参赛队员姓名：王肃羽 林语瑭 田明昊

中学：北京一零一中学

省份：北京市

国家/地区：中国

指导教师姓名：王亦洲 周宇辰

指导教师单位：北京大学 北京市一 0 一中学

论文题目：寻找失去的记忆—基于视觉时空推理的阿尔茨海默症老年患者智能寻物系统

摘要

当前全世界人口老龄化严重的情况下，为了让老年人的晚年生活更健康、安全、便利，用人工智能技术为老年人提供服务显得至关重要。阿尔茨海默症引起的记忆障碍严重影响老人，尤其是无人看护时段、独居和空巢老人的日常生活、健康甚至安全。例如，在发生心绞痛、呼吸困难等突发症状的情况下，因记忆障碍无法及时找到硝酸甘油、哮喘喷雾、制氧剂等急救药物和器械，会错失转瞬即逝的急救时间窗口，威胁老人生命安全。 现有的基于无线物联网和计算机视觉的寻物技术存在诸如无法解决物体遮挡干扰、寻物时间跨度小、部署和使用复杂等问题。我们提出并实现了一种新型的基于视觉时空推理的智能寻物方法。该方法首先利用室内部署的摄像头，通过基于深度神经网络的小物体目标检测、3D场景理解与重建、人体追踪与动作识别等技术构建典型室内家居场景多维度时空因果场景图，动态追踪和记录环境物体（如具有存放功能的家具）、目标物体（如小药瓶等）和人之间的关系和变化。以此为基础，融合人-物交互和、环境物体功能性和目标物体可见性等信息，通过概率逻辑推理技术，在时空因果场景图上实现了目标物体不可见状态下的位置推理。实验证明，与基于传统计算机视觉的寻物方法相比，本项目提出的方法能够在较长时间跨度情况下和有干扰的复杂场景中，高效准确地寻找场景中目标物体的位置，从而有效地帮助阿尔茨海默症老年患者解决因记忆障碍引发的生活困扰。

关键词：视觉时空推理，时空因果场景图，动作识别，人-物交互，阿尔茨海默症

目录

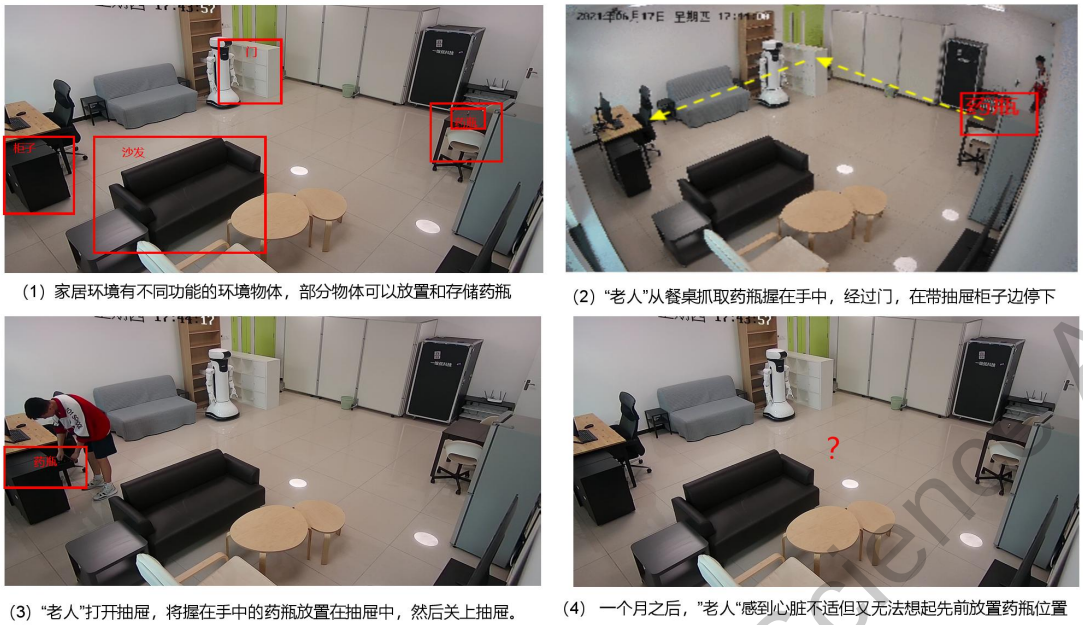
1. 引言

当前全世界人口老龄化严重。联合国预期二十一世纪人口老龄化比率会超过上一世纪。自 1950 年来，年过 60 岁的人数增加三倍，达 2000 年的 6 亿，在 2006 年超过 7 亿。预期到 2050 年，老龄化人口会达到 21 亿[39-40]。老龄化问题在中国尤为尖锐。2016 年底，中国 60 岁老龄人口已达 2.3 亿，占人口的 16.7%。其中 65 岁人口 1.5 亿，占总人口的 10.8%[41]。预计 2025 年，中国社会中 60 以上人口达到 3 亿[42]。人口老龄化将减少劳动力的供给数量、增加家庭养老负担和基本公共服务供给的压力。 为了让老年人的晚年生活更健康、安全、便利，用人工智能技术为老年人提供服务显得至关重要。

阿尔茨海默症（Alzheimer’s Disease，AD）是西方世界第一大神经退行性疾病[1]。 这是一种起病隐匿的进行性发展的神经系统退行性疾病，此病占了失智症中六到七成的成因[2-3]。2015 年，全球大约有 2980 万人患有阿尔茨海默病[3, 7]。 患者的发病年龄一般在 65 岁以上。在发达国家中，它是耗费最多社会资源的一种疾病[11-12]。 其症状可以影响大部分复杂的日常生活活动[15]，最明显的症状就是记忆障碍，主要是以难以记住最近发生的事情和无法吸收新资讯[14, 16] 。

以阿尔茨海默病为代表的记忆障碍严重影响老人日常生活、健康甚至安全，尤其是在无人看护时段和对独居和空巢老人。无法找到眼镜、钥匙、手机、身份证、银行卡、遥控器等重要物品会严重影响老人的正常生活，带来诸多不便。而无法及时找到药物和医疗用品（如血氧仪、血糖仪、血压计、制氧剂、哮喘喷雾剂等），则是对老年人的健康和生命安全产生影响。轻则会使老人无法按时服药、检测，对于糖尿病、高血压、心血管疾病等慢性病的持续治疗和指标控制产生严重影响；重则在产生心绞痛、呼吸困难等突发症状的情况下，无法及时找到硝酸甘油、哮喘喷雾、制氧剂等急救药物和器械，会错失转瞬即逝的急救时间窗口，威胁老人生命安全。

现有老人辅助寻物系统大多以无线物联网技术为主，但这类系统需要对目标物体附加定位设备，部署操作复杂，无法覆盖所有可能目标物品，而且随着动态增加的物品会使得其成本持续增加。现有计算机视觉技术如目标检测和基于目标



(1) 家居环境有不同功能的环境物体，部分物体可以放置和存储药瓶

(2) “老人”从餐桌抓取药瓶握在手中，经过门，在带抽屉柜子边停下

(3) “老人”打开抽屉，将握在手中的药瓶放置在抽屉中，然后关上抽屉。

(4) 一个月之后，“老人”感到心脏不适但又无法想起先前放置药瓶位置

图 1　目标场景示例

检测的追踪技术只能实现在较短时间的连续视频中对不被遮挡的可见物体进行追踪，无法处理因遮挡、覆盖或收纳而不可见复杂场景，并且无法支持如数周或数月的长时间跨度物体位置查询，应用场景受到极大限制。

针对上述问题，面向以图 1 为例的阿尔茨海默症辅助室内寻物场景，本项目设计和实现了一种基于视觉时空推理技术的阿尔茨海默症老年患者智能寻物系统。利用室内部署的摄像头，通过目标检测识别和 3D 重建等技术构建典型室内家居场景动态时空因果场景图，以人体动作识别、人-物交互和环境物体功能性建模为基础，通过概率逻辑推理技术实现了目标物体不可见状态下的位置推理。主要包括以下技术：

