PVNet: Pixel-wise Voting Network for 6DoF Pose Estimation

作者：彭思达 学号：21821103

这篇技术报告是对我2019年CVPR的论文"PVNet: Pixel-wise Voting Network for 6DoF Pose Estimation"的评述。

Abstract

本文讨论了在严重遮挡或截断下单个RGB图像对6DoF姿态估计的挑战。最近的许多研究表明，两阶段方法：首先检测关键点，然后解决了姿势估计的Perspective-n-Point（PnP）问题，实现了卓越的性能。然而，这些方法中的大多数仅通过回归其对遮挡和截断敏感的图像坐标或热图来定位一组稀疏关键点。相反，我们引入像素投票网络（PVNet）来回归指向关键点的逐像素向量，并使用这些向量来投票给关键点位置。这为本地化被遮挡或截断的关键点创建了灵活的表示。这种表示的另一个重要特征是它提供了关键点位置的不确定性，可以通过PnP求解器进一步利用。实验表明，所提出的方法大大优于LINEMOD，Occlusion LINEMOD和YCB-Video数据集的现有技术水平，同时对于实时姿态估计是有效的。我们进一步创建一个Truncation LINEMOD数据集来验证我们的截断方法的稳健性。

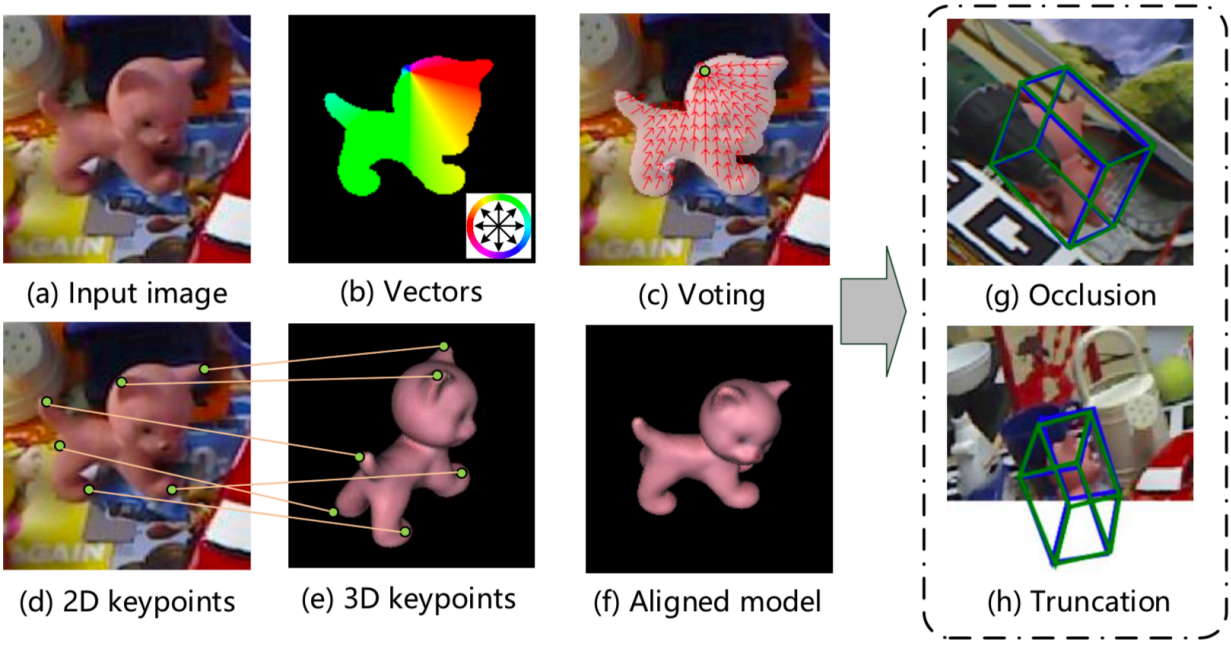
1. Introduction

物体姿态估计旨在检测物体并估计它们相对于标准坐标系的方向和平移[39]。 准确的姿态估计对于各种应用是必不可少的，例如增强现实，自动驾驶和机器人操纵。 例如，快速而稳健的姿态估计在亚马逊挑选挑战中至关重要[6]，其中机器人需要从仓库货架中挑选物品。 本文重点介绍从该对象的单个RGB图像中恢复对象的6DoF姿势（即，3D中的旋转和平移）。 从许多角度来看，这个问题非常具有挑战性，包括严重遮挡下的物体检测，光线和外观的变化以及杂乱的背景物体。

