视觉-语言-行为：视觉语言融合研究综述*

李睿，郑顺义，王西旗

(武汉大学 遥感信息工程学院，武汉 430079)

摘 要：通过语言给予智能体指示使其完成通用性的任务是人工智能领域的愿景之一。近年来有越多越多的学者试图通过融合计算机视觉与自然语言处理领域的相关技术以期实现此目标。为了及时跟进相关领域的研究，把握视觉与语言融合方向前沿方向，对于视觉-语言-行为最新进展进行综述。首先简单介绍了融合视觉与语言的弱耦合尝试，之后重点综述了视觉-语言-行为这一最新方向相关的视觉语言导航、具身问答及其相似研究，最后总结了制约此领域发展的关键问题及可能的解决方案。

关键词：计算机视觉；自然语言处理；深度学习；视觉语言导航；具身问答

中图分类号：TP181

Vision-Language-Action: A Survey of the Integration of Vision and Language

Li Rui, Zheng Shunyi, Wang Xiqi

(School of Remote Sensing and Information Engineering, Wuhan University, Wu Han 430079, China)

Abstract: The idea that we might be able to give general and verbal instructions to a agent and have at least a reasonable probability that it will carry out the required task is one of the long-held visions of robotics and artificial intelligence (AI). We have noticed that more and more scholars in recent years have tried to this target using the latest advances in the field of computer vision and natural language progressing. In order to follow up the research in related fields and grasp the frontier of the fusion of vision and language, the latest progress of visual-language-behavior is reviewed in this paper. We first introduce the weak coupling experiment of fusion between vision and language, then focuses on the Vision-and-Language Navigation, the Embodied Question Answering and their similar research, and finally looks forward to the future development trends in this field.

Key words: computer vision; natural language progressing; deep learning; vision and language navigation; embodied question answering

0 引言

“人工智能 (Artifi cial Intelligence) 系统应该能感知环境并且为了执行特定任务而采取行动。”[1]这是人工智能领域先驱 Russell 和 Norvig 在其经典教材 *Artificial Intelligence: A Modern Approach* 中对于 AI 给出的期待。近年来，得益于计算机尤其是显卡性能的提升，曾经因为需要大量计算而一度沉寂的深度学习 (Deep Learning)[2]方法再次得到学术界和工业界的重视，而 tensorflow[3]、Caffe[4]和 keras[5]等数十种深度学习开源框架的提出，更是显著降低了深度学习技术的使用门槛。近年来，深度学习方法已经在计算机视觉 (Computer Vision)[6]、自然语言处理 (Natural Language Processing)[7]和自动语音识别 (Automatic Speech Recognition)[8]等各个领域得到了广泛而深入的应用，推动了人脸识别、自动驾驶和语音识别等一系列技术的发展和成熟，在很多目标清晰、规则明确的任务比如物体检测[9]、目标分割[10]甚至是围棋[11]、象棋[12]领域达到甚至超越了人类的表现。

但是当前深度学习领域的研究往往局限在特定领域甚至特定的任务上，对于环境往往也有许多假设或是限制，与通用人工智能 (Artificial General Intelligence)[13]或是自主智能体 (autonomous agent)[14]的目标相去甚远。与此同时，我们注意到有越来越多的学者开始研究视觉和语言交叉领域，比如图像描述、视觉问答和文本图像生成等，但是往往缺乏对于环境的理解，而近年来陆续出现了将视觉和语言与行为联系的研究，比如视觉语言导航[15, 16](VLN, Vision-and-Language Navigation)、具身问答[17~20](EmbodiedQA, Embodied Question Answering) 和

收稿日期：2019-09-09 **修回日期**：2019-11-06 **基金项目**：国家自然科学基金(41671452)

作者简介：李睿（1996-），男，山东滨州人，硕士研究生，主要研究方向为深度学习、计算机视觉等（lironui@whu.edu.cn）；郑顺义（1973-），男（通信作者），山西襄汾人，教授，博士，博导，主要研究方向为计算机视觉、数字摄影测量等（syzheng@whu.edu.cn）；王西旗（1990-），男，山东临沂人，博士研究生，主要研究方向为深度学习、计算机视觉.

交互式问答[21](IQA, Interactive Question Answering) 等，上述研究不但需要融合视觉与语言技术，还需要智能体 (agent) 针对基于文本的问题，在虚拟的空间环境中进行路径规划和探索，相对而言是对视觉与语言的深度融合。而国内少有相关领域的介绍或研究，因此我们希望能通过文献综述的方式让更多学者了解这方面的进展。

1 计算机视觉与自然语言处理的弱耦合尝试

在介绍视觉语言导航和具身问答任务之前，我们先简单回顾下目前研究较多的几个融合视觉与语言的任务。得益于视觉中检测、分割和识别等方法和自然语言处理中自动文本生成等方法的提出和完善，陆续有学者尝试解决需要同时使用两个领域技术的问题，在语言和视觉的整合方面进行了大量的研究，将单词、短语、句子、段落、文章等不同层次的语言信息与图像、视频等不同层次的视觉信息相结合。

1.1 图像描述 (Image Caption)

最早被提出的问题是图像描述，即根据给定图片自动生成语言描述[22]。初期解决方案分为图像预处理、特征提取和文本生成三个模块[23]，比如图像算子提取特征，SVM 检测可能存在的目标，根据目标属性生成句子[22, 24]，但是对于目标属性定义的依赖限制了描述的生成。近年来则大多基于深度学习提出解决方案，2015 年谷歌 DeepMind 团队和李飞飞团队分别提出了基于编码-解码 (Encoder-Decoder) 框架的 Show and Tell[25] 和 Neural Talk[26] 模型，均使用 CNN+RNN 的模式；生成对抗网络 (Generative Adversarial Networks) [27]、深度强化学习 (Deep Reinforcement Learning) [28]和注意力机制(Attention Mechanism) [29]也被陆续引入相关研究。随着解决方案的成熟，图像描述任务也不断扩展，比如基于群组的图像描述方法[30]和生成文本的风格化问题[31]等。

1.2 视觉问答 (Visual Question Answer)

