引用格式：桑晗博，林巍峣，叶龙.基于深度三维模型表征的类别级六维位姿估计[J].中国传媒大学学报（自然科学版），2022，29(04)：50-56.
文章编号：1673-4793(2022)04-0050-07

基于深度三维模型表征的类别级六维位姿估计

桑晗博[1]，林巍峣[1*]，叶龙[2]
（1. 上海交通大学电子信息与电气工程学院，上海 200241；2. 中国传媒大学媒体融合与传播国家重点实验室，北京 100024）

摘要：类别级物体六维位姿估计在机器人操作、自动驾驶和增强现实等领域有着广泛的应用。相较于实例级任务，类别级六维位姿估计的难点主要在于类别先验特征基础上的类内差异。本文采用一种基于有向距离场(Signed Distance Field,SDF)的深度三维模型表征提取出类别级先验共享信息，同时依据输入深度图像的几何形状特征搜索最优的形状隐变量，两者结合重建出标准空间内的完整实例模型。通过学习深度点与标准化实例模型的点对匹配关系，即可求解出物体的六维位姿参数。实验证明本文提出的类别级六维位姿估计架构具有良好的性能和对类内新物体的泛化能力。
关键词：类别级物体六维位姿估计；深度三维模型表征
中图分类号：TP183　　**文献标识码：**A

3D deep implicit function for category-level object 6D pose estimation

SANG Hanbo[1], LIN Weiyao[1*], YE Long[2]
（1. Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200241, China; 2. State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China）

Abstract: Category-level object 6D pose estimation is important for the task of robot manipulation, autonomous driving and augmented reality. Compared with instance-level one, the challenge of category-level 6D pose estimation mainly lies in the intra-class variation given a category prior. In this paper, a deep implicit function for representing 3D model based on SDF is adopted to extract the shared category-level prior. At the same time, the optimal shape latent code is predicted according to the geometric feature extracted from the input depth image. Both of the shared prior decoder and the specific shape latent code are combined together to reconstruct the complete instance in the normalized canonical space. Then the 6D pose could be solved by estimating the point matching between the depth point cloud and the canonical instance. Experiments show that the proposed framework for category-level 6D pose estimation achieves relatively good performance as well as generalization ability for novel instances within the same category.
Key words: category-level 6D object pose estimation; deep implicit function

基金项目：媒体融合与传播国家重点实验室开放课题（SKLMCC2021KF007）
作者简介（*为通讯作者）：桑晗博（1998-），男，博士研究生，主要从事人工智能研究。Email: hanbosang@sjtu.edu.cn；林巍峣（1980-），男，教授、博士生导师，主要从事人工智能研究。Email: wylin@sjtu.edu.cn；叶龙（1981-），男，教授、博士生导师，主要从事人工智能研究。Email: yelong@cuc.edu.cn

1　引言

1.1　研究背景及意义

六维物体位姿估计是计算机视觉中的一项重要任务，它能够由输入图像恢复出物体在当前相机中心坐标系下的三维位置和方向，被广泛应用于机械臂操作[1]、自动驾驶[2]和虚拟增强现实[3]等领域。

目前，大多数现有的工作都集中在实例级位姿估计上，即预先提供测试物体的精确三维模型；然而在真实场景中，很难为每个新对象建立精确的三维模型。因此，实例级位姿估计大多只能处理训练集中出现过的物体。近期，类别级六维物体位姿估计方法得到了广泛的关注和研究，它可以在没有显式给定三维模型的情况下处理训练中未见过的新对象（该对象必须属于训练集中的某一已知类别）。与实例级位姿估计方法相比，类别级位姿估计对实际应用场景具有更强的适应能力，但同时也对模型设计提出了更大的挑战：第一，未知物体的数据分布和训练数据之间存在差异，需要模型有更强的泛化能力；第二，物体位姿是相对于物体坐标系下的物体初始位姿定义的，需要给出每个物体实例的初始位姿先验信息；第三，由于不提供物体的三维模型，对于增强现实等可视化任务，无法显式渲染出三维场景。本文通过对三维物体的深度表征提供获取类别级先验信息和减小类内形状差异的一种解决思路。