- **时空因果场景图**

时空因果场景图以时间序列方式识别和记录场景中目标物体状态的动态变化情况，包括空间位置变换、物体的开合状态变化等。本项目将场景物体划分为具有特定功能属性的环境物体（如柜子、桌子、沙发等）和需要追踪的目标物体（如药瓶），并分别对其进行定位。对环境物体开发了基于 2D 目标检测的 3D 场景理解和重建技术，支持在不同数量摄像机的应用场景中识别物体位置及其尺寸；对于目标物体，针对尺寸小不易检测的特点，筛选和优化识别算法。以此为基础构建的时空因果场景图支持长时间跨度的物体寻找，避免了进行回溯查找，

有效的避免了传统视觉方法存储空间大和时间跨度长的问题，如有效查找 12 小时前服用的药品位置，甚至常年不使用的非常用应急用品（如急救药品和用具等）。

- **不可见物体位置推理**

人-物交互是物体可见状态不连续情况下推理目标物体位置的主要线索。针对人体动作和环境物体的多样性，通过结合基于深度学习的人体动作识别结果、物体功能属性、距离等多维信息，不仅定性地推理出同一场景中人-物交互可能的多种语义，并同时计算其概率，为目标位置推理提供支持。物体位置推理基于人-物交互推理结果，以及时空因果场景图中随时间变化的目标和环境物体当前和前序状态，根据规则推理出目标物体不可见状态下可能的位置。例如，当人以较大概率执行抓取动作后，目标物体可能处在人的手中或口袋中并随这人移动；当人以较大概率执行放置动作后，目标物体处于周围距离最近的且具有存放功能的环境物体中（如柜子和抽屉等），并给出其概率值。该技术以推理方式实现了当前视觉检测和追踪技术在视线被部分遮挡或物体位置隐藏的情况下无法有效识别位置的缺陷，通过推断潜在隐藏目标物体，提高寻物准确性和场景适应性。

本项目提出的智能寻物系统结构如图 2 所示。时空因果场景图用于记录基于深度学习的 3D 场景理解与重建、目标物体检测、人体动作识别结果。目标位置推理根据场景图中不同时间切片中动作与状态变化的时序与因果关系，推理出不可见状态下目标物体位置，并进一步通过可视化等方式支持长时间跨度的目标物

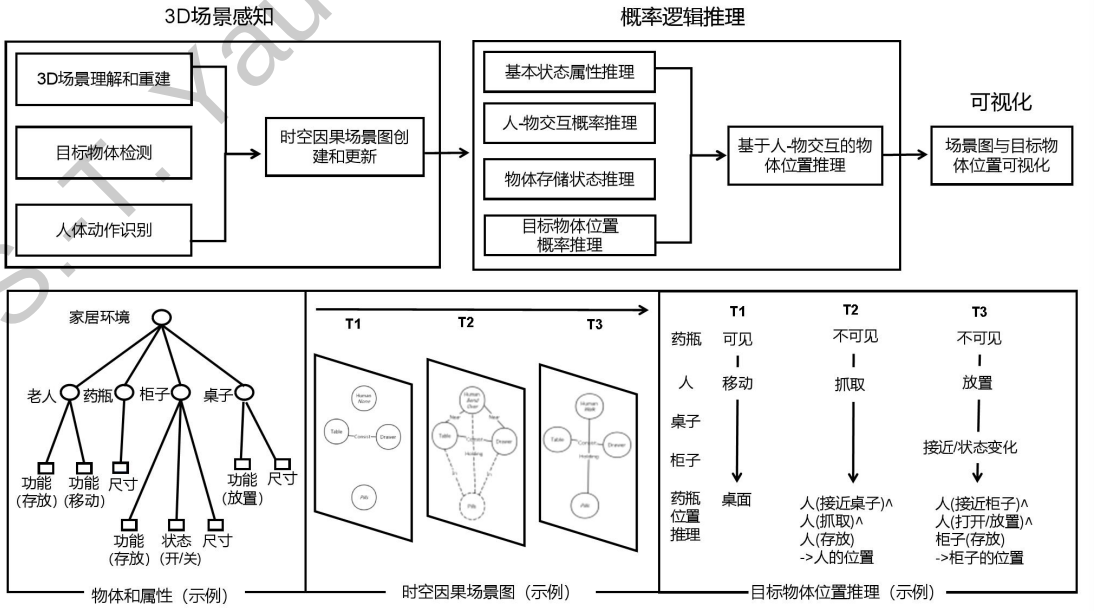


图 2 本项目提出的智能寻物系统结构

体位置查询。考虑家居环境视频可能涉及个人隐私，涉及视频内容的检测与识别的模块可以部署于家居本地，而只包含抽象数据的场景图和推理等模块可以根据需要部署进行本地或远程部署。

2. 相关研究工作

无线物联网是目前老人辅助寻物系统普遍采用的技术。早期的研究人员尝试通过射频识别（Radio Frequency Identification, RFID）、近场通讯（Near Field Communication, NFC）等技术实现物体定位和检测[37-38]。由电子标签、读写器等元件组成的 RFID 系统可以实现便携、易于操作且成本较低的物体检测系统，但 RFID 系统会受到多路效应、不稳定的接收信号强度等技术问题的影响[37]，并且当场景中物体的数量较大时，为每个目标物体均附着电子标签不具备可操作性，尤其是在长期使用过程中持续添加新的目标物体的真实场景。同时系统部署的复杂性和价格也影响了这类系统的普及。因此，研究人员开始寻找能够自动化地检测大量物体、简化更新的解决方案。

计算机视觉是目前主流的物体检测、追踪手段之一。随着深度学习和神经网络的发展，研究人员近年来已经在诸如物体检测[17-21]、动作识别[22-24]、人体姿态估计[25-27]等领域取得较大进展[28-29]。然而，尽管在大量训练数据与 YOLO[45]、Faster R-CNN[46]等物体识别网络的帮助下，人们能够在上述任务中基于 COCO 等通用数据集实现较高的准确率，但这些模型却无法有效地学习、利用时间跨度较长的数据信息和深层次的空间关系[31]。更具体地，简单的物体检测模型难以编码物体之间的相互关系，无法给出存在于摄像设备视野之外的物体的相关信息，从而也无法判断是否存在被遮挡的物体，以及获取被遮挡物体的位置信息。除此之外，以上的计算机视觉模型本身无法全面利用当前时刻之前和之后的视频信息，并且难以快速处理时间跨度较长的视频流。如果利用这一类算法模型实现在一段较长的时间内追踪某个物体的（受人类活动影响而产生的）行动轨迹，那么模型必须从最末时刻开始，逐帧向前回溯，并在每一帧中执行一次物体检测、人体动作以及人-物交互，导致在时间流中寻找物体的速度被大大降低。

为了记录某一特定场景中物体间的相互关系，进而学习更深层的空间视觉信息，研究人员提出了场景图（Scene Graph）的概念，即一种编码了某场景中的

物体实体、物体属性和物体间相互关系的数据结构[30，32]。目前大部分工作集中在图像的场景图构建，部分时序视频流处理技术也是以构建静态场景图为目标[34]，一些研究通过与或图的方式描述场景图中物体的状态[57]。；而目现有工作大多用以描述物体间的高层次抽象关系，未发现精确定位场景图中不同物体动态关系和 3D 距离的实用性实现。

感知与推理都是人工智能中的关键研究问题，当前以深度学习为代表的统计机器学习极大地推动了感知能力发展，然而其推理能力却极为有限，很多应用在试图突破感知瓶颈的同时却忽略了认知推理的重要作用。传统基于规则的推理在感知能力不足的前提下显得极为脆弱，当前结合神经网络感知与符号逻辑推理的研究工作方兴未艾，即所谓的神经符号推理（Neural-Symbolic Reasoning），这种感知与推理融合的技术能够很好地解决传统的基于规则推理系统的问题[55,58]。