视觉问答可以视作图像描述问题的逻辑推理扩展，任务形式通常是，给定一幅图片和基于图片的问题，输出问题的正确答案，包括是或否的二元逻辑问题和多项选择[32]以及图像中的文本信息等[33]。解决方法基本可划分为四类：联合嵌入模型 (Joint Embedding Models)、注意力机制模型 (Attention Mechanisms)、模块化组合模型 (Compositional Models) 和知识库增强模型 (Knowledge Base-enhanced Models)[34]。联合嵌入方法将图像和文字在公共特征空间学习，注意力机制使用局部图像特征对不同区域的特征加权解决噪声问题，模块化组合模型引入不同功能的神经网络模块 (Neural Module Networks)，知识库增强模型通过外部知识库解决需要先验知识的问题。作为视觉问答的拓展领域视频问答 (Video Question Answer) 也越来越多受到学者的关注[35]。

1.3 文本图像生成 (Text to Image Synthesis)

文本图像生成则正好是图像描述的逆向问题，从给定文本描述生成图像。变分自编码器 (Variational Auto Encoders)[36]、基于流的生成模型 (flow-based generative model)[37]和近似 PixelCNN[38]等方法都曾用于解决此问题。但是自生成对抗网络 (Generative Adversarial Networks, GAN)[39]引入文本图像生成以来，因其卓越表现已成为主流方法。当前基于 GAN 的优化方向主要有：其一是调整网络结构，比如增加网络深度[40]或者引入多个判别器[41]，其二是充分利用文本信息，比如注意力机制[42]和 MirrorGAN[43]等工作，其三是增加额外约束，比如 Condition-GAN[44]机制等工作，其四是分阶段生成，比如李飞场景图 (Scene Graphs) [45]和语义中间层 (semantic layout) 等工作[46]。同样文本图像生成任务形式也得到了进一步拓展，比如基于多段落生成系列图片的故事可视化[47] (Story Visualization) 任务和文本生成视频等[48]。

1.4 视觉对话 (Visual Dialog)

视觉对话可以视为图像描述问题的对话扩展，在 2017 年 CVPR 会议上由佐治亚理工学院的 Das A 等人提出，与视觉问答中单次交互不同，视觉对话要求智能体基于视觉内容与人类进行多次交流[49]。具体讲，就是在给定图像、对话历史记录和关于图像问题的条件下，智能体必须基于图像内容，从历史记录中推断上下文，并准确地回答该问题。与此相似与此相似的还有‘Guess What?!’任务但是其仅限于答案为“是”或“否”的布尔型问题[50]，Alamri H 等人则进一步引入了视频对话 (Video Dialog) 的任务[51]。视觉对话目前的解决方案主要有：基于深度强化学习 (Deep Reinforcement Learning) 的模型[52]，注意力机制[53]，条件变分自编码器方法[54]，和基于神经网络模块 (Neural Module Networks)的架构[55]等。

1.5 多模态机器翻译 (Multimodal Machine Translation)

多模态机器翻译则是对机器翻译工作的扩展，其目标是给定描述图片的源语言和图片本身，根据文本内容和图像提供的额外信息翻译成目标语言[56]，同时 Specia 定义了两类任务，其一是单句源语言描述图片，其二是多句源语言描述图片，Elliott 等人进一步将任务二扩展到多种源语言[57]（比如关于同一图片英语、法语和德语描述），Wang Xin 等人则进一步把任务扩展到视频层面[58]。研究方向主要有：引入注意力机制[59]，分解任务目标[60]，充分发掘图片的视觉特征[61]，强化学习方法的使用[62]，无监督学习模型的扩展[63]等。

1.6 小结

除了以上任务之外，还有定位视频中文本位置的视频文本定位 (Video Scene Text Spotting)[64]任务，判断文本描述和图片内容是否匹配的视觉蕴含 (Visual Entailment)[65]任务，问题必须基于图片内容进行推理才能回答的视觉推理 (Visual Reasoning)[66]任务等。包括上述问题在内的大部分早期研究往往是在视觉和语言的层次上不断扩展，比如将图片扩展到视频，从句子扩展到段落等，或者在此基础上加入逻辑层面的推理等。但是我们注意到，在一定意义上讲，上述任务仅仅是计算机视觉和自然语言处理两个任务的弱耦合，甚至部分任务可以把视觉部分和语言部分可以完全分离地进行训练，将其中一部分的

输出作为另一部分的输入就能实现任务的要求，因此没有真正的发掘视觉与语言的内在联系，并且其更多的侧重于特定任务的完成，对于环境的感知是被动甚至缺失的，因此我们尝试姑且称之为计算机视觉和自然语言处理的弱耦合尝试。

2 视觉语言导航、具身问答及其相似研究

与上述任务有所不同的是，视觉语言导航(Vision-and-language navigation)和具身问答(Embodied Question Answering)需要智能体(agent)不但能够综合使用视觉与语言能力，还必须不断通过与环境主动地交互获取所需要的信息，在交互中完成对环境的理解，进而完成指定的任务，即在计算机视觉(Computer Vision)和自然语言处理(Natural Language Processing)之外，还加入了行为规划(Action Planning)的部分，下面分别介绍。

2.1 视觉语言导航(Vision-and-language navigation)

2.1.1 视觉导航和语言导航相关研究

基于视觉的导航往往需要环境的先验信息[67, 68]，或者需要使用激光雷达、深度图或从运动中获取的数据以纯几何方法构建三维地图[69, 70]，或者需要人类指导的地图构造过程[71, 72]。并且在地图构造的过程中，即使环境有明显的模式或特征，但是在被完全建模之前也是不能被观察到的[73]。环境构建与路径规划之间的分离使得系统变得脆弱，因此越来越多的研究开始转向端到端的学习方式——不需要显式的模型或状态估计便可实现从环境图像到路径行为的转换[73, 74]。