1.2　国内外技术发展现状与研究

随着深度学习的发展，早期的研究聚焦于实例级物体六维位姿估计取代传统的模板匹配方法，即已知测试物体的精确三维模型。其中部分工作[4,5]由输入图像直接回归目标物体的六维位姿参数，然而这容易导致空间信息的丢失；另一类工作[6,7]首先估计图像中目标前景像素点与初始状态的物体局部坐标系下三维点坐标的点对匹配关系，再利用匹配关系求解出位姿参数。

实例级位姿估计大多只能处理训练过程中见过的物体，然而在真实场景中获取物体的三维模型是相当困难的。近期，类别级六维物体位姿估计方法得到了广泛的关注和研究，它可以在没有显式给定三维模型的情况下处理训练中未见过的新对象。NOCS[8]是第一个基于深度学习的类别级六维物体位姿估计框架，通过显式构建一类物体的NOCS共享空间为初始位姿提供先验；为解决不同实例间的类内差异，部分方法建模由类别模板先验到目标实例间的变形[9,10]，或提取出更具代表性的实例物体的几何特征[11,12]。然而由于三维模型未知，这些方法通常需要显式重建出实例物体的三维点云模型，大大增加了模型的收敛难度。本文通过深度SDF表征提取类别级先验信息，并通过重建实例的归一化模型减小类内形状差异，提高模型的性能和泛化能力。

1.3　研究内容与主要贡献

本文研究未知三维模型情况下的类别级物体六维位姿估计问题。为解决同一类别内物体间的形状差异（类内差异），现有方法致力于充分利用输入包含深度的空间形状信息，重建出输入物体的点云几何形状[11]或初始位姿下的点云模型[9]。但是，点云模型天然的稀疏性导致这些方法难以重建出连续的表面以及精确的三维结构，不仅无法准确提取出物体的形状特征，且不便于进行渲染，难以用于增强现实等可视化需求。因此，本文采用一种深度隐式三维模型DeepSDF[13]进行三维形状特征的表征，从而同时保留共享的类别级先验信息以及每个类内实例独特的形状特征。本文的主要贡献有以下三点：

（1）. 提出了一个基于深度三维模型表征的类别级物体六维姿态估计架构；

（2）. 通过构建类别级先验解码器以及归一化模型重建模块，充分提取出类别级先验信息，同时减小了类内形状差异；

（3）. 在合成和真实数据集上都取得了良好的效果，具有一定的泛化能力。

2　模型架构设计

类别级物体六维位姿估计的目的是依据输入的RGB-D图像I估计其中包含的所有物体的旋转参数$R \in SO(3)$和平移参数$t \in R^3$，R和t构成的位姿参数描述物体由局部坐标系下预定义的初始位姿变换到当前相机坐标系下。本文采用四元数表征三维旋转参数。完整的模型架构设计如图1所示，首先用已预训练好的检测分割模块获取图像中每一物体的前景区域，分割出的深度点重构出相机坐标系下的可见部分点云M_{vis}^{c}；其次通过估计得到的类别选取类别先验解码器（2.1节），隐式地提取出一类物体的共享先验特征；随后通过M_{vis}^{c}的视觉信息重建对应实例模型的隐变量（2.2节），利用先验解码器重建出对应实例在局部物体坐标系下的归一化精细模型M_{rec}^{o}；最后用NOCS作为监督估计M_{vis}^{c}到M_{rec}^{o}的密集点对匹配（2.3节），根据点对匹配关系求解出六维位姿以及尺度因子。

