正投影相关，但是我们的算法也可以处理透视投影。

能量公式. 我们主要依据图像轮廓来制定目标函数。举个例子，一个实例的外形应该总是在它的轮廓内预测，并且应该与关键点（利用 NRSfM 提升到三维）一致。所以我们定义相应的能量项如下（这里 $P(S)$ 对应外形 S 的二维投影，C^{mask} 表示轮廓 O 的二值模板的倒角距离场，$\Delta^k(p;Q)$ 被定义为集合 Q 中，p 点到它最邻近的 k 个点的平方平均距离）

轮廓一致性. 轮廓一致性使一个实例的预测外形映射到它的轮廓内部。我们可以实现这一点，通过利用到轮廓的距离来隐藏那些映射到实例外部的点。在我们的符号 Δ 中可以被写成如下形式：

$$E_s(S,O,P)=\sum_{C^{mask}(p)>0}\Delta^1(p;O) \qquad (3)$$

轮廓覆盖. 单独使用轮廓一致性只能使映射到外面的点往轮廓里移动。但是这不能确保目标轮廓被充满，就是说可能会矫枉过正。我们的处理是利用一个能量项来引导轮廓上的点把映射到周围的点往它们拉。这可以被规范的表示为：

$$E_c(S,O,P)=\sum_{p\in O}\Delta^m(p;P(S)) \qquad (4)$$

关键点一致性. NRSfM 算法利用相机视点给我们提供了单独的三维关键点。我们使用训练集合中的这些稀少的对应关系来改变外形以适应这些三维点。对于每一个实例，相对应

的能量项对于外形和三维关键点 KP 之间的偏差进行约束。这可以被明确地写成：

$$E_{kp}(S,O,P)=\sum_{\kappa\in KP}\Delta^m(\kappa;S) \qquad (5)$$

局部一致性. 除了以上的数据项之外，我们使用一个简单的外形正则化来限制随意的形变，具体是在每个点和它的临近点之间强加一个二次的形变约束。我们也给形变加上了一个相似的约束来保证局部平整。参数 δ 表示相邻点的平均平方位移，并且它使所有的面都具有相似的大小。这里 V_{ki} 是第 k 个基础的第 i 个点。

$$E_l(\bar{S},V)=\sum_i\sum_{j\in N(i)}((\|\bar{S}_i-\bar{S}_j-\delta)^2+\sum_k\|V_{ki}-V_{kj}\|^2) \qquad (6)$$

标准平滑度. 自然世界中的产生的外形往往具有局部的光滑性。我们通过在外形中局部领域的变化的法线方向放置一个代价预先获得这一点。我们的标准平滑度能量被规定为：

$$E_n(S)=\sum_i\sum_{j\in N(i)}(1-\vec{N}_i\vec{N}_j) \qquad (7)$$

这里，$\vec{N}_i$ 代表外形 S 中第 i 个点的标准，可以通过拟合平面上的局部点领域来计算。实质上我们的先验状态，局部点领域应该是平的。注意到这一点，结合我们之前的能量，自然而然地就可以使常用的先验标准与遮挡轮廓[4]中的观察方向相垂直。

我们的总能量在等式 8 中给定。此外，我们还约束形变参数 L_2 的模 α_i 来避免非自然的较大形变。

$$E_{tot}(\bar{S},V,\alpha)=E_l(\bar{S},V)+\sum_i(E_s^i+E_{kp}^i+E_c^i+E_n^i+\sum_k(\|\alpha_{ik}V_k\|_F^2)) \qquad (8)$$