同时学者很早就开始关注对于自然语言的理解[75, 76]，引入语言指引的导航策略也受到过许多关注[77~79]，但是其往往对于语言或环境做出了一定程度的抽象，比如语言指令限制在特定范围或假设语言命令有固定的结构以及将环境中的物体做特定标记[80, 81]，或者将智能体限制在只需要有限知觉的视觉受限环境中[75, 78, 79]。近年来虽然有很多新的多模态非结构化的仿真平台比如 House3D[82]、AI2-THOR[83] 和 HoME[84]等，但是其基于人工合成而非真实图像的模型一定程度上限制了环境建模的准确性和丰富性。

2.1.2 视觉语言导航任务内容

Qi Wu 等人在 2018 年 CVPR 会议上提出了视觉语言导航(Vision-and-Language Navigation)任务，要求智能体在给定语言指令的情况下，在作者提供的 Matterport3D Simulator 仿真环境中，从随机初始位置到达目标位置，并且其仿真环境构建于包含大量基于真实图像生成的 RGB-D 全景图的数据集 Matterport3D[85]。但是其相对复杂和具体的语言描述与实际不太相符。因此在 2019 年，Qi Wu 等人进一步提出被称为 RERERE(Remote Embodied Referring Expressions in Real indoor Environments)的任务，精简指令的同时引入了对于环境的理解[86]。

2.1.3 视觉语言导航任务最新进展

Qi Wu 提出任务的同时，同时提出了将智能体建模为基于长短期记忆(Long Short Term Memory, LSTM)序列到序列结构(sequence-to-sequence architecture)注意力机制循环神经网络[15]的解决方案和随机移动策略和最短路径策略两种基线算法以及人类在此任务中的表现（成功率 86.4%）。

视觉语言导航任务也可以视为在给定语言指导条件下寻找从起始点到目标点最佳路径的轨迹搜索(trajectory search)问题，基于此 Fried D 提出 speaker-follower 系统[87]，系统中的 speaker 模型用于学习路径描述，follower 模型用于预测和执行路径，并使用全景行为空间(panoramic action space)代替视觉运动空间(visuomotor space)的方式使得智能体可以感知当前位置 360° 全景视觉。

为解决视觉语言导航任务中的解决跨模态基标对准(cross-modal grounding)问题和增强泛化能力(generalization)，Xin Wang 等人提出基于强化学习(Reinforcement Learning)和模仿学习的(Imitation Learning)的策略[16]，引入了强化跨模态匹配(Reinforced Cross-Modal Matching)方法和自监督模仿学习(Self-Supervised Imitation Learning)方法。

在之前的研究中，视觉语言导航任务中主要评价指标是任务完成度即最终位置与目标位置之间的关系，因此语言指示在导航任务中所发挥的作用难以量化。谷歌研究院的 Jain V 等人因此提出可刻画预测路径与语言指示之间契合度的评价标准 CLS(Coverage weighted by Length Score)，并根据此指标扩展了 R2R 数据集，提出包含更多节点和更多样化路径的 R4R(Room-for-Room)数据集[88]。

在实际导航场景中，使用者更倾向于利用简练的语言给定任务的内容而非具体详尽地描述路径的所有信息，因此 Qi Wu 等人进一步提出 Remote Embodied Referring Expressions in Real indoor Environments(RERERE)的任务[86]，其中包含类似“去带条纹墙纸的卧室(Go to the bedroom with the striped wallpaper)”的导航(Navigation)部分和类似“把放在凳子旁边的枕头拿给我(Bring me the pillow that is laying next to the ottoman)”的指称表达(Referring expression)[89]部分，并提供了被称为导航-指向模型(Navigator - Pointer Model)的基线算法。

2.2 具身问答

2.2.1 具身认知概念

具身认知(Embodied cognition)这一概念是随着哲学、人工智能和相关领域的发展关于认知的本质被重新思考和定义的过程中诞生的，新的研究越来越倾向于认为大多数现实世界的思考常常发生在非常特殊通常也十分复杂的环境中，出于非常实际的目的，并且利用外部事物的可交互性(interaction)和可操作性(manipulation)，即认知是一种非常具体化(embodied)和情景化(situated)的活动[90]。身体的解剖学结构、身体的活动方式、身体的感觉和运动体验都决定了我们怎样认识和看待世界[101]。简而言之，具身认知理论认为人的生理体验与心理状态之间是有着深刻的内在联系。因此具身相关任务的内涵，就

是将任务具体化到可交互的场景中，而非传统的静态图片或无法互动的视频。

2.2.2 具身问答任务内容

具身问答 (Embodied Question Answering) 是 Das 等人在 2018 年 CVPR 会议上提出的任务[17]，将智能体随机安放在三维环境中的某个位置，并且以语言的形式提出类似“汽车的颜色是什么 (What color is the car?)”或者“有多少个房间里有椅子 (How many rooms contain chairs?)”等类似需要环境信息的问题，为了得到问题的答案，智能体 (agent) 需要自主地对环境进行探索并且收集所需要的信息，最后对问题做出解答。智能体仅依靠单目全景 RGB 摄像头与环境交互，而没有类似环境地图、自身定位的全局表示或类似物体信息、房间描述的结构表示，当然也没有关于任务本身的额外信息，即先验知识几乎为零，需要智能体充分理解任务内容的情况下，通过与具体环境的不断交互，实现对环境的理解，进而完成问题的回答。

2.2.3 具身问答任务最新进展

Das 等人提供的基线算法中智能体视觉、语言、导航和回答 (vision, language, navigation and answering) 四个部分的实现[17]，其中视觉部分基于通过 CNN 将 RGB 图像生成固定大小的表示，语言部分使用 LSTM 编码，导航部分引入包含选择动作（前进，左转，右转）的规划模块和指定执行次数（1，2…）的控制模块的自适应倍率计算 (Adaptive Computation Time) 方法，问答部分计算智能体轨迹最后五帧的图像-问题相似性的视觉编码与问题的 LSTM 编码进行比较并输出结果。

在上述研究的基础上，受人类将行为概念化为一系列更高层次语义目标（比如为了吃夜宵，人类会将其抽象为“离开卧室-走到厨房-打开冰箱-找到甜点”而不会详尽地规划路线）的启发，Das 等人进一步提出了模块化学习策略 (Neural Modular Controller)[19] ，将学习目标加以分解。