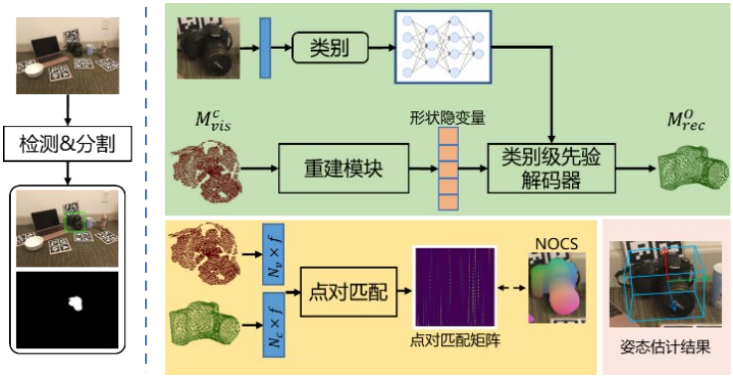

图1 模型总体架构

2.1 基于深度SDF的类别级先验学习模块

由于类别级物体位姿估计问题中物体的三维模型和初始位姿均未知,需要对其在局部坐标系下的形状和位姿进行先验建模。受NOCS[8]启发,虽然不同实例的形状有所不同,但同一类别的对象通常具有共同的属性和语义结构,人类可以从丰富的经验和先验知识中轻松推断出没见过的物体的三维形状;例如水杯通常是圆柱形的,水杯之间的差异通常只是圆柱体的高度和直径。与NOCS直接将所有同类别物体的尺寸和方向显式归一化在同一共享空间不同,本文希望隐式地构建该先验信息,使其更适应深度学习架构。

一系列方法[9, 14]利用自编码器为同类别的多个实例点云模型提取高维特征作为隐变量表征原物体的形状,并将这些高维特征的均值作为类别先验。本文使用深度SDF表征物体三维模型,相对于点云能够重建出连续的表面结构和更细致的几何架构,提供更丰富的类别先验。

本文首先训练一个DeepSDF解码器,从数据集中学习类别级形状先验知识(该数据集可以和六维姿态估计数据集的物体实例不同)。SDF表示有符号距离函数,其中点的SDF值表示到曲面边界的距离,符号表示区域是在形状的内部(-)还是外部(+),模型表面即为SDF值为0的决策边界。深度SDF模型通过建立空间内离散的三维查询点 $x = \{x_1,...,x_N\}$ 到SDF值的映射关系构建三维表面的隐空间模型:

$$f_D(x; z) = d \tag{1}$$

其中 $d = \{d_1,...,d_N\}$ 为查询点 x 对应的SDF值,f_D 为包含了类别先验信息公共属性的DeepSDF解码器,z 为表征实例形状特征的隐变量,即每个实例映射到 z。测试时通过解码器,物体表面可以隐式地由 $F(z) = 0$ 的等值面表示并重建出精细的三维模型[15]。

为最大程度的区分不同类别的先验信息,本文为每一类别单独训练一个深度SDF解码器。在训练过程中,首先将所有三维模型标准化,即对齐坐标原点和旋转轴,并缩放到相同尺度。因此,解码器重建的三维模型 M_{rec}^{o} 是在物体局部坐标系下的归一化模型,其中 O 表示物体局部坐标系。

2.2 基于形状差异表征的隐变量重建模块

2.1节中的类别级解码器隐式包含了类别级共享先验信息,使得模型初步具备处理新物体的能力,但由于类内差异较大,类别级先验特征无法完全表征当前输入物体。类内差异是指尽管属于同一类别、具有共享的相似特征,但类别内各实例间仍存在较大差异,例如颜色差异、形状差异、尺寸差异等。如图2所示,对于相机这一类别的实例,相对于模板的形状各异。