针对现有系统在长时场景寻物任务中存在的问题，本项目结合深度神经网络感知和符号逻辑推理提出了一个鲁棒的智能寻物系统，该系统能够在长时环境下对场景目标状态进行推理，实现因遮挡等因素造成的不可见目标的有效追踪。

3. 时空因果场景图

时空因果场景图对实验环境中的环境物体、目标物体和人以及之间的位置关系进行识别，并对实验环境的变化进行动态识别和记录。它是目标物体不可见状态下位置推理的基础，也是大时间跨度下目标物体位置查询的基础。

3.1 时空因果场景图的定义

下面用一个典型室内家居场景来阐述时空因果场景图的定义。如图 3 所示是该场景内的物体关系及其属性，譬如家居环境中有老人、药瓶、门、餐桌、滑轮柜、沙发、茶几等，在此家居环境中，需要规定不同物体各自具有的多维属性与特征，包括 3D 空间位置（x，y，z）、存放功能（有/无）、置放功能（有/无）、可见性（能/否）、状态（开/关/无）和尺寸（长，宽，高）等，然后通过物体识别和定位算法将物体类别和其他属性从场景里面提取并储存。

考虑一个已知场景 $\mathbb{E}$ ，对于其中的物体 $O_i \in \mathbb{E}$ ，有如下属性：

空间位置：$L_i = O_{i,loc}$

尺寸大小：$S_i = O_{i,size}$

功能属性：$\mathbb{F}_i = func_i = O_{i,func}$

其他状态：$\mathbb{S}_i = state_i = O_{i,state}$

对于 $\mathbb{F}$ 与 $\mathbb{S}$,具体功能和状态有：

开关、存放、置放：$\mathbb{F}_U = \{F_{switch}, F_{store}, F_{place}, F_{take}\}$

开、关等其他状态：$\mathbb{S}_U = \{S_{other}, S_{open}, S_{close}\}$

对于物体间的关系 $\mathbb{R}$,本项目主要关注储存关系(Contain)、远近关系(Near)等其他关系：

$$\mathbb{R}_U = \{R_{other}, R_{contain}\}$$

在目标场景中，患有阿尔茨海默症的老年人家居环境中拿药、放药瓶并最终忘记药瓶所在位置的情景。该过程中做出“拿”，“放”，“打开”，“关闭”等一系列动作，如打开抽屉，把药瓶放到衣服兜里，从柜子里拿出药瓶等。同时

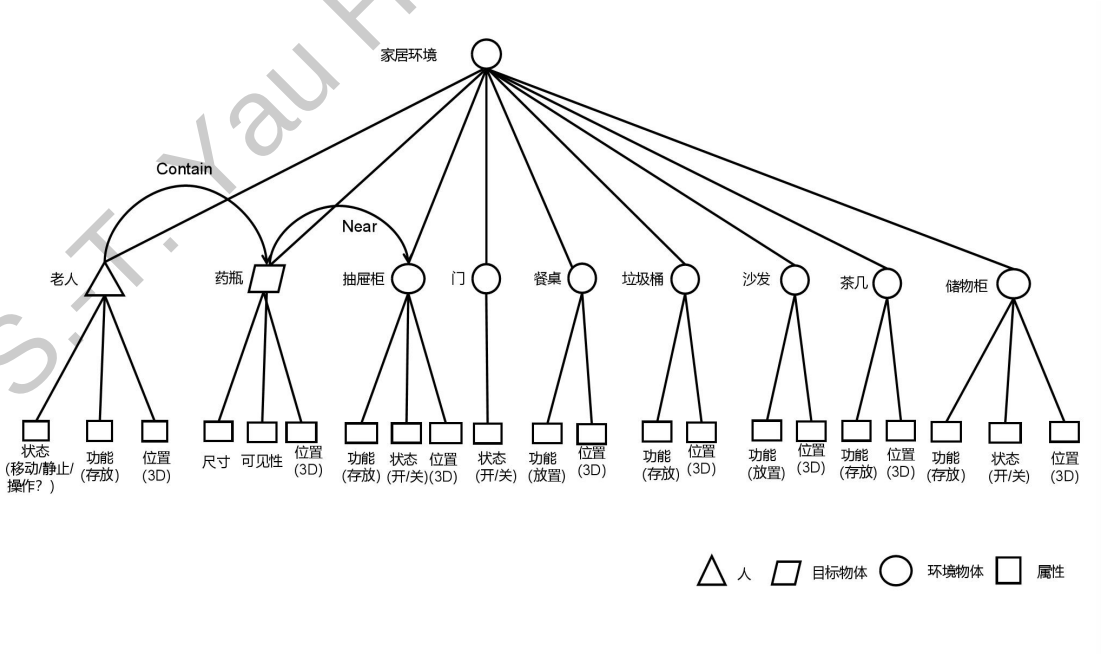


图 3 时空因果场景图中的物体和属性

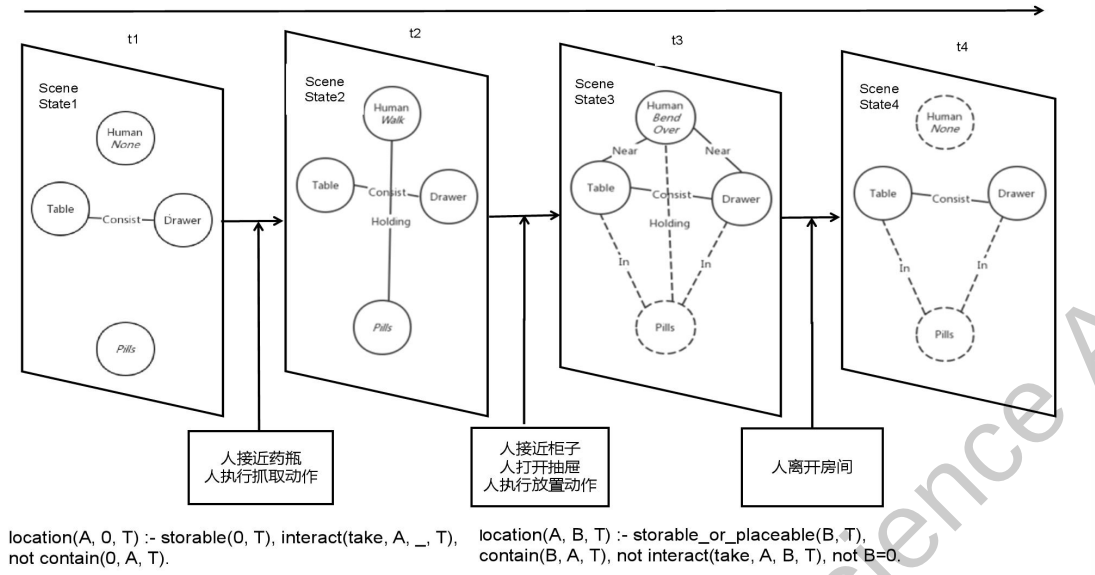


图 4 时空因果场景图动态变化过程示例

在该场景中，药瓶在中途某些时刻以及最终的时刻是不可见的。一个简单场景中由人的动作、目标物体可见性和推理规则触发的时空因果场景图动态变化如图 4 所示。

3.2 3D 场景理解与重建

目标场景中需要进行识别的物体分为环境物体和目标物体。其中，目标物体为系统需要定位和查询的动态可移动物体，如本场景中的小药瓶；环境物体是指能够存放或者放置目标物体的物体，如本场景中的餐桌、柜子等。因此，在构建场景图时，需要对环境物体进行识别，并利用二维的识别结果重建出环境物体的三维位置和大致尺寸（见 3.3.2）。值得注意的是，环境物体在大多数时间处于静止状态，通常环境物体的识别只需要在初始化场景图和环境物体发生变化时执行。从这一点可以看出，使用时空场景图可以节省运行时间和算力资源。

因为本场景的环境物体均为室内的桌子、茶几、抽屉等的物体，所以使用 MS-COCO 和 Open Images 子集上的预训练模型[47][48]进行 2D 物体检测，获取到每个环境物体在四个摄像头下各自的边界框（Bounding box），包括左上角和右下角的坐标、置信度和物体类别。