Yu L 等人则把 EQA 任务扩展为 MT-EQA(Multi-Target EQA) 即在问题形式中引入了多目标[20]，比如类似“卧室里的梳妆台比厨房里的烤箱更大么 (Is the dresser in the bedroom bigger than the oven in the kitchen)”这样的问题。

Wijmans E 等人设计了基于三维点云格式的具身问答数据集 MP3D-EQA[18]，设计并测试了多达 16 种不同的导航策略组合，提出损失加权方案 Inflection Weighting 以提高行为模仿的有效性。

2.3 其他相似研究及分类探讨

受 Tomasello、Skinner 和 Bruner 等经验主义者“语言可以通过使用来学习”观点的启发，百度研究院的 Yu H 和 Zhang H 等人提出用于智能体在交互中学习自然语言的虚拟环境 XWORLD[91~93]，包括导航 (navigation) 和问答 (question answering)两种交互模式，智能体可通过交互的方式习得主动搜索和记忆新物体特征的可迁移技能[134]和通过内插（interpolation，在同样的情况下使用已掌握单词的新组合）和外推（extrapolation，使用从其他情况或模型中转换的新词）方法解释包含新的单词组合甚至新单词的句子[130]的能力。在随后的工作中，Yu H 和 Zhang H 等人进一步将环境扩展到三维环境 XWORLD3D[94]。

Gordon D 等人在 2018 年 CVPR 会议上提出交互式问答 (Interactive Question Answering) 任务[21]，其任务形式与具身问答相似，智能体需要在与环境的交互中完成对给定问题的回答。作者同时提出被称为 HIMN(Hierarchical Interactive Memory Network) 的解决方法。但是一方面相关研究较少，另一方面也缺乏数据集的后续更新，目前影响力和知名度较低。

参考相关文献中的分类方法[17]，我们可以从视觉、语言和行为三个方面对现有的视觉和自然语言结合的任务进行简单的分类。如图 1 所示，视觉可划分为基于图像和基于视频，语言可划分为问答形式和对话形式，行为可划分为被动感知和主动感知。

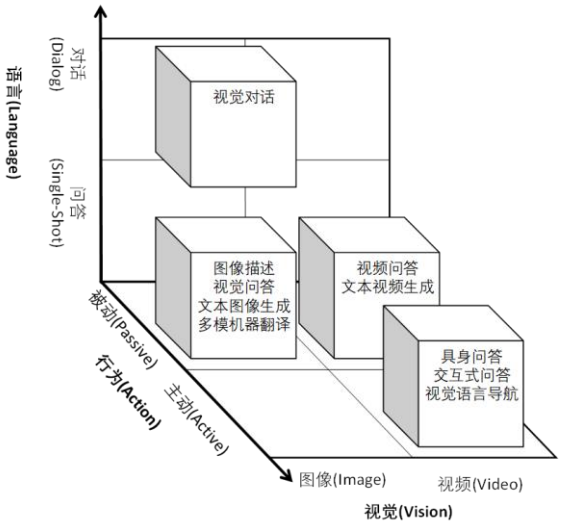

图 1 视觉和自然语言结合任务简单分类

Fig. 1 A simple classification the task combining computer vision and natural language processing

3 相关数据集介绍

视觉语言导航任务主要包含 3 个数据集，其一是 Qi Wu 等人在提出视觉语言导航任务时开源的 R2R(Room-to-Room) 数据集[15]，其二是 Jain V 等人在改进任务评价方法时开源的 R4R(Room-for-Room) 数据集[88]，其三是 Qi Wu 等人提出 RERERE 任务[86]时建立的数据集（暂未开源）。表 1 是三个数据集的简单对比，从对比中可以发现，因为 R4R 数据集更倾向于使得智能体运动轨迹更加符合导航指令而非最短距离，因此参考路径的长度要大于最短路径的长度；而 RERERE 任务则倾向于使用更加简洁的指令，因此指令平均长度要小于 R2R

表 1 视觉语言导航任务常用数据集对比

Table 1 Comparison of common data sets of the Vision-and-Language Navigation

数据集	分类	Homes	Floors	Unq Qns	Tol Qns
EQA v1	训练集	643	\	147	4246
	验证集	67	\	104	506
	测试集	57	\	105	529
MT-EQA	训练集	486	\	2030	14495
	验证集	50	\	938	1954
	测试集	52	\	1246	2838
MP3D-EQA	训练集	57	102	767	174
	验证集	10	16	130	88
	测试集	16	28	239	112

1 可见验证集：包含关于与训练集重叠的环境的新指令。

2 不可见验证集：验证集与训练集完全分离。

3 R2R: https://bringmeaspoon.org

R4R: https://github.com/google-research/google-research/tree/master/r4r

具身问答任务数据集主要包括 3 个数据集，其一是 Das 等人开源的 EQA(Embodied Question Answering) v1 数据集[17]，其二是 Yu L 等人引入多目标任务时提出的 MT-EQA(Multi-Target EQA) 数据集[131]，其三是 Wijmans E 等人将任务中的数据类型替换为点云时提出的数据集 MP3D-EQA 数据集[18]，后两个数据集暂时未开源。表 2 是三个数据集的内容对比，表 3 是三个数据集问题类型对比。需要注意的是，数据集中包含被称为 Unique Question 的问题，是指可能产生歧义的问题，比如房间中同时存在两台冰箱时，问题 ‘What room is the air conditioner located in?’ 就会产生歧义。

4 计算机视觉与自然语言处理融合未来方向展望

4.1 真实环境迁移与泛化能力

视觉与自然语言结合的任务取得了令人瞩目的进展，从早期简单将两部分技术简单串联加和的形式扩展到需要智能体借助视觉和语言理解环境并且采取行动的深度融合，但是绝大部分任务都是基于现有的数据集在模拟的环境中进行。诚然，考虑到目前表现较好的算法均是基于需要大量试错的深度强化学习方法，在真实环境中训练的确会消耗大量的时间与精力，但是在模拟环境表现完美的模型迁移到真实环境中也可能会遇到很多意料之外的问题，而现有的绝大部分研究只是在数据集上达到了较高的精度（比如视觉语言导航任务中 SOTA 算法在可见验证集和不可见验证集上分别达到了 73.0%和 61.3%的成功率[16]），仅有少数学者将算法在实际环境中加以验证[95]。因此未来研究重要方向之一是如何将模型迁移到真实环境中。在此过程中，泛化能力又是其中关键，即智能若遇到训练集中未出现的环境或者未遇到的物体，能否根据过往经验做出较为合理