学习. 我们在等式 9 中解决了优化的问题，得到了我们的外表模型 $M=(\bar{S},V)$。在训练集合中，平均外形和形变基础可以使用次梯度计算的方法通过 $(\bar{S},V)$ 和 α 中降序的块坐标来推断。我们限制 $\|V_k\|_F$ 成为一个常量来解决公式中 V 和 α 尺度模糊的问题。为了处理不完善的分割和错误的关键点估计，我们使用以上能量的缩短版本以减少极端值的碰撞。在 PASCAL VOC 的刚体类别中，使用我们的算法所得的平均外形如图 4 所示。注意到除表示一个类别粗略外形细节之外，模型还得到纤维结构，如椅子的腿和自行车把手这类在形变上更突出的目标。

$$\min_{S,V,\alpha} \quad E_{tot}(\bar{S},V,\alpha)$$
$$subject\ to:\ S^i=\bar{S}+\sum_k\alpha_{ik}V_k \qquad (9)$$

我们的训练目标是高度的非凸和非平滑以及对于初始化的灵敏度高。我们按照[14]的方案，利用所有的训练实例计算一个柔软的可视外壳，并

以此来初始化我们的平均外形。形变基数和形变权重被随机地初始化。

3. 在野外进行重建

我们从大的图片向下来进行目标的重建，就像雕塑家首先捶打出大块雕塑，然后凿出细节部分。在检测和分割出场景中的目标之后，我们推断出它们粗略的三维位姿，然后以此来拟合我们对于有噪声分割掩膜制定的自上而下的外表模型。最后，我们从阴影线索还原出高频的外形细节。我们现在开始逐个介绍这些部分。

初始化. 在推断过程中，我们首先在图像中检测并分割出目标[20]，然后使用基于 CNN 的系统[34]（扩充估计子类别）预测出目标的视点（旋转矩阵）和子类别。我们的学习模型是在一个典型的边框尺度里，所有的目标在训练过程中都首先被调整到一个特殊的宽度。我们给出预测的边框，从而拓展得到的预测子类别的平均外形。最后，平均外形按照预测的视点被旋转，并且平移到预测边框的中心。

外形推断. 在初始化之后，我们解决了形变权重 α（初始化为 0），同时也通过优化等式（9）得到确定的 $\overline{S},V$。解决了所有的相机投影参数（尺度，平移和旋转）。需要注意的是在测试的是偶我们没有使用标注的关键点位置，这个“关键点一致性”的能量 E_{kp} 在优化过程中被忽略。

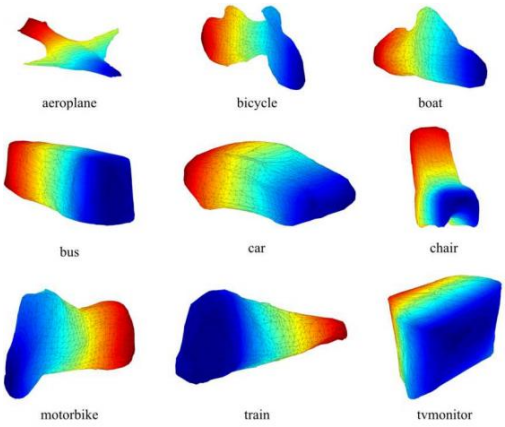

图 4：使用我们的基础形状公式在 PASCAL VOC 中得到的刚体的平均外形，图像显示了从正面观察时的颜色编码深度。

自下而上的外形改善. 上述的优化使我们得到了了一个基于类级模型的自上而下的三维重建，推断的目标轮廓，视点和我们的先验外形。我们提出了一个额外的处理步骤来恢复高频外形信息，通过改写固有的 Barron and Malik[5, 4]的图像算法 SIRFS，这个算法是利用的是外形，反射率和光照之间的统计规律。规范的 SIRFS 算法被制定为如下的优化问题：

$$\min_{Z,L} imize \quad g(I - S(Z,L)) + f(Z) + h(L)$$

这里 $R = I - S(Z,L)$ 是 log-反射率图像，Z 是一个深度图，L 是光照的球形函数模型。$S(Z,L)$ 是产生对数阴影

图像的渲染引擎。g, f 和 h 是与反射率，外形和光照各自对应的损失函数。

我们通过一个额外的损失项将当前的粗略的外形估计合并入SIRFS算法中：

$$f_o(Z, Z') = \sum_i ((Z_i - Z_i')^2 + \varepsilon^2)^{\gamma_0}$$

这里 Z' 是初始的粗略的外形，ε 是一个新加的参数来使损失项处处可微。我们通过渲染适合的三维外表模型的一个深度图来指导这个高度非凸成本函数的优化。这个自下而上的改善的输出是我们想要保留的反射率，外形和亮度图。