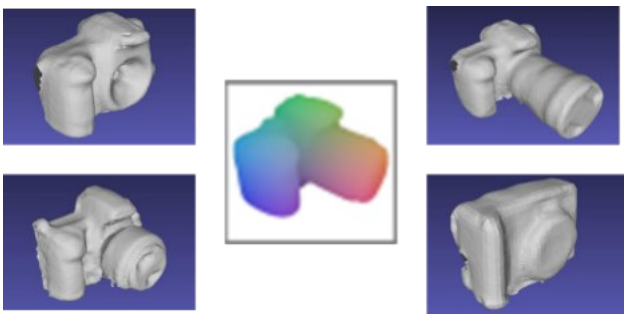

图2 类内差异示意图

相对于实例级位姿估计方法常用的视觉特征和深度特征密集融合[5],本文仅使用深度图的几何信息进行位姿估计,从而消除颜色差异。为减小类内形状

差异对泛化性能的影响，本文根据先验解码器中包含的类别模板信息在局部物体坐标系中重建出当前输入物体的归一化实例模型，从而充分提取目标物体的形状特征。

为了由输入深度图重构的可见模型点 M_{vis}^{c} 重建出局部物体坐标系下的模型 M_{rec}^{o}，一部分方法[9]设计复杂的网络估计每个点的偏移，模型复杂度高且重建的点云较为粗糙。利用深度 SDF 表征有利于提高重建精度、降低模型复杂度。一种简单直观的重建思路是直接由 M_{vis}^{c} 估计出最优的隐变量 $\hat{z}$，并用 $\hat{z}$ 的真实标注进行监督（隐变量的真实值由该实例对应的物体坐标模型获取）。这一做法实质上是将低维空间密集点的重建转换为高维隐空间中隐变量的重建，通过特征降维降低模型复杂度，但是在训练过程中仍然需要三维模型标注，并且容易使重建网络过度“记住”输入的三维模型，影响其泛化到新物体上的能力。

为此，本文利用 SDF 标注学习隐变量重建网络的模型参数。依据训练过程中已知的 NOCS 坐标，可以获取输入实例在归一化局部物体坐标系下的可见点坐标；为这些可见点构建 SDF 样本，沿表面发现方向在内外侧 γ 距离的位置各取一个点，构成模型表面内外的采样点 x_{nocs}，SDF 值分别为 $+\gamma$ 和 $-\gamma$。用构造的 SDF 样本对隐变量重建网络进行监督：

$$L_z(d)=\sum_{i}^{2N_v}\left|\left|f_D\left(x_{nocs}^{i},z\right)-d_{\gamma}^{i}\right|\right|_1+\frac{1}{\sigma^2}||z||_2 \tag{2}$$

其中，$f_D(\cdot)$ 为类别级先验解码器，$d_{\gamma}^{i}\in\{+\gamma^{i},-\gamma^{i}\}$ 为每个 NOCS 采样点的 SDF 标注。实验表明，相对于直接用隐变量的值进行显式监督，该方法具有更好的性能。

2.3　相机坐标系到物体坐标系的点对匹配模块

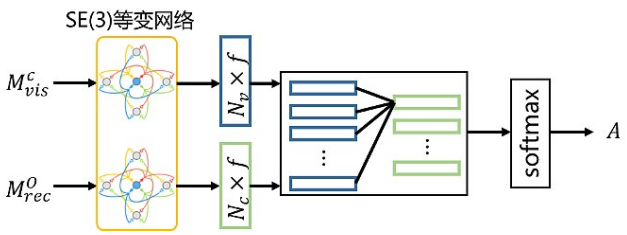

图 3　点对匹配模块

图 1 中所示点对匹配模块求解当前相机坐标系下的观测物体 $M_{vis}^{c}\in R^{N_v\times 3}$ 相对于局部物体坐标系中重建的 $M_{rec}^{o}\in R^{N_c\times 3}$ 的密集点对匹配关系矩阵 $A\in R^{N_v\times N_c}$，从而求解出六维姿态变换。网络结构设计如图 3 所示，通过 SE(3) 等变网络[16]提取出两者的密集点特征，其中 SE(3) 等变网络能够更好的保留输入模型的旋转信息；然后构建两两点对特征之间的交叉注意力，softmax 后获取匹配关系 A。将 A 和 M_{rec}^{o} 相乘即可获得输入的可见点对应的 NOCS 坐标估计值 $\hat{P}\in R^{N_v\times 3}$，利用 NOCS 坐标对点对匹配网络进行监督学习：

$$L_{cor}\left(\hat{P},P\right)=\frac{1}{N_v}\sum_{x\in\hat{P},y\in P}\begin{cases}5|x-y|^2, & |x-y|\leqslant 0.1\\ |x-y|-0.05, & \text{otherwise}\end{cases} \tag{3}$$

依据当前相机坐标系下的观测点 M_{vis}^{c} 相对于局部物体坐标系中重建 M_{rec}^{o} 的密集点对匹配，本文采用 Umeyama 算法[17]求解从物体坐标系到相机坐标系的六维位姿 $R|t$。