表 2 具身问答常用数据集对比

Table 2 Comparison of common data sets of the Embodied Question Answering

数据集	分类	样本数	参考路径平均长度	最短路径平均长度
R2R (指令平均长度 29 个单词)	训练集	14039	9.91	9.91
	可见验证集	1021	10.2	10.2
	不可见验证集	2249	9.50	9.50
R4R	训练集	233613	20.6	10.5
	可见验证集	1035	20.4	11.1
	不可见验证集	45162	920.2	10.1
RERERE (指令平均长度 19 个单词)	训练集	5181	\	\
	可见验证集	1668	10.3	10.3
	不可见验证集	2174	9.56	9.56

1 Homes：数据集中涉及的房间环境。

2 Floors：数据集中涉及的楼层数。

3 Unq Qns：环境中可能出现歧义的问题。

4 Tol Qns：数据集中所有问题。

5 EQA v1: https://embodiedqa.org/

的反应，可能的解决方案是借鉴已经在视觉对话、常识推理和事实预测等方向得到广泛使用和验证的外部知识库方法，即利用事实性或常识性的先验知识提高智能体对于环境的理解和认知能力。

4.2 与环境更强大的交互能力

目前已经开源的数据集中，智能体与环境之间的交互相对有限，仅涉及打开微波炉、移动物体[74, 95]或到达指定位置等基本操作，并且可采取的运动形式限制在特定范围（比如前进、左转和右转），虽然在最新的研究中已经涉及类似“把放在凳子旁边的枕头拿给我 (Bring me the pillow that is laying next to the ottoman)”[89]这类相对较为复杂的交互形式，但是显然与真实环境的交互方式和运动形式有较大的差距，并且简化了真实环境中的诸多物理性限制，比如“去厨房拿一个鸡蛋”和“去厨房拿一把勺子”语言指示，在真实的环境中智能体需要考虑分别以何种的力度夹取鸡蛋和勺子，而现有的数据集并不考虑此类区别。另一个比较有前景的方向是与物联网的深度结合，电视、空调和冰箱等对于人类而言需要后天习得交互方式的电器，却因其规则明确和易于联网的性质能够与智能体直接交互。最后就是对环境中其他信息的利用，比如利用声音信息对不可见物体的非视距重建[96]、使用工具达成指定目标甚至与环境中其他智能体的对话交流等。这些与环境的相对复杂的交互是目前研究所欠缺的，但也是未来智能体在真实环境中运行所需要的。

4.3 推理能力的引入

目前无论是视觉语言导航还是具身问答，所给的任务都相对直接（比如根据语言提示到达某个房间或者回答环境中某物体是什么颜色等），但是现实生活中更多是是需要推理能力的问题，比如类似视觉推理[66]任务中的比较、属性识别和逻辑运算

等初级推理能力，以及演绎、归纳和类比等高级推理能力。虽然在部分研究中已经涉及推理能力[20]，但仍相对简单，未来可能会引入类似“房间装修是什么风格？”或者“到书房中取一本散文集。”这种涉及相对高级推理能力的任务，前者需要智能体基于房间的整体特征比如吊灯的样式、桌椅的摆放和墙纸的花饰等信息归纳推理得出装修风格的答案，后者则需要智能体能够区分散文、小说或诗歌等不同的文体。当然目前视觉和自然语言方面的进展距离解决此类问题仍有较大空间，但是推理能力尤其是高级推理能力的研究不失为一个值得关注的研究方向。

4.4 三维数据的使用

三维点云数据可以提供比图像更丰富和准确的信息，Wijmans E 等人发现在具身问答任务中点云信息可以提升智能体避障能力的学习[18]，Wang Y 等人甚至发现仅仅将二维的双目视觉图像转换为三维点云数据就能大幅提高目标检测的准确度[97]，因此点云数据可能不单在信息内容方面甚至是在数据表示方面均提供了更多的信息。但是一方面受制于点云数据获取的成本和难度，成本百元的相机模组在短短几秒钟内便可获取千万像素级别的高精度图像，但是点云获取设备往往动辄数十万获取时间也往往需要数分钟甚至数小时。另一方面基于点云的深度学习研究相对滞后于图像，虽然得益于 Point Net++[98]、ASCN[99]、和 SplatNet[100]等方法的提出，点云数据固有的无序性和旋转性不再是应用深度学习技术的障碍，但是学术界对于点云数据的研究仍远远少于图像数据。因此不论是点云数据集的构建还是基于点云数据的研究均不同程度的存在一些困难。后续的研究可能需要更多的引入点云格式的环境信息，为了弥补目前点云数据获取困难的状况，基于双目视觉的三维重建可能是很有希望的辅助手段之一。

4.5 学习目标的优化

建构主义者认为，学习是学习者在与环境交互作用的过程中主动地建构内部心理表征的过程。而我们现在已经拥有了多个可交互的模拟环境[82~84]，因此后续的研究可以在不断地交互进行比如对自然语言的理解[91~93]或者对环境中工具的使用等能力的学习和提升。此外从表 1 的分类中可以看出，视觉语言导航、具身问答以及交互式问答等在语言层面仍停留于“问答”阶段，即针对单一问题给出正确的答案，未来的研究中很有可能将目标优化到“对话”层面，即针对多个有内在逻辑联系的问题分别给出正确答案，同时问题之间的内在联系也有助于智能体更好地理解环境。

5 结束语

a) 简要回顾了图像描述、视觉问答、文本图像生成、视觉对话和多模态机器翻译等融合计算机视觉和自然语言处理的早期研究，分别介绍了其任务内容、解决方案和前沿方向等，并分析了其存在的问题。

b) 重点综述了包括视觉语言导航和具身问答等在内的“视觉—语言—行为”相关研究，包括其任务内容、与之研究的异同以及最新进展，并对比分析了相应任务的数据集，从视觉、语言和行为三个维度对目前的不同任务简单分类。

c) 对制约视觉与语言融合相关技术发展的关键问题包括泛化能力、交互能力、推理能力、数据集和学习目标等做出了分析，并给出了可能的解决方案。

参考文献

[1] Russell S J, Norvig P. Artificial intelligence: a modern approach[M]. Malaysia; Pearson Education Limited,, 2016.

