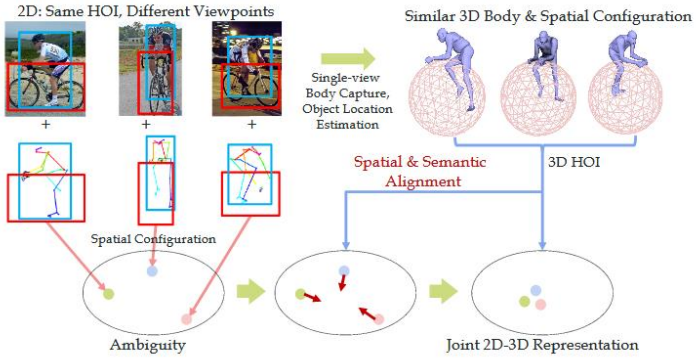
人-物交互的含细节二维-三维联合表示

问题。Human – Object Interaction（HOI，此处译为人-物交互[1]）问题是视觉关系问题的一个子任务，旨在定位活动的人类和对象并同时推断人正在做的动作。HOI可以提高对人活动的认知能力、并可以应用于模仿学习。

具体而言，模型接受一张或多张二维图片，推断其中人和人正在交互的物体的位置、人的动作以及物体的类别。

为什么需要二维-三维联合表示。基于二维的 HOI 检测可能在不同视角下具有二义性。三维下的HOI表示鲁棒性更佳。但是三维下重建出的粗糙的人体姿态不具有重建出检测动作必须的细节。因而需要二维-三维的联合表示。



前人的工作。

人-物交互检测。近几年在人-物交互的检测中有了重大进展。为了推动这一领域机器学习的发展，人们制作和发布了大规模的数据集。同时人们提出了大量基于深度学习的方法。Chao等人提出了一个多数据流的模型框架(HO – RCNN)；未采用直接视觉方法的GPNN是一种基于图的模型，使用消息传递的模型来处理图像和视频上的人-物交互检测；Gkioxari等人采用了动作密度分布图来估计交互物体的二维位置；iCAN利用了自我注意力机制将上下文与人-物交互相关联；TIN 是一个显式的交互度学习模型，它可以识别不是人-物交互焦点的人和物体，并在网络推导时减少这些物体的影响；HAKE提出了一种基于人的身体部分的状态的新的学习范式。这些方法主要依赖于视觉表现和人-物空间上的相对位置关系。也有一些利用了二维姿态估计。但是二维下不同视角下的二义性并未在HOI问题下被详细探究。

基于三维姿态的动作识别。近几年来，基于深度学习的三维姿态估计方法有了显著的进步。在上述提到过的基于二维姿态估计的动作识别之外，很多工作也引入了三维的人体姿态。Yao等人由二维外观和三维姿态重建了 2.5D 的场景图，并从中选取不同动作下典型的图，对动作进行进行分类识别。在Discovering object functionality一文中，二维的姿态被映射到三维的姿态，随后通过对三维姿态的相似性比较，对动作进行分类。Luvizon等人估计了二维/三维的姿态并使用一个统一的模型进行图像和视频的动作识别。Wang等人利用RGB – D数据获取三维人体关节位置，并采用了一种动作组件组合的方法进行 HOI 学习。最近Pham等人提出了一种多任务模型来同时从视频数据进行三维姿态识别和动作识别。大多数基于三维姿态的方法使用一种基于RNN的框架进行时空上的动作识别，但很少有人聚焦于从单张RGB

[1] 未找到被广泛认可的译名，通常即用 HOI 代称。

图片进行HOI识别。

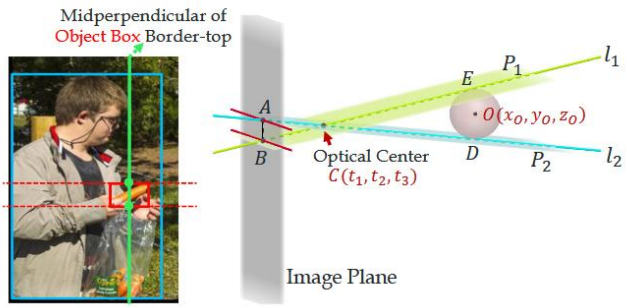
单视角三维体态恢复。在深度学习和大规模扫描的三维人体数据库的帮助下，我们已经能够从单张RGB图片直接恢复出三维的身体模型和姿态。SMPLify－X是一个有效的从二维身体、手和脸的姿态捕捉相应三维姿态的模型。为了获得更加准确和实际的人物体型，SMPLify－X利用在MoCap数据集上训练的VPoser模型。这个模型具有大量的人体姿态知识。这个模型让我们从 HOI 图片恢复出详细的三维人物姿态，并嵌入更多的人体姿态知识。