实现细节. 涉及到对外形和投影参数的优化，梯度的计算是效率极高的，我们使用近似的KNN算法来实行‘轮廓覆盖’的梯度和杠杆倒角距离场，来获得‘轮廓一致性’梯度。在使用单核CPU的情况下，我们总体的计算只花费了2秒来重建一个新的实例。我们的训练管道也一样的高效，只需几分钟的时间来学习一个给定目标类别的外表模型。

4. 实验

所做的实验主要评估了两个方面：1）通过重建所得的三维模型和底层训练数据的三维模型的匹配情况来评估模型的表现。2）研究对于图像的嘈杂自动分割和位姿预测的灵敏度。

数据集合. 对于所有的实验，我们使用来自于PASCAL VOC 2012具有挑战性的数据集[15]，它包含了10个刚体类别的图像（如表1所列）。我们使用公开的具有参考性的特定类别的关键点[9]和目标分割[19]。因为对于PASCAL VOC和其他大部分的检测数据集，参考的三维外形标准是无法获得的，所以我们在我们成功获得的最好的PASCAL 3D+数据集上来评估模型的表现，在PASCAL VOC中，这个数据集针对刚性类别，有多达10种的三维CAD模型。PASCAL 3D+对于“电视显示器”和“火车”提供了4种不同的模型，对于“汽车”和“椅子”提供了10种不同的模型。不同的匹配主要区分不同的子类别，但是仍然可能会产生冗余（对于四轮轿车就有超过3种的吻合）。在训练数据中，我们通过合并不同的案例来得到子类别标签，这还可以帮助我们抓住一些子类别的稀少的数据。对于PASCAL数据集，我们过滤掉闭塞的实例得到了一个子数据集，有70张沙发的图像，500张飞机和汽车的图像在此论文中没有提及。我们会将所有的图像集和我们的实现方法公开。

指标. 通过和PASCAL 3D+模型的比较，我们使用了两个指标量化了我们的三维模型的品质，（1）利用参考模

	Classes	aero	bike	boat	bus	car	chair	mbike	sofa	train	tv	mean
Mesh	KP+Mask	**5.00**	6.27	9.94	**6.22**	5.18	**5.20**	4.98	6.58	**12.60**	9.64	**7.16**
	Carvi[36]	5.07	**6.03**	**8.80**	8.76	**4.38**	5.74	**4.86**	**6.49**	17.52	**8.37**	7.60
	Puffball[35]	9.73	10.39	11.68	15.40	11.77	8.58	8.99	8.62	23.68	9.45	11.83
Depth	KP+Mask	**9.25**	7.87	12.36	**11.77**	7.22	7.51	8.97	**9.70**	30.91	6.84	**11.24**
	Carvi[36]	9.39	**7.24**	**11.43**	18.42	**6.86**	**7.39**	**8.06**	12.21	**29.57**	**5.75**	11.63
	SIRFS[4]	12.98	12.31	16.03	29.21	21.58	15.53	16.30	18.08	38.54	21.36	20.19

表 1：我们的三维模型的学习表现：使用 PASCAL VOC 中的参考关键点和编码的方法[36，35]和我们方法的对比。注意到[36]中对参考标注进行了操作，并且重建了一个图像集，我们的方法也是完成了同样的任务，可以从论文中获得细节信息。