[2] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436.

[3] Abadi M, Agarwal A, Barham P, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems[J]. arXiv preprint arXiv:1603.04467, 2016.

[4] Jia Yangqing, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]. Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.

[5] Chollet F. Keras[J]. https://github.com/keras-team/keras, 2015.

[6] 邓柳, 汪子杰. 基于深度卷积神经网络的车型识别研究[J]. 计算机应用研究, 2016, 33(3): 930-932. (Deng Liu，Wang Zijie．Deep convolution neural networks for vehicle classification[J]．Application Research of Computers，2016，33(3)：930—932.)

[7] 李阳辉, 谢明, 易阳. 基于深度学习的社交网络平台细粒度情感分析[J]. 计算机应用研究, 2017, 34(3): 743-747. (Li Yang-hui, Xie Ming, Yi Yang. Fine-grained sentiment analysis for social network platform based on deep learning model[J]. Application Research of Computers,2017(3):743 - 747.)

[8] Dong Yu, Li Deng. Automatic Speech Recognition: A Deep Learning Approach[M]. Automatic speech recognition: a deep learning approach. Springer, 2014.

[9] Hu Xiaowei, Lei Zhu, Fu C W, et al. Direction-aware spatial context features for shadow detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7454-7462.

[10] 宋焕生, 张向清, 郑宝峰, 等. 基于深度学习方法的复杂场景下车辆目标检测[J]. 计算机应用研究, 2018 (2018 年 04): 1270-1273. (SONG Huansheng, Zhang Xiangqing, Zheng Baofeng, et al. Vehicle detection based on deep learning in complex scene[J]. Application research of computers, 2018, 35(4): 1270-1273. DOI:10.3969/j.issn.1001-3695.2018.04.067)

[11] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550(7676): 354.

[12] Campbell M, Hoane Jr A J, Hsu F. Deep blue[J]. Artificial intelligence, 2002, 134(1-2): 57-83.

[13] Goertzel B, Wang Pei. A foundational architecture for artificial general intelligence[J]. Advances in artificial general intelligence: Concepts, architectures and algorithms, 2007, 6: 36.

[14] Franklin S, Graesser A. Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents[C]. International Workshop on Agent Theories, Architectures, and Languages. Springer, Berlin, Heidelberg, 1996: 21-35.

[15] Anderson P, Wu Qi, Teney D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3674-3683.

[16] Wang Xin, Huang Qiuyuan, Celikyilmaz A, et al. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6629-6638.

[17] Das A, Datta S, Gkioxari G, et al. Embodied question answering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018: 2054-2063.

[18] Wijmans E, Datta S, Maksymets O, et al. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6659-6668.

[19] Das A, Gkioxari G, Lee S, et al. Neural modular control for embodied question answering[J]. arXiv preprint arXiv:1810.11181, 2018.

[20] Yu Licheng, Chen Xinlei, Gkioxari G, et al. Multi-target embodied question answering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6309-6318.

[21] Gordon D, Kembhavi A, Rastegari M, et al. Iqa: Visual question answering in interactive environments[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4089-4098.

[22] Farhadi A, Hejrati M, Sadeghi M A, et al. Every picture tells a story: Generating sentences from images[C]. European conference on computer vision. Springer, Berlin, Heidelberg, 2010: 15-29.

[23] Bai Shuang, An Shan. A survey on automatic image caption generation[J]. Neurocomputing, 2018, 311: 291-304.

[24] Kulkarni G, Premraj V, Ordonez V, et al. Babytalk: Understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-2903.

[25] Vinyals O, Toshev A, Bengio S, et al. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(4): 652-663.

[26] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3128-3137.

[27] Dai B, Fidler S, Urtasun R, et al. Towards diverse and natural image descriptions via a conditional gan[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 2970-2979.

[28] Ren Zhou, Wang Xiaoyu, Zhang Ning, et al. Deep reinforcement learning-based image captioning with embedding reward[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 290-298.

[29] Long Chen, Zhang Hanwang, Jun Xiao, et al. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5659-5667.

[30] Chen Fuhai, Ji Rongrong, Sun Xiaoshuai, et al. Groupcap: Group-based image captioning with structured relevance and diversity constraints[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1345-1353.

[31] Mathews A, Xie Lexing, He Xuming. Semstyle: Learning to generate stylised image captions using unaligned text[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8591-8600.

[32] Antol S, Agrawal A, Lu Jiasen, et al. Vqa: Visual question answering[C]. Proceedings of the IEEE international conference on computer vision. 2015: 2425-2433.

[33] Singh A, Natarajan V, Shah M, et al. Towards vqa models that can read[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 8317-8326.

[34] Wu Qi, Teney D, Wang Peng, et al. Visual question answering: A survey of methods and datasets[J]. Computer Vision and Image Understanding, 2017, 163: 21-40.

[35] Tapaswi M, Zhu Yukun, Stiefelhagen R, et al. Movieqa: Understanding stories in movies through question-answering[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4631-4640.

[36] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[37] Dinh L, Krueger D, Bengio Y. Nice: Non-linear independent components estimation[J]. arXiv preprint arXiv:1410.8516, 2014.

[38] Reed S, van den Oord A, Kalchbrenner N, et al. Parallel multiscale autoregressive density estimation[C]. Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 2912-2921.

[39] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]. Advances in neural information processing systems. 2014: 2672-2680.

[40] Zhang Han, Xu Tao, Li Hongsheng, et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(8): 1947-1962.