在对环境物体和目标物体从四个视角分别进行物体识别后，我们便能得到至多四个边界框。为了利用 2D 边界框重建物体在现实世界中的 3D 坐标，并将物体用长方体边界框表示出来，我们使用了基于 OpenCV[51]的计算机视觉三维重建算法。因为 OpenCV 的三维重建算法多为双目摄像头而设计，我们首先将四个摄像头拍摄的图片分成两组，针对两组图片分别进行一次重建。

具体地，为构建物体的 3D 点云，我们使用 Scale-Invariant Feature Transform（SIFT）算法提取两张图片的特征点，然后使用基于 Fast Library for Approximate Nearest Neighbors（FLANN）的 k-Nearest Neighbors（kNN）方法进行特征点匹配，最后使用 Ratio Test 过滤质量较差的配对。

在获得物体的相互配对的特征点后，我们利用基于 OpenCV 的 pymvg[54]所提供的多摄像头系统下的三角化方法，将成对的 2D 坐标映射为真实世界中物体的 3D 坐标，从而构建起物体的 3D 点云。例如，在图片没有扭曲的情况下，我们可以通过解出如下方程中的 X，从而获取摄像头中某一点的 3D 坐标：

$$x = A[R|t]X$$

其中 $\mathrm{x} \in \mathbb{R}^3$ 为这一点的 2D（射影）坐标；$\mathrm{A} \in \mathbb{R}^{3\times3}$为相机的内参矩阵，即

$$A = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

其中f_x和f_y为相机 x、y 轴的焦距，(c_x, c_y)为相机的主点；$[R|t]$为相机的外参矩阵，其中$R \in \mathbb{R}^{3\times3}$为旋转矩阵，$t \in \mathbb{R}^{3\times1}$为平移矩阵；$X \in \mathbb{R}^4$为这一点的 3D（射影）坐标。

对于通过上述方法获取到的一系列 3D 坐标，我们取它们在 x、y、z 轴上的最大和最小坐标。通过这六个坐标我们可以定义能包含整个 3D 点云的最小长方体。这一长方体综合了物体在真实世界中的位置信息和大致的尺寸信息。

3.3 目标物体检测

本项目主要聚焦在药瓶、钥匙等小目标物体的检测与识别，而对于小目标物体的检测是系统实现中的难点之一。在对目标物体进行物体识别的过程中，我们

发现作为目标物体的小药瓶（速效救心丸等）由于本身尺寸较小，在全部四个摄像头的视角内往往其长宽均小于 32 个像素点，相对视频帧大小（1920x1080 像素）已然属于小目标范畴。一般在 MS-COCO 上进行训练的 YOLOv3 网络（实际输入大小为边长 416 像素的正方形）无法直接在 1920x1080 像素的原始图片帧上识别小药瓶。因此，我们尝试将原图切割成近似边长为 144 像素的正方形，并使用 Faster RCNN、Mask RCNN、YOLOv3、SSD[52]、FCOS[53]五种单阶段和双阶段的物体识别网络在每一个小正方形上进行物体识别。为提高神经网络识别结果的置信度，我们在切分的基础上使用 Lanczos 插值算法在 8x8 像素邻域上将正方形图片分别放大 1～5 倍，然后再执行物体检测算法。根据实验我们发现 Faster RCNN 在检测白色小药瓶的任务上平均来看取得最高置信度。因此，在实际使用中，我们选择 Faster RCNN 作为小物体检测模型，并在识别前将切分的小正方形运用 Lanczos 算法放大 3 倍，以提高识别准确率和置信度。

3.4 人体追踪与动作识别

人体动作是物体位置变化的主要原因。对场景中的人进行姿态识别与行为识别，是后续人-物交互和位置推理的基础。首先我们采用 EasyMocap[56]进行姿态评估，通过四个视角的 RGB 图像估计 17 个人体关节点在空间中的坐标。在此基础上，我们采用 LSTM 模型进行行为识别。

行为识别模型接受人体的三维关节点坐标作为输入，将人体分为 Trunk，Left arm，Right arm，Left leg，Right leg 五个部分，采用 LSTM 处理时序信息，

首先，我们通过仿射变换将肢体的关节点坐标映射到特征空间，然后每个肢体将通过不同的 LSTM 网络，融合时空特征。并获得 LSTM 层的 Cell State，进行拼接输入最终的 LSTM 网络，取得 Cell State，并使用 MLP 进行分类。

行为识别模型的损失函数采用交叉熵损失函数（Cross Entropy Loss）以衡量两个概率分布的距离：

$$L = -\frac{1}{N}\sum_{i}\sum_{C=1}^{M} y_{ic}\log(p_{ic})$$

其中 M 为类别的数量，N 为 Batch 数量，y_{ic} 当真实类别为 c 时取 1，其他情况取 0，p_{ic} 则是样本 i 属于类别 c 的预测概率。

我们采用 AdamW 优化方法，BatchSize 为 128 进行 10 个 Epoch 的训练。最终

在测试集上达到了 99.8% 的准确率。

4. 基于人-物交互的物体位置推理

基于人体动作识别和人物交互的物体位置推理引擎是一个逻辑推理系统，能够在给定场景中结合时间与空间的信息来概率性的推断物体的位置，以及物体之间的关系。推理引擎将会结合物体的位置、功能（如可放置、可储存）、状态（如开、关）以及人的行为来做出推理，推导出物体间可能的关系变化。

在根据规则进行推理前，需要两条基本的假设：第一，环境中的物体的位置、功能、状态的变化都是由于人的行为直接或间接导致的。第二，环境中物体的位置，在知道初始位置后，一定可以通过它的属性变化与物体（人）的关系推知。以上这两条假设适用于绝大多数情况，并且是显而易见的，在这两条假设的基础上，我们提出了物体位置的推理逻辑。

我们在实现中选择了 Problog[57]做为规则引擎，Problog 支持 Prolog 合取范式的谓词逻辑表达方式，同时进行了概率计算扩展。我们在下文中使用 Problog 对推理过程中的事实和规则进行描述。

4.1 基本状态属性推理

依据 3.1 中物体属性的定义，我们需要知道如果一个具有存储功能的物体，其存放操作需要满足的条件，该物体应该或可被放置，或有存储功能且有开关功能且处于开的状态，或有存储功能且无开关功能。我们定义如下包含概率的谓词作为推理的事实基础：

```
p(_)::close(Index1, T).

p(_)::open(Index1, T).

p(_)::placeable(Index1, T).

p(_)::switchable(Index1, T).

p(_)::storable(Index1, T).

p(_)::takeable(Index1, T).
```

上述谓词描述了索引号为 Index1 的环境物体在 T 时刻的功能、状态概率（如 $F_{switch}, F_{store}, F_{place}, F_{take}$ 等）。谓词 close(Index1, T)和 open(Index1, T)表示 T 时刻环境物体对应于 3.1 中所定义的开/关状态,谓词 placeable(Index1, T)，switchable(Index1, T)和 storable(Index1, T)表示索引号为 Index1 的环境物体对应于 3.1 中所定义的功能属性（放置，打开，存放）。谓词 takeable(Index1, T)表示索引号为 Index1 的目标物体具有可以被抓取和移动的属性。我们通过物体的索引号来区分人与环境或目标物体，索引号为 0 的物体为人，索引号为非 0 值的物体为环境。p(_)是对应谓词为真的概率，本项目根据物体 Index1 的类别概率来确定，譬如有多大概率为抽屉就有相应的概率可存放（storable），有多大概率为其他小物品就有多大概率可移动拿取（takeable），物体 Index1 的类别概率可从场景感知中得到。

鉴于物体之间的距离是后续人物交互、物品存放的一个重要因素，我们定义了物品之间的邻近谓词 neighbor，同时根据其距离确定了该谓词为真的概率：

```
p(_):: neighbor(Index1, Index2, T).

neighbor(Index1, Index2, T) :- neighbor(Index2, Index1, T).
```