	Classes	aero	bike	boat	bus	car	chair	mbike	sofa	train	tv	mean
Mesh	KP+Mask	5.13	6.46	10.46	5.89	5.07	5.34	5.15	15.07	12.16	11.69	8.24
	KP+SDS	4.96	6.58	10.58	4.67	4.97	5.40	5.21	15.08	12.78	12.18	8.24
	PP+SDS	6.58	14.02	14.43	6.65	7.96	7.47	7.57	15.21	15.23	13.24	10.84
	Puffball[35](SDS)	9.68	10.23	11.80	15.95	12.42	8.28	9.45	9.60	23.38	9.26	12.00
Depth	KP+Mask	9.02	7.26	13.51	12.10	8.04	8.02	10.00	23.05	25.57	7.48	12.41
	KP+SDS	9.07	7.98	13.57	9.90	7.98	7.96	9.99	22.57	23.59	7.64	12.03
	PP+SDS	10.94	11.64	12.26	15.95	13.17	10.06	12.55	21.19	36.37	8.98	15.31
	SIRFS[4]	11.80	11.83	15.98	29.15	21.64	15.58	16.91	19.64	37.58	23.01	20.31

表 2：在测试时，对于 PASCAL VOC 中的目标，我们的方法对多种类别标注进行的分离学习。可以看到，我们的方法对于松散的标注有着缓慢的下降。注意到这些实验都是在一个训练/测试 的设置中进行的，并且数字与表 1 中不同，可以从论文中获得详细的信息。

型的三维边框尺寸标准化的 Hausdorff 距离和（2）深度图误差。这两个指标用来评估重建可视目标表面的质量，并且可以通过重建深度和参考深度之间的绝对平均距离来衡量：

$$Z-MAW(\hat{Z},Z^{*})=\frac{1}{\eta\cdot\gamma}\cdot \min_{\beta}\sum_{x,y}|\hat{Z}_{x,y}-Z^{*}_{x,y}-\beta| \quad (10)$$

这里 $\hat{Z}$ 和 $Z*$ 分别代表预测的深度图和参考的标准深度图。通过分析，β 可以通过计算 $\hat{Z}-Z*$ 的中值来的到，γ 是用来解释目标绝对距离（我们使用的边界框的对角线）的归一化因子。注意我们的深度图误差是平动的并且是尺度不变的。

4.1. 学习三维模型的表现

我们按照 Vicente et al[36]的建立过程把我们的三维模型用在相同的整个数据集上（没有 训练/测试分离）。表 1 比较了我们在 PASCAL VOC 上的重建和最近针对这个课题提出的其他方法的重建效果（这不是专门为嘈杂数据设计的），也和一种类无关艺术状态的外形膨胀的方法进行了比较，这种方法也是基于单个轮廓的重建。在两个基准下，我们的模型

都展现了具有竞争力的表现，对于火车和公共汽车上的透视缩短效果，我们的模型显示了个号的稳健性。类别无关的方法——Puffball[35]和SIRFS[4]，在基准上的表现持续地变差。某些类别，例如船和电视显示器是特别难的，分别因为较大的组内差异和稀少数据。

4.2 灵敏度分析

为了分析模型对于有噪声输入的灵敏度，对于视点估计，我们分别比较了我们方法的各个版本，例如使用参考标准方法（Mask）/不完善的分割（SDS）和关键点（KP）/位姿预测器（PP）等。对于位姿预测，我们使用了基于CNN的系统[34]，并且在测试时利用它来预测子类型，它的实现是通过训练[34]中所描述的系统，同时加上从 PASCAL 3D+中得到的子类别标签。为了从边界框中得到一个近似的分割，我们使用的是[20]中提出的最先进的联合检测与分割系统，并对它进行了改善。

这里，我们使用了一个训练/测试的设置，只使用数据的一个子集来训练模型，然后利用模型从边界框中重建数据。表2展示了我们的结果，从完全标注到全自动设置。我们方法对于一些错误的分割有很好的鲁棒性，这是因为在面对有噪声的轮廓时，我们的模型会防止外形产生非自然的弯曲。当面对一些不完善的位姿初始化时，我们的模型精确性会产生轻微的下降，即使我们的映射参数优化会在一定程度上处理这个问题，也没有收到非常好的效果。得到了预测的位姿之后，我们可以观察到有时我们的重建看起来与实物和接近，但是错误率很高，说明指标对于比较差的对齐有着很高的灵敏度。数据稀少的问题在沙发的例子中尤为明显，在表2中可以看出，在训练数据量减少时（只有34个实例），数字结果出现了显著的下降。注意到我们并没有评价PASCAL 3D+提供的自下而上的组成部分，这部分没有同样展现出实例的高频外形细节。我们将在下一个小章节中给出定性的结果。