方法。

我们的目标是学习二维-三维联合的 HOI 表示。因而我们需要先得出在三维空间中表示 HOI 的方法。我们先在一张静态图片上应用物体识别和姿态估计来得出二维的物体包围盒和人物姿态；随后我们使用三维人体捕捉来从上述二维检测结果估计三维的人体姿态，并在三维空间中估计物体的大小和位置来重建三维空间的形态体积。

单视角三维体态捕捉。粗糙的三维姿态不足以区分各种动作，尤其是与日常物体的复杂交互。因而我们需要更加精细的三维人体信息作为线索。在此我们采用了一种功能整体性的三维身体捕捉方法来从单张 RGB 图片重构精细的三维人体。给定了图像 I 的二维检测结果（例如，二维的人和物的包围盒b_h, b_0，二维的人体姿态$\theta^{2D} = \{\theta_b^{2D}, \theta_f^{2D}, \theta_h^{2D}\}$其中 b 代表身体，f 代表脸部的骨骼节点，h 代表手部的手指骨骼节点。我们将其输入SMPLify－X中来恢复出三维的人体框架，并根据结果重建三维人体模型。

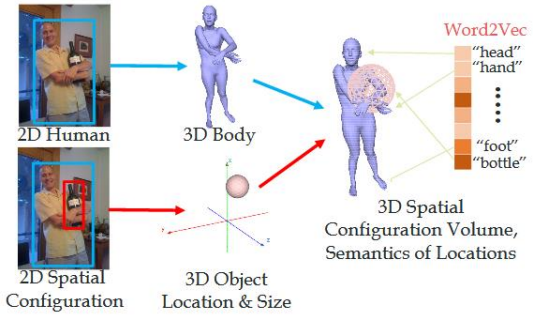
三维形态体积的获取。在获得了三维的人体模型后，我们在三维空间中进一步表示人-物交互信息，即估计三维物体的位置和大小。为了稳定性和高效性，我们不重构物体的形状，而是用一个空心球体去代表物体。这样我们也避免了单视角下困难的六维姿态估计问题。我们采用SMPLify－X估计产生的相机参数，固定焦距f为 5000，并忽略相机导致的变形等影响。我们通过二维图片在图片平面上的位置将球限制在两个视角投影平面之间。随后我们通过对COCO数据集中 80 种物体的大致大小进行手动标注，确定球的半径大小，从而利用与投影平面相切的性质最终求出球体的位置和大小。在标注过程中，我们为较小的物体或者在各个维度上长度相似的物体标注了球半径和人肩宽的比；对严重受到旋转影响的物体（例如船，雪橇）我们用人和物体包围盒的相对比例为参照。另外，由于球的深度对半径大小极为敏感，我们在此基础上加上球的深度限制，即球与人的包围盒相交。最后我们以人体骨盆为坐标原点，两肩连线为 x 轴，重力为 z 后建立坐标系，并在三维人体和球体上分别采样 916 和 312 个点，最后以瞳距为坐标单位对场景进行归一化，得到最终三维形态体积。



二维-三维联合学习。我们提出了含细节的联合表示网络(Detailed Joint Representation Network DJ − RN)。它包括两个模块，二维表示网络(2D − RN)和三维表示网络(3D − RN)。我们用这两个网络来分别提取特征，并将二维空间和三维空间的特征进行对齐，并用身体部分注意力的一致性和语义一致性来指导学习。

二维特征提取。我们用在COCO数据集上预训练的Faster − RCNN模型和iCAN区块来获取二维的人体特征f_H^{2D}和物体特征f_O^{2D}；另外，尽管外观携带了重要的信息，它也引入了噪声和从某些角度观察导致的误导性的模式。因而额外地，人-物空间形态可以用于区分各种特征。空间数据流主要考虑二维人-物交互相对位置。我们将二维姿态图和空间图输入空间区块(空间区块由卷积和全连接层组成）提取出空间特征f_{sp}^{2D}。空间图包括了人和物体的两个通道，均由人和物体的包围盒生成，其大小均为64×64。在盒子内则值为1否则为0。姿态图包含了由OpenPose计算而来的 17 个关节的64×64热度图。

三维特征提取。体积区块。在三维空间数据流中，我们采用PointNet网络来提取三维空间的特征f_{sp}^{3D}。首先我们预训练这个网络从前述采样点云中分离人和物体的点。从而它可以学习区分人体和物体的几何性态差异。为了进一步在其中嵌入三维位置的语义信息，我们将空间特征与相应的语义信息（例如Word2Vec Embedding后对应物体和身体部分的词向量)。我们首先将点云划分为 18 类：17 个人体部分类和 1 个物体空心球类。随后，我们将部分类的特征和经过主成分分析(PCA)削减后的词向量连接起来，得到f_{sp}^{3D}如图。