其中 p(_)可以如此定义：

$$p(\_) = 1 - \min(dist(O_m, O_n)/V, 1)$$

$dist(O_m, O_n)$是物体 O_m, O_n的空间距离，V 是控制距离概率的常数。可以看到，neighbor 谓词是基于时空因果场景图中目存储的 T 时刻目标检测和 3D 重建识结果和距离概率公式计算出的索引号为 Index2 和 Index1 的物体间距离表示，如果距离很近，则大概率两个物体很接近，如果两个物体距离很远，则邻近概率接近为 0。此外，我们还定义了 neighbor 谓词的对称性规则。同时，为了增强模型的概率预测能力，谓词 neighbor(Index1, Index2, T)使用概率的方法结合了概率空间与场景图空间的坐标信息，参考上一时刻的位置概率，决定当前时刻的距离概率。

4.2 人-物交互概率推理

我们首先对输入的四个视角图像进行人体姿态估计，随后将人体姿态数据送入行为识别网络，生成 6 种预设行为及其概率作为事实，包括“开、关、放、取、增加物品、其他”，该动作集合表示如下：

$$\mathbb{B}_U = \{B_{other}, B_{take}, B_{put}, B_{open}, B_{close}\, B_{add\_item}\}$$

系统生成的动作谓词和概率示例如下，表示了人对索引号为 Index2 的物体执行相应动作，其中动作谓词间的 Annotated Disjunction 关系“;”表示同一时刻 6 类动作中至多只有一个可以发生：

```
p(_)::action(other, Index2);

p(_)::action(take, Index2);

p(_)::action(put, Index2);

p(_)::action(open, Index2);

p(_)::action(close, Index2);

p(_)::action(add_item, Index2).
```

与状态属性推理中概率的定义类似，这里的 p(_)也从 3D 场景感知的行为识别和目标识别中得到。

基于上述动作相关事实，我们定义人-物交互规则如下：

```
storable_or_placeable(B, T) :- storable(B, T).

storable_or_placeable(B, T) :- placeable(B, T).

interact(take, A, B, T) :- can_change_storge(B, T), contain(B, A, T),
neighbor(A, O, T), action(take, T), storable(O, T).

interact(put, A, B, T) :- can_change_storge(B, T), contain(O, A, T),
neighbor(A, B, T), action(put, T).

interact(open, A, T) :- switchable(A, T), close(A, T), action(open,
```

T), neighbor(A, O, T).

interact(close, A, T) :- switchable(A, T), open(A, T), action(close, T), neighbor(A, O, T).

谓词 storable_or_placeable(B, T)相关的两条规则表示物体 B 在 T 时刻是可以被存储或放置的。谓词 contain(O, A, T)表示在 T 时刻目标物体由人携带。谓词 interact(take, A, B, T)表示 T 时刻人将物体 A 从 B 抓取及其概率。若使 interact(take, A, B, T)为真，当且仅当物体 B 可以改变存储内容，物体 B 存储物体 A，物体 A 与人的距离满足关系，并且这时人做出 Take 动作并且人不存储着 A。interact(put, A, B, T)表示 T 时刻人将物体 A 放入 B 及其概率。interact(open, A, T)表示 T 时刻人打开物体 A 及其概率。interact(close, A, T)表示 T 时刻人关闭物体 A 及其概率。

4.3 物体存储状态推理

为了后续进行较为复杂的推理，需要对物体的位置给出基本推理逻辑。考虑目标物体处于人或环境物体相同的位置，需要满足物体间的距离条件，环境物体是具有放置或存放功能的，并且当前时刻目标物体依附于环境物体。我们定义如下规则:

can_change_storge(B, T) :- storable(B, T), (not switchable(B, T); switchable(B, T), open(B, T)).

can_change_storge(B, T) :- storable(B, T), switchable(B, T), T1 is T-1, (interact(open, B, T1).

can_change_storge(B, T) :- placeable(B, T).

谓词 can_change_storge(B, T) 表示是否可以改变某个物体的存储内容。上述规则对于具有不同功能（如放置和存放）的环境物体存放内容是否可改变做出推理。特别地，对于第二条规则，需要对物体的开关状态进行判断，我们使用上个时刻中人是否与物体存在开的交互进行判断。

4.4 目标物体位置概率推理

目标物体在可见状态下的位置由 3.2 和 3.3 中的目标检测和 3D 重建过程得出。当目标物体不可见状态时，我们通过推理处理如下两种可能性。

第一种可能性，目标物体被人抓取并跟随人移动，如拿在手中或放置于衣物口袋中等。这种情况下目标物体位置可以认为是人的位置。当由于人的移动造成某个摄像头短时遮挡造成不可见，可以根据 4 个摄像头中剩余其他不同位置摄像头图像进行定位，其状态仍然为可见。当目标物体被移动人体全方位短时遮挡之后又出现在初始位置附近，我们认为位置没有发生改变，不会进一步改变场景图中目标物体位置。

第二种可能性，目标物体被人放置于具有存放功能的环境物体如柜子，导致状态变为不可见。这种情况下，我们认为目标物体位置与存放它的环境物体相同。

针对这两种可能性，我们在人物交互概率推理与基本状态属性推理的基础上进行物体的位置概率推理。

对于目标物体被存放于环境物体内部的情况，即上述第二种可能性，我们使用谓词 location(A, B, T)描述在 T 时刻目标物体 A 在环境物体 B 处的概率，定义如下规则：

location(A, B, T) :- storable_or_placeable(B, T), contain(B, A, T), not interact(take, A, B, T), not B=0.

location(A, B, T) :- storable_or_placeable(B, T), neighbor(A, B, T), not contain(_, A, T), not B=0.

location(A, B, T) :- storable_or_placeable(B, T), interact(put, A, B, T), not B=0.

可以看出，目标物体 A 在环境物体 B 的位置的概率要通过一系列推理条件计算，包括 B 可放置或可储存概率，A、B 距离概率计算。满足上述条件后，B 储存 A，那应当 A 在 B 的位置。但是我们应考虑到人物交互影响，因此除 contain(B, A, T)外需要 not interact(take, A, B, T)除去人拿 A 的影响。同理，当 B 不

存放 A 时，也应考虑到人物交互的影响，因此需要考虑 interact(put, A, B, T)。

同理，对于目标物体被存放于环境物体内部的情况，即上述第一种可能性，我们定义如下规则进行推理和概率计算：

```
location(A, O, T) :- storable(O, T), contain(O, A, T),
not interact(put, A, _, T).
location(A, O, T) :- storable(O, T), interact(take, A, _, T),
 not contain(O, A, T).
```

对于 B 的 Index 为 0 的物体（即人这个物体），则需要考虑的只有人物交互的规则。

通过对于三种情况的处理，我们可以同通过目标物体静止可见、被人所携带、放置于环境物体中覆盖目标场景中的绝大多数状态，从而获得目标物体的连续位置信息并存储于场景图中以备查询。而多个位置推理结果的概率值则可以帮助我们处理复杂场景中识别和定位不准确的问题，为老人提供多个按优先级排列的可能位置。

5. 实验与讨论

5.1 数据集采集

我们在实验室家居环境中通过设在屋角的四台摄像机采集共 30GB 的 17 组目标场景样本数据。每组数据包括四台摄像机从不同方向拍摄的视频，图像分辨率为 1920*1280，视频帧率为 25/FPS。场景中的部分环境物体和目标物体如图 5 所示。为增加算法和模型对不同场景的适应性，并确定可能算法改进方向，数据采集过程中我们尽可能考虑了数据在不同维度的多样性，包括：

- 目标物体：采用不同颜色（白色、黄色、蓝色、绿色等）和不同尺寸（大，中，小）的药瓶；
- 重点环境对象：选用了不同颜色（黑色、白色）和不同形态（门柜和抽屉柜）作为容器，用于存储的目标物体；