2.4　实现细节

对于输入图像和深度，采样 $N_v=1024$ 个点作为当前相机坐标系下的可见点。训练深度类别级先验解码器时，表征实例形状特征的隐变量 z 的维度设为 16，每个解码器训练迭代 1500 轮。对于隐变量重建网络和点对匹配网络的联合训练，迭代 200 轮，取初始学习率为 0.0001。

对于对称物体类别（例如瓶子、碗和罐头），本文忽略了围绕对称轴的旋转误差。对于容易因为自遮挡转变对称性的物体类别（如杯子），在把手不可见的情况下，将杯子视为对称对象；在把手可见的情况下，将杯子视为非对称对象。

3　实验结果与分析

3.1　数据集和评估指标

本文使用合成数据集 CAMERA[8]进行训练，真实数据集 REAL 进行测试。CAMERA 数据集是通过虚拟引擎以上下文感知的方式渲染三维对象并将其合成为真实场景而生成的，总共有 300K 张合成图像，其中 25K 用于评估。其中包含 1085个对象实例。其评估集包含 184个不同的实例。REAL 数据集通过相机拍摄真实场景和物体构成，共计 4300 张包含 7个场景的真实图像用于训练，2750 张包含 6个场景的真实图像用于评估。每组包含 18个真实对象实例。两个数据集的物体实例均从 6个不同类别中选择——瓶子、碗、相机、罐头、笔记本电脑和杯子，如图 4 所示。所有物体都预先进行了物体坐标系下中心点和旋转方向的对齐，并对尺度进行了归一化。

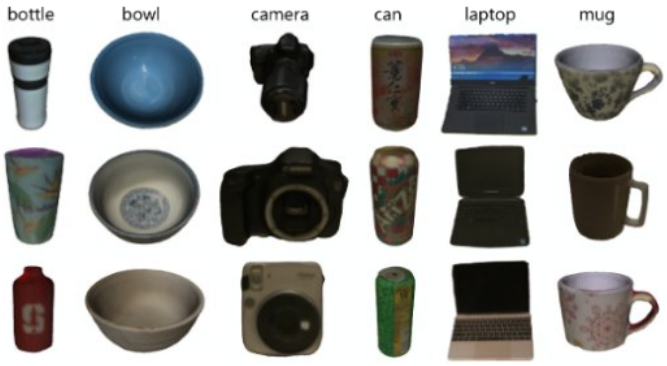

图4 数据集展示

表1 实验结果与方法对比

	CAMERA			REAL		
	5°2 cm	5°5cm	10°2 cm	5°2 cm	5°5cm	10°2 cm
NOCS[8]	32.3	40.9	48.2	7.2	10.0	13.8
SPD[9]	54.3	59	73.3	19.3	21.4	43.2
SGPA[10]	**70.7**	**74.5**	**82.7**	35.9	39.6	61.3
FS-Net[11]	–	–	–	–	28.2	–
DualPose[12]	64.7	70.7	77.2	29.3	35.9	50.0
OURS	69.2	72.3	78.9	**36.5**	**44.8**	**63.6**

本文使用位姿误差 $n^{\circ}mcm$ 对三维位姿估计结果进行评估。$n^{\circ}mcm$ 计算所有测试样本中旋转估计偏差 $\Delta R \leq n^{\circ}$ 以及位移估计偏差 $\Delta t \leq m$ 的占比，选用 $5^{\circ}2cm$、$5^{\circ}5cm$ 和 $10^{\circ}2cm$ 三个阈值标准。

3.2 模型效果和对比

将本文所提方法的结果与NOCS[8]、SPD[9]、SGPA[10]以及DualPose[12]分别在CAMERA和REAL数据集上进行对比，与FS-Net[11]在REAL数据集上进行对比，结果如表1所示。其中，$n^{\circ}mcm$ 反映六维位姿估计效果。

根据 $n^{\circ}mcm$ 评估六维位姿估计的模型表现。对于CAMERA合成数据集，本文超过了NOCS、SPD和DualPose，与SGPA的表现较为接近（在 $5^{\circ}2cm$、$5^{\circ}5cm$ 和 $10^{\circ}2cm$ 上仅分别降低2.1%、3%和4.6%）。由于SGPA使用了复杂的Transformer结构估计点对匹配，性能较好，但耗时相对较长。