二维-三维空间对齐。由于二维空间的特征缺乏鲁棒性，我们提出将二维的特征对齐到三维坐标体系下。对于前述的f_{sp}^{2D}，我们从三维空间特征的训练集中随机选取一个有相同 HOI 标签的正样本f_{sp+}^{3D}和有具有不相交标签的(因为一个人可能同时做多件事因而有多个标签)负样本f_{sp-}^{3D}。对于一个人-物对，我们用triplet loss来对齐他们的二维空间特征，即

$$L_{tri} = \left[d\left(f_{sp}^{2d}, f_{sp+}^{3D}\right) - d\left(f_{sp}^{2D}, f_{sp-}^{3D}\right) + \alpha\right]_+$$

其中$d(\cdot)$表示欧式距离，$\alpha = 0.5$是边界值。这一损失函数将会使具有相同 HOI 但是不同空间形态的二维样本在空间中聚集。

联合的人体注意力估计方法。

二维。我们将前述的三类（人，物，空间）特征连接得到二维的全体特征，随后使用全局平均池化的方式得到全局的特征向量f_g^{2D}。将其与二维全体特征求内积得出二维的注意力图$att^{2D} = Softmax\left(f_g^{2D} \cdot f^{2D}\right)$。随后对于身体部分，我们通过加权平均的方式额外对周围的点提升一定的注意力。

$$a_i^{2D} = \frac{\sum_{u,v} att_{u,v}^{2D} / (1 + d[(u,v),(u_i,v_i)])}{\sum_{u,v} 1 / (1 + d[(u,v),(u_i,v_i)])}$$

并对其进行归一化处理得到最终每点的二维注意力。

三维。输入$f_{sp}^{3D}:[1228,384]; f_H^{3D}:[1024]$。利用广播将$f_H^{3D}:[1024] \xrightarrow{tile} [1228,1024]$并连接得到$f^{3D}:[1228,1408]$. 应用全局平均池化，并通过两层大小为 512 的全连接网络，再经Softmax得到 17 个关节的注意力。

注意力一致性。我们在二维和三维的各个关节的注意力上计算$KL-Divergence$来得到一项损失函数。这可以使两个网络生成相似的部分重要性，从而实现注意力机制的一致性。随后我们在二维和三维的特征上应用注意力得到加权后的特征。

语义。在提取出特征后，我们将其连接至两层全连接网络并通过Sigmoid激活，得到HOI分类结果。为了二维和三维的语义具有一致性，我们在其上加上了L2正则化损失。最后将得到的语义结果相加，得到最终HOI各类的分数并应用多标签交叉熵作为损失函数。

实验。我们在$HICO-DET$数据集上进行了验证，并且在测试集中手动挑选了一个二义性较高的子集作为新的Benchmark目标$Ambiguous-HOI$，使用mAP作为测试指标。

	Default			Known Object		
Method	Full	Rare	Non-Rare	Full	Rare	Non-Rare
Shen *et al.* [56]	6.46	4.24	7.12	-	-	-
HO-RCNN [9]	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [19]	9.94	7.16	10.77	-	-	-
GPNN [51]	13.11	9.34	14.23	-	-	-
iCAN [17]	14.84	10.45	16.15	16.26	11.33	17.73
Interactiveness [31]	17.03	13.42	18.11	19.17	15.51	20.26
No-Frills [21]	17.18	12.17	18.68	-	-	-
PMFNet [58]	17.46	15.65	18.00	20.34	17.47	21.20
Julia *et al.* [48]	19.40	14.60	20.90	-	-	-
$\mathcal{S}^{2D}$	19.98	16.97	20.88	22.56	19.48	23.48
$\mathcal{S}^{3D}$	12.41	13.08	12.21	16.95	17.74	16.72
$\mathcal{S}^{Joint}$	20.61	17.01	21.69	23.21	19.66	24.28
DJ-RN	**21.34**	**18.53**	**22.18**	**23.69**	**20.64**	**24.60**

Table 1. Results comparison on HICO-DET [9].

Method	Ambiguious-HOI
iCAN [17]	8.14
Interactiveness [31]	8.22
Julia *et al.* [48]	9.72
DJ-RN	**10.37**

Table 2. Results comparison on Ambiguous-HOI.

上述分别为$HICO-DET$和$Ambiguous-HOI$上的结果。可以看到本文提出的模型达到了目前的State of the art。