传统方法[24,20,15]已经表明，可以通过建立对象图像和对象模型之间的对应关系来实现姿态估计。 它们依赖于手工制作的功能，这些方法对图像变化和背景杂乱不稳定。 基于深度学习的方法[33,17,40,4]训练端到端神经网络，其将图像作为输入并输出其相应的姿势。 然而，泛化仍然是一个问题，因为不清楚这种端到端方法是否学习了姿势估计的充分特征表示。

最近的一些方法[29,30,36]使用CNN首先回归2D关键点，然后使用Perspective-n-Point（PnP）算法计算6D姿势参数。 换句话说，检测到的关键点用作姿势估计的中间表示。 由于关键点的强大检测，这种两阶段方法实现了最先进的性能。 但是，这些方法在处理遮挡和截断对象时很困难，因为它们的部分关键点是不可见的。 尽管CNN可能通过记忆类似的模式来预测这些看不见的关键点，但泛化仍然很困难。

我们认为解决遮挡和截断需要密集预测，即最终输出或中间表示的像素或图像块估计。 为此，我们提出了一种使用像素投票网络（PVNet）进行6D姿态估计的新颖框架。 基本思想如下图所示.



PVNet不是直接回归关键点的图像坐标，而是预测表示从对象的每个像素朝向关键点的方向的向量。 然后，这些指示基于RANSAC [9]投票选择关键点位置。 这种投票方案的动机来自于刚性物体的特性，一旦我们看到一些局部部分，我们就可以推断出与其他部分的相对方向。

我们的方法实质上为关键点定位创建了矢量场表示。 与基于坐标或热图的表示形成对比，学习这样的表示会强制网络关注对象的局部特征和对象部分之间的空间关系。 因此，可以从可见部分推断出不可见部分的位置。 此外，该矢量场表示能够表示甚至在输入图像之外的对象关键点。 所有这些优点使其成为被遮挡或截断对象的理想表示。 Xiang等人[40]提出了类似的想法来检测对象，在这里我们使用它来定位关键点。

所提出方法的另一个优点是密集输出为PnP求解器提供了丰富的信息，以处理不准确的关键点预测。 具体而言，基于RANSAC的投票可以修剪异常值预测，并为每个关键点提供空间概率分布。 关键点位置的这种不确定性使得PnP求解器能够更自由地识别用于预测最终姿势的一致对应关系。 实验表明，不确定性驱动的PnP算法提高了姿态估计的准确性。

我们在LINEMOD [15]，Occlusion LINEMOD [2]和YCB-Video [40]数据集上评估我们的方法，这些数据集是用于6D姿态估计的广泛使用的基准数据集。 在所有数据集中，PVNet展示了最先进的性能。 我们还展示了我们的方法处理名为Truncation LINEMOD的新数据集上的截断对象的能力，该数据集是通过随机裁剪LINEMOD的图像而创建的。 此外，我们的方法是高效的，在GTX 1080ti GPU上运行25 fps，可以用于实时姿态估计。

2. Related work

给定图像，一些方法旨在估计单次拍摄中物体的3D位置和方向。传统方法主要依赖于模板匹配技术[16,12,14,42]，这些技术对杂乱的环境和外观变化很敏感。最近，CNN显示出对环境变化的显着稳健性。作为先驱，PoseNet [19]引入了CNN架构，可以直接从单个RGB图像中回归6D相机姿势，这一任务类似于物体姿态估计。然而，由于缺乏深度信息和+大搜索空间，直接在3D中定位对象是困难的。为了克服这个问题，PoseCNN [40]定位2D图像中的对象并预测它们的深度以获得3D位置。然而，直接估计3D旋转也是困难的，因为旋转空间的非线性使得CNN不那么普遍。为了避免这个问题，[38,33,23,35]将旋转空间离散化并将3D旋转估计投射到分类任务中。这种离散化会产生粗糙的结果，后期细化对于获得准确的6DoF姿势至关重要。