4.3. 完全自动重建

我们在图 5 中定性地展示了以0.5的IoU重叠在PASCAL VOC整个图像上进行自动检测和分割实例的重建过程[20]，可以看见我们的方法有能力处理一定程度上的错误分割。我们的一些失败的例子包括无法获得目标正确的尺度和位姿，因此在一些例子中与轮廓吻合得很差。我们的子类别预测也在一些实例中失败了（CRT vs 平面显示屏），导致不正确的重建。我们在补充材料中加入了更多的这样的图像，以便读者查看。

5. 结论

我们提出了在一个现实的数据集

合中实现基于单张图像的完全自动的目标重建方法，它可能是第一个解决这个问题的方法。关键点在于，我们的可变三维外表模型是可以从一些易获得的二维参考标注中自举的，从而绕过了手动的匹配设计或者是三维扫描，并且让我们可以在较大的真实世界的数据集合中方便地使用这些模型（PASCAL VOC）。我们在PASCAL VOC 三维基准数据集[39]上进行了一个对于三维学习模型的广泛的评估，并且与那些特殊类别的利用标准分割作为输入的方法相比，我们方法的结果也是相当不错的，具有很强的竞争力，同时我们的方法还能在自动目标探测器的基础上用于野外的图像重建。

大量的研究成果摆在我们的面前，无论是在测试阶段的质量提高和重建的鲁棒性加强（自下而上和自上而下两个部分），还是在训练过程中发展联合识别与重建的基准，以及对于标注的需求，所有的这些组成了未来工作的重要的方向。更具表现力的非线性外表模型[38]可能是很有帮助的，分割与重建之间更为紧密的整合也会发挥作用。

致谢

这个工作得到了来自于 NSF Award IIS-1212798 和 ONR MURI-N00014-10-1-0933 的部分支持。Shubham Tulsiani 是由 Berkeley fellowship 支持的，Joao Carreira 是由 Portuguese Science Foundation，FCT 支持的。我们非常感谢 NVIDIA 公司为该项研究捐赠的 Tesla GPUs。

Reference

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In ACM Trans. Graph., 2005. 2

[2] P. Arbel´aez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. CVPR, 2014. 1

[3] N. Aspert, D. Santa-Cruz, and T. Ebrahimi. Mesh: Measuring errors between surfaces using the hausdorff distance. In ICME, 2002. 6

[4] J. T. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. ECCV, 2012. 4, 5, 6

[5] J. T. Barron and J. Malik. Shape, illumination, and re fl ectance from shading. Technical Report UCB/EECS-2013117, EECS, UC Berkeley, May 2013. 5

[6] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and P. Sayd. Coarse-to-fine