图 5 环境物体和目标物体多样性示例

- 人与动作：选取 3 个不同身高、衣着颜色的“演员”，在抓取药瓶后保持在身体不同部位（手中、衣兜）。

此外，我们对数据集中的目标物体位置和人体行为动作进行了标注，用于训练和测试子模块的性能。

5.2 基于视觉的场景感知

视觉感知模块主要目标是通过多视角图像提取场景中可见物体的状态，包括了场景三维重建、目标物体检测与定位、人体位姿估计与行为识别等。

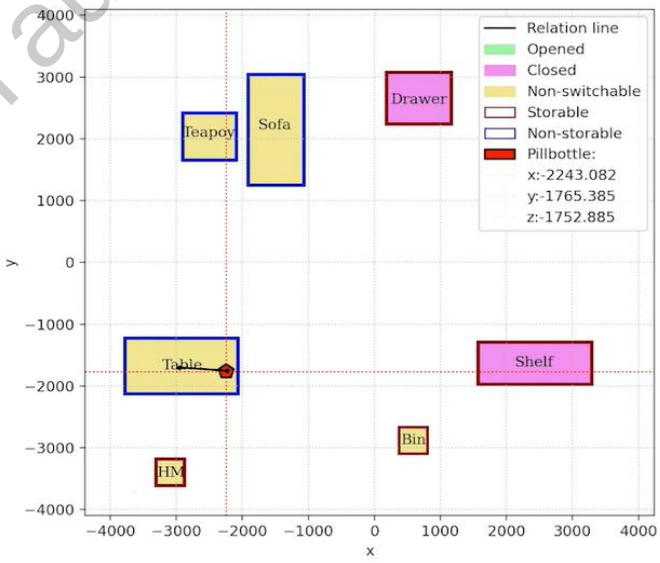


图 6 时空因果场景图俯视图

1920*1080像素原始图像中的目标物体（药瓶）　　放大后的144*144像素切片

图 7. 目标物体图像切片

首先，我们构建场景中静态物体的语义场景图。通过四个摄像机 2D 图像的目标检测，SIFT 特征点提取以及 3D 重建生成物体的 3D Bounding Box，作为时空因果场景图的基础。场景图的俯视图如图 5 所示，除了尺寸和相对位置关系，进一步通过内部和边框颜色动态展示静态属性（如存放、置放功能）和动态属性（如开、关等）。

由于目标物体尺寸较小，因此还需要克服小物体检测问题。如图 7 左侧所示，原始分辨率图像中的药瓶尺寸很小，直接使用现有的目标检测算法（如 3.3 所述 Faster RCNN、Mask RCNN、YOLOv3、SSD、FCOS）都不能很好地检测出物体。我们考虑将图片分割切割后，再分别送入检测器，图 7 右侧四幅图片为切片后的图片。具体地，我们按 144*144 大小对图像进行切片，并对各切片进行插值放大送入常用的物体检测模型。实验中，我们对比了上述 5 种目标检测模型，最终选用 Faster RCNN 作为小物体检测模型，并在识别前将切分的小正方形运用 Lanczos 算法放大 3 倍。

行为识别模型的训练是在对录制数据集进行人工标注后的数据集上进行的。首先从四视角 RGB 图像中获得人的关节点坐标信息，然后从视频中生成动作序列，为了增大数据量，对于每一个序列都加入了一定的噪声。在测试集上生成的结果如图 8 所示。我们将上述识别结果保存在时空因果场景图的对应时间切片中。

图 8. 通过四个摄像机识别出的 17 个人体骨架关键点

5.3 目标物体位置推理

基于视觉感知信息，我们进一步生成时空因果场景图，作为目标物体位置推理的基础。时空因果场景图可视化效果如图 9 所示。场景图中的物体被放置在一个 8000*8000*2500 的三维虚拟空间中，其中 x 轴与 y 轴的范围都为[-4000,4000]，z 轴的范围为[0，2500]，原点为地面中心点。该坐标系完整地对实验场地进行了 3D 重建，通过几何体展示物体属性特征，3D 关节及骨骼展示实验中人的躯干，以及随场景图时间切片变化的位置和相对关系。

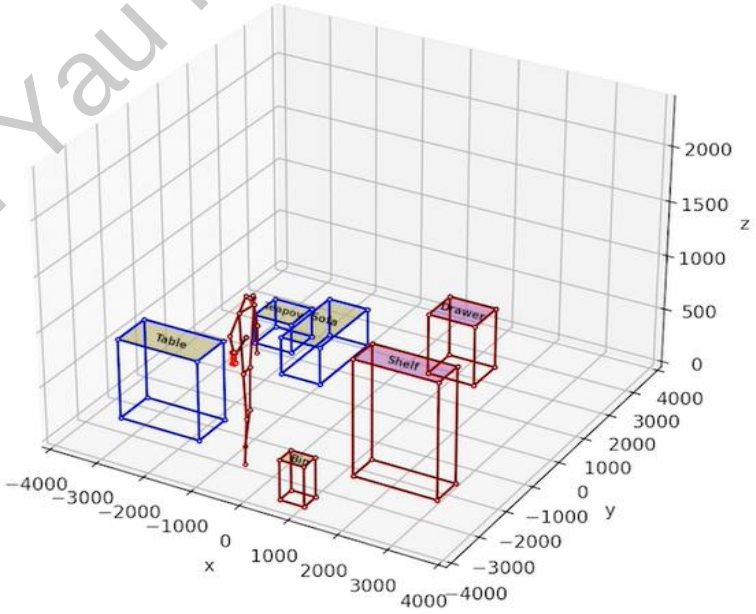


图 9. 时空因果场景图

在上述时空因果场景图时间切片序列上，我们生成 Problog 基本事实并使用第 4 节中所定义的规则进行位置推理。如图 10 所示，我们将样本视频中不同阶段的位置推理结果与基于 Faster RCNN 的目标检测寻物方法进行对比。当物体以可见状态放置于桌面时，两种方法都可以指出目标物体的位置。当人经过桌边执行完抓取动作后，目标物体处于不可见状态，位置推理方法可以推出目标物体随人移动位置发生的连续变化，而只依赖 Faster RCNN 目标检测实现的寻物方法则无法检测目标物体位置并无法继续。当人在接近柜子并执行完包括打开、防止、关闭等一系列动作后，位置推理方法可以推出目标物体大概率被放置于抽屉柜内。因此我们可以得出结论，在上述复杂场景中，本文所提出的位置推理方法比目标检测以及基于目标检测的追踪等方法具有更好的适用性。

现有原型系统的多个算法模块目前可以进行分阶段验证，未来需要进一步的集成优化，以进一步突破性能瓶颈。目标物体识别过程中，由于小物体的尺寸以及前景背景色差等情况，在现有摄像机分辨率下小物体的检测成功率仍有提升空间。当前可视化工作中指示了推理出的目标物体位置，后续工作中会进一步增加语音或文本的位置查询功能以方便使用。