基于关键点的方法不是直接从图像中获取姿势，而是采用两阶段管道：它们首先预测对象的2D关键点，然后通过与PnP算法的2D-3D对应来计算姿势。 2D关键点检测比3D定位和旋转估计更容易。对于具有丰富纹理的对象，传统方法[24,32,1]可以稳健地检测局部关键点，因此即使在杂乱的场景和严重的遮挡下，也可以高效准确地估计对象姿态。然而，传统方法难以处理无纹理对象和处理低分辨率图像[20]。为了解决这个问题，最近的工作定义了一组语义关键点，并使用CNN作为关键点检测器。 [30]使用分段来识别包含对象的图像区域并从检测到的图像区域中回归关键点。 [36]采用YOLO架构[31]来估计对象关键点。他们的网络基于低分辨率特征图进行预测。当发生全局干扰（例如遮挡）时，特征图会受到干扰[27]并且姿势估计精度下降。受2D人体姿态估计[26]成功的推动，另一类方法[29,27]输出关键点的逐像素热图，以解决遮挡问题。但是，由于热图是固定大小的，因此这些方法难以处理截断的对象，其关键点可能位于输入图像之外。相反，我们的方法使用更灵活的表示（即矢量场）对2D关键点进行逐像素预测。关键点位置由方向投票决定，这适用于截断的对象。

在这些方法中，每个像素或贴片产生对所需输出的预测，然后在广义霍夫投票方案中对最终结果进行投票[22,34,11]。 [2,25]使用随机森林来预测每个像素的3D对象坐标，并使用几何约束产生2D-3D对应假设。为了利用强大的CNN，[18,7]密集地采样图像补丁并使用网络来提取后者投票的特征。但是，这些方法需要RGB-D数据。在仅存在RGB数据的情况下，[3]使用自动上下文回归框架[37]来产生3D对象坐标的逐像素分布。与稀疏关键点相比，对象坐标为姿势估计提供了密集的2D-3D对应，这对于遮挡更加稳健。但由于输出空间较大，回归对象坐标比关键点检测更困难。我们的方法对关键点本地化进行了密集的预测。它可以被视为基于关键点和密集方法的混合，它结合了两种方法的优点。