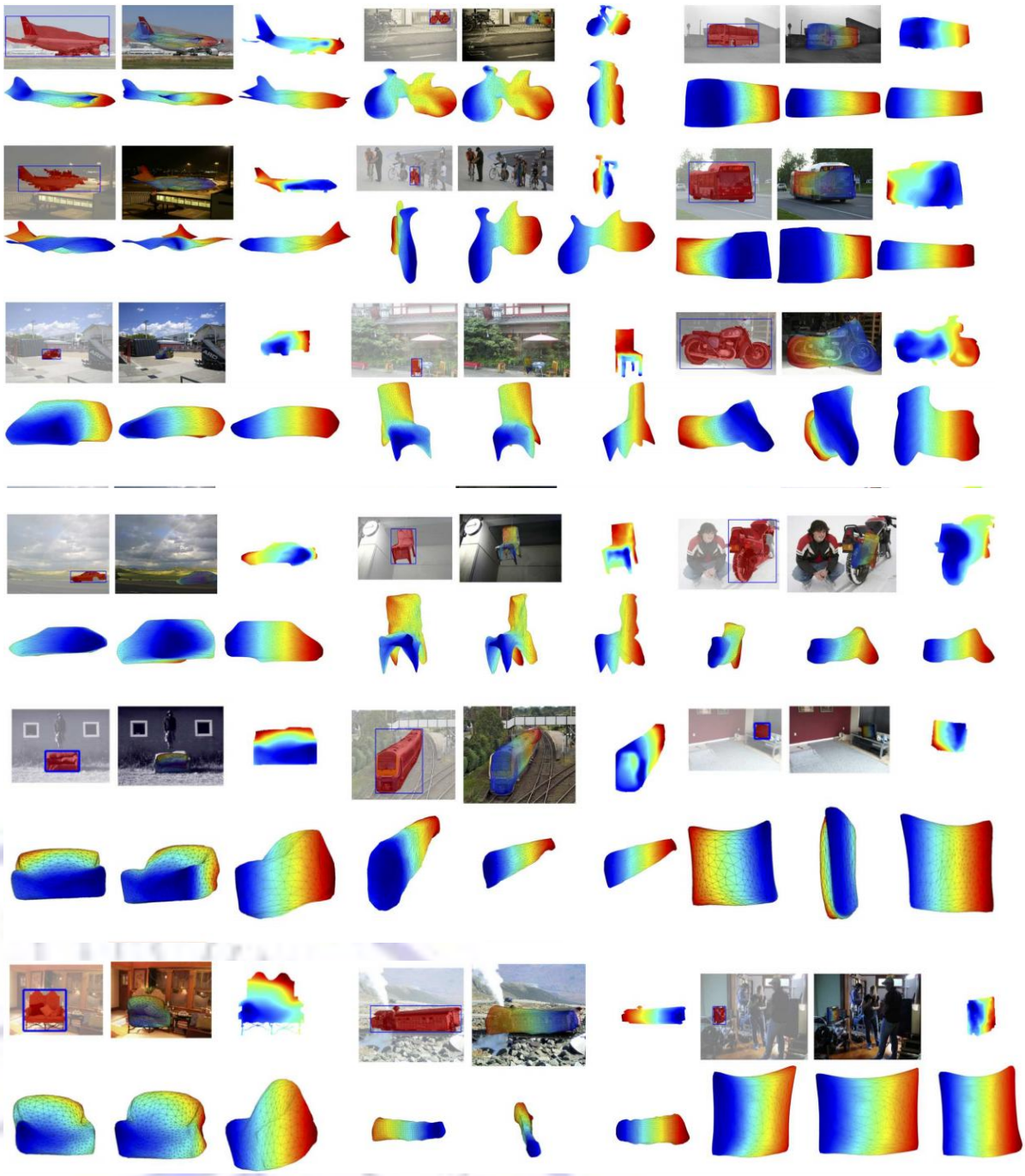

图 5:使用我们的模型在 PASCAL VOC 中的刚体类别中进行的全自动实例重建（参考值：0.5IoU）。我们展示了我们输入的实例分割，覆盖在图像上的推测外形，一个 2.5 维德深度图（在自下而上的改善步骤之后），图像视点上的网格和两个其他视点，从图中可以看出我们的方法产生了十分吻合的结果，对于单张图片，并且在实例分割含有噪声的情况，这是一个卓越的成就。图中也展示了在坐标系中的颜色编码深度（蓝色标识近处），更多的结果可以在 http://goo.gl/ImALxQ 中找到。

low-rank structure-from-motion. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8, June 2008. 2

[7] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques, 1999. 2

[8] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. TPAMI, 25(9):1063–1074, 2003. 3

[9] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In European Conference on Computer Vision (ECCV), 2010. 5

[10] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on, v。olume 2, pages 690–696 vol.2, 2000. 2