[41] Nguyen T, Le T, Vu H, et al. Dual discriminator generative adversarial nets[C]. Advances in Neural Information Processing Systems. 2017: 2670-2680.

[42] Xu Tao, Zhang Pengchuan, Huang Qiuyuan, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks[C]. Proceedings of the IEEE Conference on Computer Vision and

Pattern Recognition. 2018: 1316-1324.

[43] Qiao Tingting, Zhang Jing, Xu Duanqing, et al. MirrorGAN: Learning Text-to-image Generation by Redescription[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1505-1514.

[44] Dash A, Gamboa J C B, Ahmed S, et al. Tac-gan-text conditioned auxiliary classifier generative adversarial network[J]. arXiv preprint arXiv:1703.06412, 2017.

[45] Johnson J, Gupta A, Fei-Fei L. Image generation from scene graphs[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1219-1228.

[46] Li Wenbo, Zhang Pengchuan, Zhang Lei, et al. Object-driven Text-to-Image Synthesis via Adversarial Training[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 12174-12182.

[47] Li Yitong, Gan Zhe, Shen Yelong, et al. StoryGAN: A Sequential Conditional GAN for Story Visualization[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6329-6338.

[48] Li Yitong, Min M R, Shen Dinghan, et al. Video generation from text[C]. Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[49] Das A, Kottur S, Gupta K, et al. Visual dialog[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 326-335.

[50] De Vries H, Strub F, Chandar S, et al. Guesswhat?! visual object discovery through multi-modal dialogue[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5503-5512.

[51] Alamri H, Cartillier V, Das A, et al. Audio Visual Scene-Aware Dialog[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7558-7567.

[52] Das A, Kottur S, Moura J M F, et al. Learning cooperative visual dialog agents with deep reinforcement learning[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 2951-2960.

[53] Seo P H, Lehrmann A, Han B, et al. Visual reference resolution using attention memory for visual dialog[C]. Advances in neural information processing systems. 2017: 3719-3729.

[54] Massiceti D, Siddharth N, Dokania P K, et al. FlipDial: A generative model for two-way visual dialogue[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6097-6105.

[55] Kottur S, Moura J M F, Parikh D, et al. Visual coreference resolution in visual dialog using neural module networks[C]. Proceedings of the European Conference on Computer Vision (ECCV). 2018: 153-169.

[56] Specia L, Frank S, Sima'an K, et al. A shared task on multimodal machine translation and crosslingual image description[C]. Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. 2016: 543-553.

[57] Elliott D, Frank S, Barrault L, et al. Findings of the second shared task on multimodal machine translation and multilingual image description[J]. arXiv preprint arXiv:1710.07177, 2017.

[58] Wang Xin, Wu Jiawei, Chen Junkun, et al. VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research[J]. arXiv preprint arXiv:1904.03493, 2019.

[59] Calixto I, Liu Qun, Campbell N. Doubly-attentive decoder for multi-modal neural machine translation[J]. arXiv preprint arXiv:1702.01287, 2017.

[60] Elliott D, Kádár A. Imagination improves multimodal translation[J]. arXiv preprint arXiv:1705.04350, 2017.

[61] Calixto I, Rios M, Aziz W. Latent Visual Cues for Neural Machine Translation[J]. arXiv preprint arXiv:1811.00357, 2018.

[62] Qian Xin, Zhong Ziyi, Zhou Jieli. Multimodal machine translation with reinforcement learning[J]. arXiv preprint arXiv:1805.02356, 2018.

[63] Lample G, Ott M, Conneau A, et al. Phrase-based & neural unsupervised machine translation[J]. arXiv preprint arXiv:1804.07755, 2018.

[64] Cheng Zhanzhan, Lu Jing, Nui Yi, et al. Efficient Video Scene Text Spotting: Unifying Detection, Tracking, and Recognition[J]. arXiv preprint arXiv:1903.03299, 2019.

[65] Xie Ning, Lai F, Doran D, et al. Visual entailment: A novel task for fine-grained image understanding[J]. arXiv preprint arXiv:1901.06706, 2019.

[66] Johnson J, Hariharan B, van der Maaten L, et al. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2901-2910.

[67] Borenstein J, Koren Y. The vector field histogram-fast obstacle avoidance for mobile robots[J]. IEEE transactions on robotics and automation, 1991, 7(3): 278-288.

[68] Kim D, Nevatia R. Symbolic navigation with a generic map[J]. Autonomous Robots, 1999, 6(1): 69-88.

[69] Davison A J, Reid I D, Molton N D, et al. MonoSLAM: Real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007 (6): 1052-1067.

[70] Sim R, Little J J. Autonomous vision-based exploration and mapping using hybrid maps and Rao-Blackwellised particle filters[C]. 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2006: 2082-2089.

[71] Royer E, Bom J, Dhome M, et al. Outdoor autonomous navigation using monocular vision[C]. 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2005: 1253-1258.

[72] Kidono K, Miura J, Shirai Y. Autonomous visual navigation of a mobile robot using a human-guided experience[J]. Robotics and Autonomous Systems, 2002, 40(2-3): 121-130.

[73] Gupta S, Davidson J, Levine S, et al. Cognitive mapping and planning for visual navigation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2616-2625.

[74] Zhu Yuke, Gordon D, Kolve E, et al. Visual semantic planning using deep successor representations[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 483-492.

[75] Chaplot D S, Sathyendra K M, Pasumarthi R K, et al. Gated-attention architectures for task-oriented language grounding[C]. Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[76] Huang A S, Tellex S, Bachrach A, et al. Natural language command of an autonomous micro-air vehicle[C]. 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2010: 2663-2669.

[77] Chen D L, Mooney R J. Learning to interpret natural language navigation instructions from observations[C]. Twenty-Fifth AAAI Conference on Artificial Intelligence. 2011.

[78] Guadarrama S, Riano L, Golland D, et al. Grounding spatial relations for human-robot interaction[C]. 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2013: 1640-1647.