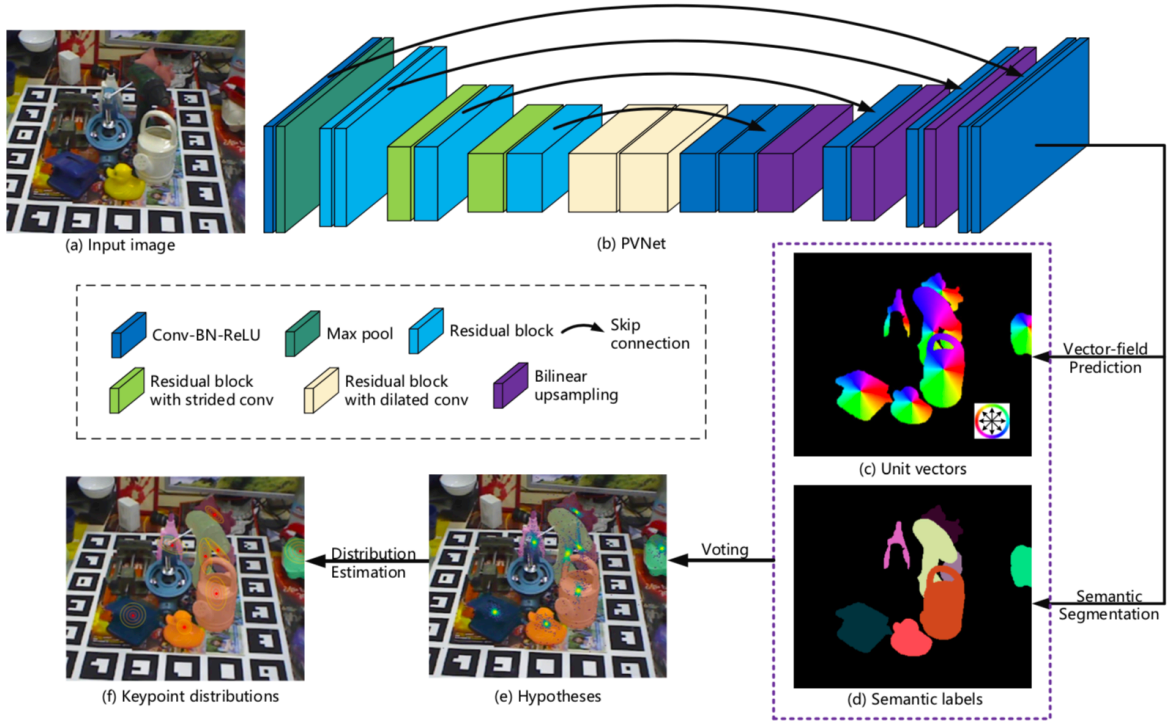
3. Proposed approach

在本文中，我们提出了一个新的6DoF对象姿态估计框架。 给定图像，姿势估计的任务是检测对象并估计它们在3D空间中的方向和平移。 具体地，6D姿势由从对象坐标系到相机坐标系的刚性变换$(R, \mathbf{t})$表示，其中R表示3D旋转，t表示3D平移。

受近期方法[29,30,36]的启发，我们使用两阶段的方法估计对象姿势：我们首先使用CNN检测2D对象关键点，然后使用PnP算法计算6D姿势参数。 我们的创新是2D对象关键点的新表示以及用于姿势估计的修改的PnP算法。 具体来说，我们的方法使用PVNet以类似RANSAC的方式检测2D关键点，它可以有效地处理被遮挡和截断的对象。 基于RANSAC的投票还给出了每个关键点的空间概率分布，允许我们用不确定性驱动的PnP估计6D姿势。

3.1 Voting-based keypoint localization

下图概述了我们定位关键点的方法。



给定RGB图像，PVNet预测逐像素对象标签和向量，表示从每个像素到每个关键点的方向。 给定来自属于该对象的所有像素的某个对象关键点的方向，我们通过基于RANSAC的投票生成该关键点的2D位置的假设以及置信度得分。 基于这些假设，我们估计每个关键点的空间概率分布的均值和协方差。

与从图像窗口[30,36]直接回归关键点位置相比，预测像素方向的任务强制网络更多地关注对象的本地特征并减轻混乱背景的影响。 这种方法的另一个优点是能够表示被遮挡或在图像外的关键点。 即使关键点不可见，也可以根据从对象的其他可见部分估计的方向正确定位关键点。

更具体地说，PVNet执行两个任务：语义分段和矢量场预测。 对于像素$\mathbf{p}$，PVNet输出将其与特定对象相关联的语义标签和表示从像素p到对象的2D关键点$\mathbf{x_k}$的方向的向量$\mathbf{v}_k(\mathbf{p})$。 矢量$\mathbf{v}_k(\mathbf{p})$可以是像素$\mathbf{p}$和关键点$\mathbf{x}_k$之间的偏移，即$\mathbf{x}_k - \mathbf{p}$。 使用语义标签和偏移量，我们获得目标对象像素并添加偏移量以生成一组关键点假设。

然而，这些偏移对物体的尺度变化敏感，这限制了PVNet的泛化能力。 因此，我们进一步提出了尺度不变的向量

$$\mathbf{v}_k(\mathbf{p}) = \frac{\mathbf{x}_k - \mathbf{p}}{\|\mathbf{x}_k - \mathbf{p}\|_2}. \tag{1}$$

它只关心对象部分之间的相对方向。

给定目标对象像素和单位向量，我们在基于RANSAC的投票方案中生成关键点假设。 首先，我们随机选择两个像素，并将它们的向量的交集作为假设$\mathbf{h}_{k,i}$给关键点$\mathbf{x}_k$。 该步骤重复N次以生成一组代表可能的关键点位置的假设$\{\mathbf{h}_{k,i} | i = 1, 2, \ldots, N\}$。 然后，对象的所有像素都投票给这些假设。 具体地，假设$\mathbf{h}_{k,i}$的投票得分$w_{k,i}$被定义为

$$w_{k,i} = \sum_{\mathbf{p} \in O} \mathbb{I}\left(\frac{(\mathbf{h}_{k,i} - \mathbf{p})^T}{\|\mathbf{h}_{k,i} - \mathbf{p}\|_2} \mathbf{v}_k(\mathbf{p}) \geq \theta \right), \tag{2}$$