[11] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In CVPR, 2010. 1

[12] T. Cashman and A. Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(1):232–244, Jan 2013. 3

[13] Y. Chen, T.-K. Kim, and R. Cipolla. Inferring 3d shapes and deformations from single views. In Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III, ECCV'10, pages 300–313, Berlin, Heidelberg, 2010. Springer-Verlag. 3

[14] C. H. Esteban and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. Comput. Vis. Image Underst., 96(3):367–392, Dec. 2004. 3, 4

[15] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascalnetwork.org/challenges/VOC/voc2012/workshop/index.html. 1, 5

[16] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In CVPR, June 2013. 2

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich

feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014. 1

[18] 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. 2

[19] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In ICCV, 2011. 5

[20] B. Hariharan, P. Arbel´aez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In European Conference on Computer Vision (ECCV), 2014. 1, 4, 6, 7

[21] M. Hejrati and D. Ramanan. Analyzing 3d objects in cluttered images. In NIPS, pages 602–610, 2012. 2

[22] E. Kalogerakis, S. Chaudhuri, D. Koller, and V. Koltun. A Probabilistic Model of Component-Based Shape Synthesis. ACM Transactions on Graphics, 31(4), 2012. 3

[23] I. Kemelmacher-Shlizerman. Internet based morphable model. In International Conference on Computer Vision (ICCV), 2011. 1

[24] A. Laurentini. The visual hull concept for silhouette-based image understanding. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 16(2):150–162, Feb 1994. 3

[25] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. In ICCV, 2013. 2

[26] C. Nandakumar, A. Torralba, and J. Malik. How little do we need for 3-d shape perception? Perception-London, 40(3):257, 2011. 1

[27] R. Nevatia and T. O. Binford. Description and recognition of curved objects. Arti fi cial Intelligence, 8(1):77–98, 1977.

[28] M. Prasad, A. Fitzgibbon, A. Zisserman, and L. Van Gool. Finding nemo: Deformable object class modelling using curve matching. In CVPR, 2010. 2

[29] L. G. Roberts. Machine Perception of Three-Dimensional Solids. PhD thesis, Massachusetts Institute of Technology, 1963. 2

[30] Y. Sahilliolu and Y. Yemez. A surface deformation framework for 3d shape recovery. In Multimedia Content Representation, Classi fi cation and Security, volume 4105 of

Lecture Notes in Computer Science, pages 570－577. Springer Berlin Heidelberg, 2006. 3

[31] S. Satkin, M. Rashid, J. Lin, and M. Hebert. 3dnn: 3d nearest neighbor. International Journal of Computer Vision, pages 1－29, 2014. 2

[32] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. Seitz. Total moving face reconstruction. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, Computer Vision ECCV 2014, volume 8692 of Lecture Notes in Computer Science, pages 796－812. Springer International Publishing, 2014. 1

[33] L. Torresani, A. Hertzmann, and C. Bregler. Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. TPAMI, 2008.

[34] S. Tulsiani and J. Malik. Viewpoints and keypoints. In CVPR. 2015. 1, 4, 6

[35] N. R. Twarog, M. F. Tappen, and E. H. Adelson. Playing with puffball: simple scale-invariant inflation for use in vision and graphics. In ACM Symp. on Applied Perception, 2012. 6

[36] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. CVPR 2014, 2014. 1, 2, 3, 6

[37] S. Vicente and L. de Agapito. Balloon shapes: Reconstructing and deforming objects with volume from images. In 3DV, pages 223－230. IEEE, 2013. 2

[38] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shape modeling. In CVPR. 2015. 7

[39] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In WACV, 2014. 5, 7

[40] J. Xiao, B. Russell, and A. Torralba. Localizing 3d cuboids in single-view images. In Advances in Neural Information Processing Systems, pages 746－754, 2012. 2

[41] S. Zhu, L. Zhang, and B. Smith. Model evolution: An incremental approach to non-rigid structure from motion. In CVPR, 2010. 2

[42] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3d representations for object recognition and modeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(11):2608－2623, 2013.