[79] Mei Hongyuan, Bansal M, Walter M R. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences[C]. Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[80] Kollar T, Tellex S, Roy D, et al. Toward understanding natural language directions[C]. Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction. IEEE Press, 2010: 259-266.

[81] Shimizu N, Haas A. Learning to follow navigational route instructions[C]. Twenty-First International Joint Conference on Artificial Intelligence. 2009.

[82] Wu Yi, Wu Yuxin, Gkioxari G, et al. Building generalizable agents with a realistic and rich 3d environment[J]. arXiv preprint arXiv:1801.02209, 2018.

[83] Kolve E, Mottaghi R, Gordon D, et al. Ai2-thor: An interactive 3d environment for visual ai[J]. arXiv preprint arXiv:1712.05474, 2017.

[84] Brodeur S, Perez E, Anand A, et al. HoME: A household multimodal environment[J]. arXiv preprint arXiv:1711.11017, 2017.

[85] Chang A, Dai A, Funkhouser T, et al. Matterport3d: Learning from rgb-d data in indoor environments[J]. arXiv preprint arXiv:1709.06158, 2017.

[86] Qi Yuankai, Wu Qi, Anderson P, et al. RERERE: Remote Embodied Referring Expressions in Real indoor Environments[J]. arXiv preprint arXiv:1904.10151, 2019.

[87] Fried D, Hu Ronghang, Cirik V, et al. Speaker-follower models for vision-and-language navigation[C]. Advances in Neural Information Processing Systems. 2018: 3314-3325.

[88] Jain V, Magalhaes G, Ku A, et al. Stay on the Path: Instruction Fidelity in Vision-and-Language Navigation[J]. arXiv preprint arXiv:1905.12255, 2019.

[89] Yu Licheng, Tan Hao, Bansal M, et al. A joint speaker-listener-reinforcer model for referring expressions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7282-7290.

[90] Anderson M L. Embodied cognition: A field guide[J]. Artificial intelligence, 2003, 149(1): 91-130.

[91] Zhang Haichao, Yu Haonan, Xu Wei. Interactive language acquisition with one-shot visual concept learning through a conversational game[J]. arXiv preprint arXiv:1805.00462, 2018.

[92] Yu Haonan, Zhang Haichao, Xu Wei. Interactive grounded language acquisition and generalization in a 2d world[J]. arXiv preprint arXiv:1802.01433, 2018.

[93] Zhang Haichao, Yu Haonan, Xu Wei. Listen, interact and talk: Learning to speak via interaction[J]. arXiv preprint arXiv:1705.09906, 2017.

[94] Yu Haonan, Lian Xiaochen, Zhang Haichao, et al. Guided feature transformation (gft): A neural language grounding module for embodied agents[J]. arXiv preprint arXiv:1805.08329, 2018.

[95] Zhu Yuke, Mottaghi R, Kolve E, et al. Target-driven visual navigation in indoor scenes using deep reinforcement learning[C]. 2017 IEEE international conference on robotics and automation (ICRA). IEEE, 2017: 3357-3364.

[96] Lindell D B, Wetzstein G, Koltun V. Acoustic non-line-of-sight imaging[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6780-6789.

[97] Wang Yan, Chao W L, Garg D, et al. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 8445-8453.

[98] Qi C R, Yi Li, Su Hao, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[C]. Advances in neural information processing systems. 2017: 5099-5108.

[99] Xie Saining, Liu Sainan, Chen Zeyu, et al. Attentional shapecontextnet for point cloud recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4606-4615.

[100] Su Hang, Jampani V, Sun Deqing, et al. Splatnet: Sparse lattice networks for point cloud processing[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2530-2539.

请填写“文章编号”等信息，然后将文档打印出来并签字，扫描后作为附件发送到 arocmag@163.com 即可

稿件保密审查证明

《计算机应用研究》编辑部：

我单位 武汉大学遥感信息工程学院 共 3 名作者的稿件

《 视觉-语言-行为：视觉语言融合研究综述 》

（稿件编号为 19090143 ），已被贵刊月刊录用。此文不涉及国家、军事、商业秘密，同意此稿在贵刊发表。

单位名称（加盖公章）： 郑顺义

2019 年 11 月 1 日

《计算机应用研究》稿件版权转让协议书

文章编号：19090143

文章名称：视觉-语言-行为：视觉语言融合研究综述

根据《中华人民共和国著作权法》及其实施条例的有关规定，本稿件全体作者同意上述稿件在《计算机应用研究》（月刊）杂志上发表，并且自本协议签订之日起，作者同意将此稿的著作权及相关财产权转让给《计算机应用研究》编辑部，即《计算机应用研究》编辑部对上述稿件具有以下专有使用权：复制权、全世界范围发行权、信息网络传播权、汇编权等；允许国内外的文献检索系统及数据库系统检索收录。该稿件出版后，《计算机应用研究》编辑部在出版一个月内向作者一次性支付稿酬，稿酬中已包括该稿件版权转让的费用。

本稿件的作者作出承诺：

1、本稿件署名无争议。多位作者署名或多个单位的文章，作者保证署名内容、顺序和单位内容、顺序无争议。在投稿之后，如变更、增加或删除作者署名内容或顺序，作者须以书面形式通知编辑部，并需有所有作者的签字和第一作者身份证证明或第一单位证明。

2、所有作者保证该文章的合法性，属于作者自己的科研成果。无抄袭、剽窃、侵权、数据伪造等不良行为，不涉及国家机密；稿件中引用他人成果已注明出处。如因不良行为造成的经济损失和社会负面影响，由作者本人负责,本刊编辑部不承担任何连带责任。

3、无论何种原因，要求撤回投稿，要求第一作者须在正式出版三个月前以书面形式通知编辑部。

4、在介绍国内外同类相关工作时，已明确标明哪些工作是作者自己的研究成果，哪些工作是对他人工作的介绍。在介绍他人工作时，须明确标明引证来源，并作为参考文献列出。

《计算机应用研究》编辑部和作者任何一方如果违反上述约定，按照《中华人民共和国著作权法》相关规定承担相应责任。

全体作者签字：[illegible] 郑顺义 王西琪

2019 年 11 月 1 日