其中$\mathbb{I}$表示指标函数，θ是阈值（在所有实验中为0.99），$\mathbf{p} \in O$表示像素$\mathbf{p}$属于对象O. 直观地，较高的投票得分意味着假设有更高的置信度，因为它拥有更多的预测方向。

由此产生的假设表征了图像中关键点的空间概率分布。 最后，关键点$\mathbf{x}_k$的平均$\boldsymbol{\mu}_k$和协方差$\boldsymbol{\Sigma}_k$通过以下方式估算：

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} w_{k,i}\mathbf{h}_{k,i}}{\sum_{i=1}^{N} w_{k,i}}, \tag{3}$$

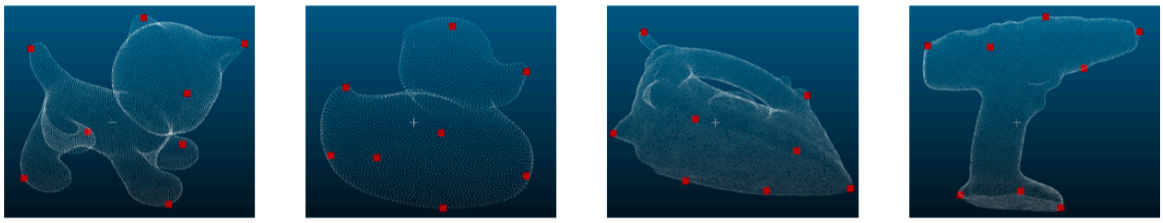
$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^{N} w_{k,i}(\mathbf{h}_{k,i}-\boldsymbol{\mu}_k)(\mathbf{h}_{k,i}-\boldsymbol{\mu}_k)^T}{\sum_{i=1}^{N} w_{k,i}}, \tag{4}$$

用于3.2节中描述的不确定性驱动的PnP。

3.1.1 Keypoint selection

需要基于3D对象模型来定义关键点。 许多最近的方法[30,36,27]使用对象的3D边界框的八个角作为关键点。 一个例子如图3（a）所示。 这些边界框角可能远离图像中的对象像素。 距离对象像素的距离越长，定位误差越大，因为关键点假设是使用从对象像素开始的向量生成的。 图3（b）和（c）分别显示了由我们的PVNet生成的边界框角和在对象表面上选择的关键点的假设。 物体表面上的关键点通常在定位方面具有小得多的变化。

因此，应该在我们的方法中在对象表面上选择关键点。 同时，这些关键点应该在对象上展开，以使PnP算法更加稳定。 考虑到这两个要求，我们使用最远点采样（FPS）算法选择K个kepoints。 首先，我们通过添加对象中心来初始化关键点集。 然后，我们在物体表面上重复找到一个距离当前关键点集最远的点，并将其添加到集合中，直到集合的大小达到K. 我们还使用不同数量的关键点来比较结果。 考虑到准确性和效率，我们根据实验结果建议K = 8。 下图显示了某些对象的选定关键点。



3.1.2 Multiple instances

我们的方法可以根据[40,28]中提出的策略处理多个实例。 对于每个对象类，我们使用我们提出的投票方案生成对象中心的假设及其投票得分。 然后，我们在假设中找到模式，并将这些模式标记为不同实例的中心。 最后，通过将像素分配给他们投票的最近的实例中心来获得实例掩码。

3.2 Uncertainty-driven PnP

给定每个对象的2D关键点位置，可以通过使用现成的PnP求解器解决PnP问题来计算其6D姿势，例如，在许多先前方法中使用的EPnP [21] [36,30]。 然而，他们中的大多数人忽略了这样一个事实，即不同的关键点可能具有不同的置信度和不确定性模式，在解决PnP问题时应该考虑这些因素。