对于REAL真实数据集，本文在 $5^{\circ}2cm$ 和 $5^{\circ}5cm$ 上均达到了最佳表现，其中在 $5^{\circ}5cm$ 方面超过其它方法10%以上。值得注意的是，REAL中包含训练和测试数据，NOCS、SPD、SGPA和DualPose都采用了混合训练策略，即以一定的比例利用CAMERA合成数据和REAL中的训练数据一起作为模型训练集，并在REAL测试数据上进行验证；FS-Net只在REAL训练数据上进行训练；因此，这些方法的训练数据和测试数据存在分布交叠，并不能完全拟合类别级物体六维位姿估计的实际情况（完全没见过的新实例）。本文所提方法仅在CAMERA合成数据上进行训练，在REAL测试数据上验证模型效果，测试数据和训练数据的分布差距更大，因此更能证明该方法的泛化能力。

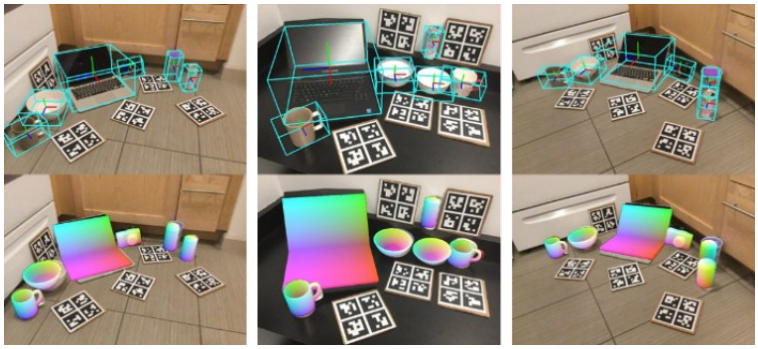

图5 位姿估计结果可视化

图5展示了本文的方法在REAL数据集上的可视化结果。其中，上面一行展示了标准化包围框按六维位姿投影到图像上的结果（标准化包围框为每一类别实例共享）。下面一行展示由输入深度点重建实例在

标准化空间的三维模型的重建效果（将重建的模型按位姿渲染到相机坐标平面上）；可以看出，基本能重建出完整的几何形状，对于部分细节（如水杯的把手）也能较好的提取。

3.3　消融实验

表 2 消融实验结果

	CAMERA		REAL	
	$5^\circ 2cm$	$5^\circ 5cm$	$5^\circ 2cm$	$5^\circ 5cm$
MeanPri	42.2	50.8	10.4	16.7
PRec	61.9	65.4	25.2	30.6
EmbRec	69.6	73.1	31.9	38
MLP	66.5	70.7	34.1	41.6
OURS	**69.2**	**72.3**	**36.5**	**44.8**

表 2 给出了各设计模块的消融实验。其中，OURS 表示最终的模型架构，MeanPri 表示使用隐变量的均值重建类别模板模型作为 M_{rec}^{O}，PRec 表示使用通过观测到的输入 M_{vis}^{c} 直接在三维空间重建实例点云，EmbRec 表示使用隐变量的真实值监督重建网络，MLP 表示使用多层感知网络取代点对匹配模块中的 SE(3) 等变网络。

MeanPri 根据预先得到的类内所有实例的隐变量取均值，求得类别模板先验，并用解码得到的模型直接送入点对匹配模块，即类内所有实例共用同一标准化模型。由于类内形状差异大，这种做法无益于提取实例独特的形状特征，会极大降低模型的性能。

PRec 和 EmbRec 根据类别先验分别进行实例的点云模型和对应的隐变量的显式重建。点云重建由于模型复杂度大大增加，性能下降较明显；隐变量重建在 CAMERA 数据集上的效果反而提升了，但在 REAL 上有明显降低。这可能是因为 CAMERA 测试数据和训练数据分布差距不大，而 REAL 测试数据是训练中完全未知的新物体。这进一步表明本文提出的 SDF 间接监督方法有助于提高泛化能力。