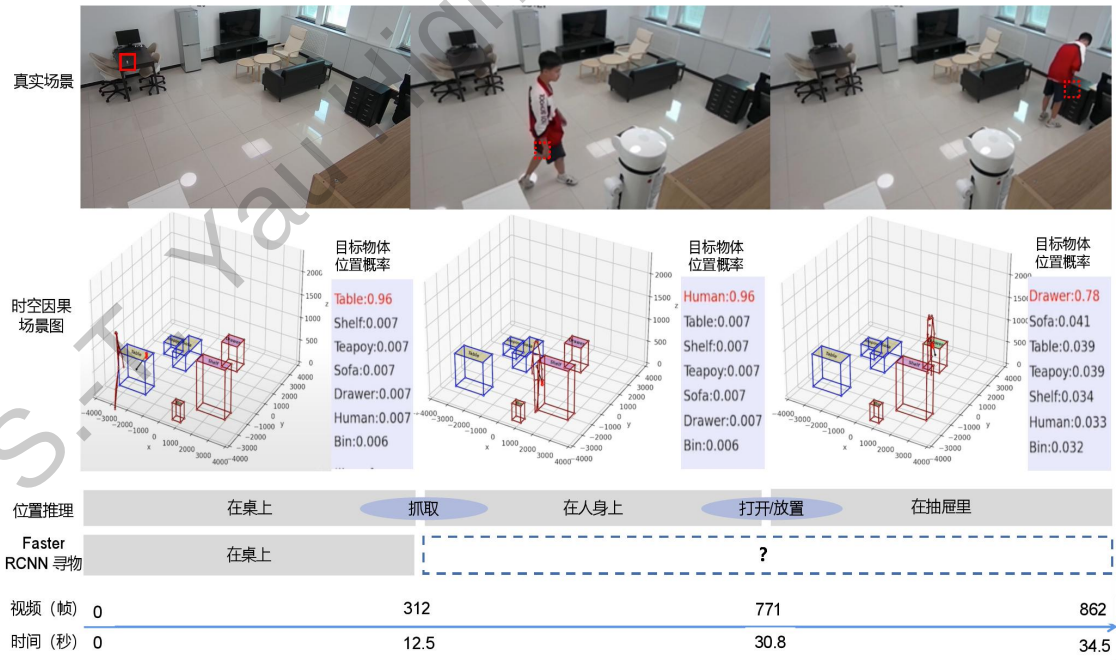


图 10. 目标物体位置推理与基于 Faster RCNN 目标检测寻物比较

6. 结论

本文提出一种基于视觉时空推理技术的智能寻物方法。利用室内部署的摄像机，通过小目标检测识别和 3D 重建等技术构建典型室内家居场景多维度动态时空场景图，以人体动作识别、人-物交互和环境物体功能性建模为基础，通过 Problog 概率推理技术在动态场景图时间序列的基础上实现目标物体不可见状态下的位置推理。实验证明这一方法能够在较长的时间跨度情况下，在有遮挡、出现不可见状态等复杂场景中高效准确地寻找场景中目标物体的位置，有效地帮助阿尔茨海默症老年患者。未来工作会针对当前算法的局限性进一步进行优化。

参考文献

[1] Hirtz, D. et al. How common are the “common” neurologic disorders? Neurology 68, 326–337 (2007).

[2] Burns, A., & Iliffe, S. (2009). Alzheimer’s disease. BMJ 338, b158.

[3] Dementia. (2019, September 19). Retrieved from https://www.who.int/news-room/fact sheets/detail/dementia

[4] Querzfurth, H. W. (2010). Review article. Mechanism of disease Alzheimer’ disease. New England Journal of Medicine, 362, 329-344.

[5] Todd, S., Barr, S., Roberts, M., & Passmore, A. P. (2013). Survival in dementia and predictors of mortality: a review. International journal of geriatric psychiatry, 28(11), 1109-1124.

[6] Manera, V., Petit, P. D., Derreumaux, A., Orvieto, I., Romagnoli, M., Lyttle, G., ... & Robert, P. H. (2015). ‘Kitchen and cooking,’ a serious game for mild cognitive impairment and Alzheimer’s disease: a pilot study. Frontiers in aging neuroscience, 7, 24.

[7] Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., & Murray, C. J. L.(2016). Global, regional, and national incidence, prevalence, and years lived with disability for, 310, 1990-2015.

[8] Mendez, M. F. (2012). Early-onset Alzheimer’s disease: nonamnestic subtypes and type 2 AD. Archives of medical research, 43(8), 677-685.

[9] Wang, H., Naghavi, M., Allen, C., Barber, R. M., Bhutta, Z. A., Carter, A., ... & Bell, M. L. (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. The lancet, 388(10053), 1459-1544.

[10] Berchtold, N. C., & Cotman, C. W. (1998). Evolution in the conceptualization of dementia and Alzheimer’s disease: Greco-Roman period to the 1960s. Neurobiology of aging, 19(3), 173-189.

[11] Bonin-Guillaume, S., Zekry, D., Giacobini, E., Gold, G., & Michel, J. P. (2005).

The economical impact of dementia. Presse Medicale (Paris, France: 1983), 34(1), 35-41.

[12] Meek, P. D., McKeithan, E. K., & Schumock, G. T. (1998). Economic considerations in Alzheimer's disease. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 18(2P2), 68-73.

[13] Waldemar, G., Dubois, B., Emre, M., Georges, J., McKeith, I. G., Rossor, M., ... & Winblad, B. (2007). Recommendations for the diagnosis and management of Alzheimer's disease and other disorders associated with dementia: EFNS guideline. European Journal of Neurology, 14(1), e1-e26.

[14] Bäckman, L., Jones, S., Berger, A. K., Laukka, E. J., & Small, B. J. (2004). Multiple cognitive deficits during the transition to Alzheimer's disease. Journal of internal medicine, 256(3), 195-204.

[15] Nygård, L. (2003). Instrumental activities of daily living: a stepping - stone towards Alzheimer's disease diagnosis in subjects with mild cognitive impairment?. Acta Neurologica Scandinavica, 107, 42-46.

[16] Arnáiz, E., & Almkvist, O. (2003). Neuropsychological features of mild cognitive impairment and preclinical Alzheimer's disease. Acta Neurologica Scandinavica, 107, 34-41.

[17] R. Girshick, "Fast R-CNN," in Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV ' 15), pp. 1440 - 1448, December 2015.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137 - 1149, 2017.

[19] B. Hariharan, P. Arbela ez, R. Girshick, and J. Malik,"Simultaneous detection and segmentation," in Computer Vision—ECCV 2014, vol. 8695 of Lecture Notes in Computer Science, pp. 297 - 312, Springer, 2014.

[20] J. Dong, Q. Chen, S. Yan, and A. Yuille, "Towards unified object detection and semantic segmentation," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics):

Preface, vol. 8693, no. 5, pp. 299–314, 2014.

[21] Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler, "SegDeepM: Exploiting segmentation and context in deep neural networks for object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 4703– 4711, USA, June 2015.

[22] K. Makantasis, A. Doulamis, N. Doulamis, and K. Psychas, "Deep learning based human behavior recognition in industrial workflows," in Proceedings of the 23rd IEEE International Conference on Image Processing, ICIP 2016, pp. 1609–1613, September 2016.

[23] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "DevNet: A Deep Event Network for multimedia event detection and evidence recounting," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 2568–2577, USA, June 2015.

[24] A.S. Voulodimos, D.I. Kosmopoulos, N.D. Doulamis, and T.A. Varvarigou, "A top-down event-driven approach for concurrent activity recognition," Multimedia Tools and Applications, vol. 69, no. 2, pp. 293–311, 2014.

[25] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, pp. 1653–1660, USA, June 2014.

[26] A. Jain, J. Tompson, and M. Andriluka, "Learning human pose estimation features with convolutional networks," in Proceedings of the ICLR, 2014.

[27] J. J. Tompson, A. Jain, Y. LeCun et al., "Joint training of a convolutional network and a graphical model for human pose estimation," in Proceedings of the NIPS, 2014.

[28] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, X. Chen, and X. Wang, "A comprehensive survey of neural architecture search: Challenges and solutions," arXiv preprint arXiv:2006.02903, 2020.

[29] ——, "A survey of deep active learning," arXiv preprint arXiv:2009.00236, 2020.

[30] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L.

Fei-Fei, “Image retrieval using scene graphs,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3668 - 3678.

[31] Zhu, Y., Gao, T., Fan, L., Huang, S., Edmonds, M., Liu, H., ... & Zhu, S. C. (2020). Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. Engineering, 6(3), 310-345.

[32] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese, “3d scene graph: A structure for unified semantics, 3d space, and camera,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5664 - 5673.

[33] H. Qi, Y. Xu, T. Yuan, T. Wu, and S.-C. Zhu, “Scene-centric joint parsing of cross-view videos,” in Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 7292 - 7299.

[34] R. Wang, Z. Wei, P. Li, Q. Zhang, and X. Huang, “Storytelling from an image stream using scene graphs,” pp. 9185 - 9192, 2020.

[35] C. Yu-Wei, L. Yunfan, L. Xieyang, Z. Huayi, and D. Jia, “Learning to detect human-object interactions,” WACV2018, pp. 381 - 389, 2018.