如3.1节所述，我们基于投票的方法估计每个关键点的空间概率分布。 给定$k = 1, \cdots, K$的估计平均$\boldsymbol{\mu}_k$和协方差矩阵$\boldsymbol{\Sigma}_k$，我们通过最小化Mahalanobis距离来计算6D姿势$(R, \mathbf{t})$：

$$\begin{aligned} \underset{R,\mathbf{t}}{\text{minimize}} \sum_{k=1}^{K} (\tilde{\mathbf{x}}_k - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\tilde{\mathbf{x}}_k - \boldsymbol{\mu}_k), \\ \tilde{\mathbf{x}}_k = \pi(R\mathbf{X}_k + \mathbf{t}), \end{aligned} \tag{5}$$

其中$\mathbf{X}_k$是关键点的3D坐标，$\tilde{\mathbf{x}}_k$是$\mathbf{X}_k$的2D投影，π是透视投影函数。 参数R和$\mathbf{t}$由EPnP [21]基于四个关键点初始化，其协方差矩阵具有最小的轨迹。 然后，我们使用Levenberg-Marquardt算法求解（5）。 在[8]中，作者还通过最小化近似Sampson误差来考虑特征不确定性。 在我们的方法中，我们直接最小化重投影错误。

4. Experiment

我们的方法在LINEMOD，Occlusion LINEMOD，Truncaiton LINEMOD和YCB-Video数据集上测试。

LINEMOD上2D projection metric的结果

	w/o refinement			w/ refinement
methods	BB8 [30]	Tekin [36]	OURS	BB8 [30]
ape	95.3	92.10	**99.23**	96.6
benchwise	80.0	95.06	**99.81**	90.1
cam	80.9	93.24	**99.21**	86.0
can	84.1	97.44	**99.90**	91.2
cat	97.0	97.41	**99.30**	98.8
driller	74.1	79.41	**96.92**	80.9
duck	81.2	94.65	**98.02**	92.2
eggbox	87.9	90.33	**99.34**	91.0
glue	89.0	96.53	**98.45**	92.3
holepuncher	90.5	92.86	**100.0**	95.3
iron	78.9	82.94	**99.18**	84.8
lamp	74.4	76.87	**98.27**	75.8
phone	77.6	86.07	**99.42**	85.3
average	83.9	90.37	**99.00**	89.3

Occlusion LINEMOD上2D projection metric的结果

methods	Tekin [36]	PoseCNN [40]	Oberweger [27]	OURS
ape	7.01	34.6	**69.6**	69.14
can	11.20	15.1	82.6	**86.09**
cat	3.62	10.4	65.1	**65.12**
duck	5.07	31.8	61.4	**61.44**
driller	1.40	7.4	**73.8**	73.06
eggbox	-	1.9	**13.1**	8.43
glue	4.70	13.8	54.9	**55.37**
holepuncher	8.26	23.1	66.4	**69.84**
average	6.16	17.2	60.9	**61.06**

Truncation LINEMOD上2D projeciton和ADD(-S)的结果

objects	ape	benc-hvise	cam	can	cat	driller	duck
2D Projection	52.59	58.19	54.87	57.44	61.66	43.27	54.23
ADD(-S)	12.78	42.80	27.73	32.94	25.19	37.04	12.36
objects	eggbox	glue	holep-uncher	iron	lamp	phone	avg
2D Projection	87.23	86.64	53.84	46.53	46.94	51.35	58.06
ADD(-S)	44.13	38.11	22.39	42.01	40.91	30.86	31.48

YCB-Video上2D projection和ADD(-S) AUC的结果

methods	PoseCNN [40]	Oberweger [27]	OURS
2D Projection	3.72	39.4	**47.4**
ADD(-S) AUC	61.0	72.8	**73.4**

References

[1] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006. 2

[2] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6d object pose estimation using 3d object coordinates. In *ECCV*, 2014.

[3] E. Brachmann, F. Michel, A. Krull, M. Ying Yang, S. Gumhold, et al. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *CVPR*, 2016.