MLP 使用多层感知网络取代点对匹配模块中的 SE(3) 等变网络对三维数据进行特征提取；由于等变网络能够更好的保留输入中的姿态信息，有利于匹配网络的学习，因此使用等变网络作为特征提取骨架有一定的优势。

4　总结

本文采用一种基于 SDF 的深度三维模型表征提取出类别级先验共享信息，同时依据输入深度图像的几何形状特征搜索最优的形状隐变量，从而重建出标准空间内的完整实例模型。通过学习深度点与标准化实例模型的点对匹配关系，即可求解出物体的六维位姿参数。本文提出的类别级六维位姿估计架构具有良好的性能以及对一定的类内新物体的泛化能力。

参考文献（References）：

[1] Azad P, Asfour T, Dillmann R. Stereo-based 6d object localization for grasping with humanoid robot systems [C]// 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2007: 919-924.

[2] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012: 3354-3361.

[3] Marchand E, Uchiyama H, Spindler F. Pose estimation for augmented reality: a hands-on survey [J]. IEEE Transactions on Visualization and Computer Graphics, 2015, 22 (12): 2633-2651.

[4] Xiang Y, Schmidt T, Narayanan V, et al. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes[DB/OL]. arXiv:1711.00199,.

[5] Wang C, Xu D, Zhu Y, et al. Densefusion: 6d object pose estimation by iterative dense fusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3343-3352.

[6] Peng S, Liu Y, Huang Q, et al. Pvnet: Pixel-wise voting network for 6dof pose estimation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4561-4570.

[7] He Y, Sun W, Huang H, et al. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11632-11641.

[8] Wang H, Sridhar S, Huang J, et al. Normalized object coordinate space for category-level 6d object pose and size estimation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 2642-2651.

[9] Tian M, Ang M H, Lee G H. Shape prior deformation for categorical 6d object pose and size estimation[C]//European Conference on Computer Vision. Springer, Cham, 2020: 530-546.

[10] Chen K, Dou Q. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation[C]//Proceed-

ings of the IEEE/CVF International Conference on Computer Vision, 2021: 2773-2782.

[11] Chen W, Jia X, Chang H J, et al. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 1581-1590.

[12] Lin J, Wei Z, Li Z, et al. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 3560-3569.

[13] Park J J, Florence P, Straub J, et al. Deepsdf: Learning continuous signed distance functions for shape representation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 165-174.

[14] Chen D, Li J, Wang Z, et al. Learning canonical shape space for category-level 6d object pose and size estimation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11973-11982.

[15] Lorensen W E, Cline H E. Marching cubes: A high resolution 3D surface construction algorithm [J]. ACM Siggraph Computer Graphics, 1987, 21(4): 163-169.

[16] Chen H, Liu S, Chen W, et al. Equivariant point network for 3d point cloud analysis [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 14514-14523.

[17] Umeyama S. Least-squares estimation of transformation parameters between two point patterns [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1991, 13(04): 376-380.

编辑:王谦

(上接第25页)

tention networks for semantic segmentation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Washington: IEEE Press, 2019: 9167-9176.

[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Neural Information Processing Systems, New York: Curran, 2017:6000-6010.

[20] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[DB/OL]. arXiv:2010.11929.

[21] Zheng S, Lu J, Zhao H, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 6881-6890

[22] Xie E, Wang W, Yu Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers [C]// Advances in Neural Information Processing Systems, 2021: 34.

[23] Xu N, Price B, Cohen S, et al. Deep interactive object selection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 373-381.

[24] Jang W D, Kim C S. Interactive image segmentation via backpropagating refinement scheme [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5297-5306.

[25] Lin Z, Zhang Z, Chen L Z, et al. Interactive image segmentation with first click attention [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 13339-13348.

[26] Yuan Y, Xie J, Chen X, et al. Segfix: Model-agnostic boundary refinement for segmentation [C]//European Conference on Computer Vision, Springer, Cham, 2020: 489-506.

[27] Zhou B, Zhao H, Puig X, et al. Scene parsing through ade20k dataset [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 633-641.

[28] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 3213-3223.

编辑:王谦