[36] G. Gkioxari, R. Girshick, P. Dollár, and K. He, “Detecting and recognizing human-object interactions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8359 - 8367.

[37] Bai, Y. B., Wu, S., Wu, H. R., & Zhang, K. (2012). Overview of RFID-Based Indoor Positioning Technology. GSR, 2012.

[38] Coskun, V., Ok, K., & Ozdenizci, B. (2011). Near field communication (NFC): From theory to practice. John Wiley & Sons.

[39] Mba, C. J. (2010). Population ageing in Ghana: research gaps and the way forward. Journal of aging research, 2010.

[40] Issahaku, P. A., & Neysmith, S. (2013). Policy implications of population ageing in West Africa. International Journal of Sociology and Social Policy.

[41] 刘淼. (n.d.). 我国 60 岁以上老年人口超 2.3 亿人 占总人口 16.7%_数据要闻_中国政府网. Retrieved from https://www.gov.cn/shuju/2017-08/03/content_5215808.htm

[42] 中国老龄化现状 2017 中国人口结构及人口老龄化现状分析. 金融界. (n.d.). http://finance.jrj.com.cn/2017/03/31101122249888.shtml.

[43] Yong Du, Wei Wang, Liang Wang. Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. IEEE Conference on Computer Vision and Pattern Recognition. CVPR, 2015.

[44] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, Tieniu Tan. An Attention Enhanced Graph Convolutional Lstm Network for Skeleton-based Action Recognition. CVPR, 2019.

[45] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[46] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 91-99.

[47] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[48] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., ... & Ferrari, V. (2020). The open images dataset v4. International Journal of Computer Vision, 128(7), 1956-1981.

[49] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[50] Kai, C., Jiaqi, W., Jiangmiao, P.,Yuhang, C., Yu,X., Xiaoxiao,L., Shuyang,S., Wansen,F.,Ziwei,L.,Jiarui,X.,Zheng,Z.,Dazhi.C,Chenchen,Z.,Tianheng,C.,Qijie,Z., Buyu,L.,Xin,L.,Rui,Z.,Yue,W.,Jifeng,D.,Jingdong,W.,Jianping,S.,Wanli,O.,Chencha nge,L.,& Dahua,L.(2019). Open MMLab Detection Toolbox and Benchmark. arXiv preprint arXiv:1906.07155

[51] Bradski, G. (2000). The OpenCV Library. Dr.Dobb's Journal of Software Tools.

[52] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[53] Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF international conference on computer vision(pp. 9627-9636).

[54] Ross Kirsling. Modal Logic Playground. http://rkirsling.github.io/modallogic/

[55] Saeed Amizadeh, Hamid Palangi, Oleksandr Polozov, Yichen Huang, Kazuhito Koishida. Neuro-Symbolic Visual Reasoning: Disentangling "Visual" from "Reasoning". https://arxiv.org/abs/2006.11524

[56] State Key Lab of CAD&CG, Zhejiang University. EasyMocap. https://github.com/zju3dv/EasyMocap

[57] Yuanlu Xu, Lei Qin, Xiaobai Liu, Jianwen Xie, Song-Chun Zhu. A Causal And-Or Graph Model for Visibility Fluent Reasoning in Tracking Interacting Objects. https://arxiv.org/abs/1709.05437

[58] Giuseppe Marra, Sebastijan Dumančić, Robin Manhaeve, Luc De Raedt. From Statistical Relational to Neural Symbolic Artificial Intelligence: a Survey. arXiv:2108.11451.

[59] The Language Probabilistic Logic Programming. https://problog.readthedocs.io/en/latest/index.html

致谢

1. 论文的选题来源、研究背景

随着全球人口老龄化程度不断加深，同记忆障碍相关的疾病，如阿尔兹海默症等，正严重影响着老年人的正常生活乃至生命健康。例如，老年人可能会因为忘记药物、医疗用品的存放位置而无法按时服药，也可能无法找到重要的个人物品，如身份证、银行卡等。在这一背景下，人工智能作为一种逐渐广泛应用于物体检测、人体姿态估计和因果推理等方向的技术，为解决这一问题提供了新的可能性。

2. 每一个队员在论文撰写中承担的工作以及贡献

林语瑭负责撰写时空场景图的定义（3.1）部分内容、人体动作识别（3.4）、目标物体位置推理（4），王肃羽负责相关研究工作（2）部分内容、环境物体识别与 3D 重建（3.2）、目标物体检测（3.3），田明昊负责相关研究工作（2）部分内容、实验结果（5）部分内容、时空场景图的定义（3.1）部分内容的撰写。其余部分由三人共同完成。

3. 指导老师与学生的关系，在论文写作过程中所起的作用，及指导是否有偿

王亦洲老师为北京市第一 0 一中学英才学院北大前沿计算研究中心 AI 实验室校外指导老师。周宇辰老师为北京市第一 0 一中学英才学院北大前沿计算研究中心 AI 实验室负责人。三位同学为该实验室认知与感知项目组成员。

4. 他人协助完成的研究成果

本文工作所使用的已有技术及其实现在文中和参考文献中明确标出。本文中其他工作为项目组三人在导师指导下完成。

团队成员介绍

王肃羽，北京市第一 0 一中学高三“GITD 实验班”学生。自学数学分析、抽象代数、测度论、信息论、深度学习等课程，并为本年级开设线性代数选修课，为高一年级讲授人工智能课程中的 NLP 部分。AI 项目经验包括“北大元培青年学者项目-对抗样本的现象、生成和防御机制”、“基于深度学习的自闭症儿童

情绪检测手环(MobileNet、ShuffleNet)”、“基于机器学习的全唐诗情感分析和自动生成”等。

林语瑭，北京市第一 0 一中学高二“钱学森实验班”学生。入选 2021 教育部和中国科协“英才计划”清华大学人工智能研究所计算机视觉方向。AI 项目经验包括，“FGSM 对于有 BN 和无 BN 卷积网络的鲁棒性影响分析”、“基于 C++的神经网络基础系统实现（CNN 和 MLP）”等。

田明昊，北京市第一 0 一中学高二“国际英才班”学生。USACO 金级，2015 年 Robofast 冠军，2020 年澳大利亚数学竞赛前 2.5%，2020 年美国数学竞赛前 5%，曾为国内 FRC 比赛中 5737 队队员。AI 项目经验包括“Diagnose Parkinson Disease: Utilizing 1D-inception Network to Analyze Typing Data”、“基于深度学习的自闭症儿童情绪检测手环(MobileNet、ShuffleNet)”等。

指导教师介绍

王亦洲，博士，博雅特聘教授，博士生导师。现任北京大学前沿计算研究中心副主任。1996 年于清华大学获得学士学位，2005 年从加州大学洛杉矶分校(UCLA)获得计算机科学博士学位，同年任美国 Xerox Palo Alto 研究中心 (Xerox PARC) 研究员。2007 年加入北京大学计算机科学技术系。主要研究领域为计算机视觉/人工智能、统计建模与计算、认知计算、医学影像分析、计算艺术。主持国家自然科学基金项目、973 计划项目课题等科研项目十余项，在国际重要学术期刊和学术会议发表学术论文 150 余篇，多篇论文获得最佳论文奖。获 2018 年中国电子学会科学技术奖科技进步一等奖（排名第二）。2019 年任第一届人工智能医疗器械标准化技术归口单位专家组专家、特邀战略委员会专家。

周宇辰，博士，研究员，北京市第一 0 一中学英才学院人工智能导师。ACM 和 IEEE 高级会员，曾任中国计算机学会 CCF 嵌入式系统专业委员会委员。IBM 20 年技术创新经验，曾任 IBM 中国研究院人工智能感知研究主管、IBM 科学院成员、IBM 发明大师、中心技术委员会和专利评审委员会主席等，3 次获杰出技术成就奖。出版学术专著 1 部，参与国际标准 2 项，获国际专利近 50 项，发表学术论文 30 余篇。