[4] M. Bui, S. Zakharov, S. Albarqouni, S. Ilic, and N. Navab. When regression meets manifold learning for object recog- nition and pose estimation. In *ICRA*, 2018.

[5] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The ycb object and model set: Towards com- mon benchmarks for manipulation research. In *ICRA*, 2015.

[6] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman. Analysis and observations from the first amazon picking challenge. *IEEE Transactions on Automation Sci- ence and Engineering*, 15(1):172–188, 2018.

[7] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.- K. Kim. Recovering 6d object pose and predicting next-best- view in the crowd. In *CVPR*, 2016.

[8] L. Ferraz, X. Binefa, and F. Moreno-Noguer. Leveraging feature uncertainty in the pnp problem. In *BMVC*, 2014.

[9] M. A. Fischler and R. C. Bolles. Random sample consen- sus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[10] R. Girshick. Fast r-cnn. In *ICCV*, 2015.

[11] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and

pose estimation. In *ICCV*, 2011.

[12] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[14] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit. Gradient response maps for real- time detection of textureless objects. *T-PAMI*, 34(5):876– 888, 2012.

[15] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. Model based training, detec- tion and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV*, 2012.

[16] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the hausdorff distance. *T-PAMI*, 15(9):850–863, 1993.

[17] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *ICCV*, 2017.

[18] W. Kehl, F. Milletari, F. Tombari, S. Ilic, and N. Navab. Deep learning of local rgb-d patches for 3d object detection and 6d pose estimation. In *ECCV*, 2016.

[19] A. Kendall, M. Grimes, and R. Cipolla. Posenet: A convolu- tional network for real-time 6-dof camera relocalization. In *ICCV*, 2015.

[20] V. Lepetit, P. Fua, et al. Monocular model-based 3d track- ing of rigid objects: A survey. *Foundations and Trends in Computer Graphics and Vision*, 1(1):1–89, 2005.

[21] V.Lepetit,F.Moreno-Noguer,andP.Fua.Epnp:Anaccurate o (n) solution to the pnp problem. *IJCV*, 81(2):155, 2009.

[22] J. Liebelt, C. Schmid, and K. Schertler. independent object class detection using 3d feature maps. In *CVPR*, 2008.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[24] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[25] F.Michel,A.Kirillov,E.Brachmann,A.Krull,S.Gumhold, B. Savchynskyy, and C. Rother. Global hypothesis genera- tion for 6d object pose estimation. In *CVPR*, 2017.

[26] A. Newell, K. Yang, and J. Deng. Stacked hourglass net- works for human pose estimation. In *ECCV*, 2016.

[27] M. Oberweger, M. Rad, and V. Lepetit. Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In *ECCV*, 2018.

[28] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy. Personlab: Person pose estimation and in- stance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, 2018.

[29] G.Pavlakos,X.Zhou,A.Chan,K.G.Derpanis,andK.Dani- ilidis. 6-dof object pose from semantic keypoints. In *ICRA*, 2017.

[30] M. Rad and V. Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *ICCV*, 2017.

[31] J.RedmonandA.Farhadi.Yolo9000:better,faster,stronger. In *CVPR*, 2017.

[32] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *IJCV*, 66(3):231–259, 2006.

[33] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with ren- dered 3d model views. In *ICCV*, 2015.

[34] M. Sun, G. Bradski, B.-X. Xu, and S. Savarese. Depth- encoded hough voting for joint object detection and shape recovery. In *ECCV*, 2010.

[35] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *ECCV*, 2018.

[36] B. Tekin, S. N. Sinha, and P. Fua. Real-time seamless single shot 6d object pose prediction. In *CVPR*, 2018.

[37] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *T-PAMI*, 32(10):1744–1757, 2010.

[38] S. Tulsiani and J. Malik. Viewpoints and keypoints. In *CVPR*, 2015.

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

[40] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018.

[41] F. Yu and V. Koltun. Multi-scale context aggregation by di- lated convolutions. In *ICLR*, 2016.

[42] M.Zhu,K.G.Derpanis,Y.Yang,S.Brahmbhatt,M.Zhang, C. Phillips, M. Lecce, and K. Daniilidis. Single image 3d object detection and pose estimation for grasping. In *ICRA*, 2014.