

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2025)01-0198-14

论文引用格式: Feng Y K, Li X W, Liu P W, Guo F J, Long T and Li X. 2025. User-customized intention understanding agent for planar scanned images. Journal of Image and Graphics, 30(01):0198-0211(冯弋珂, 励雪巍, 刘鹏伟, 郭丰俊, 龙腾, 李玺. 2025. 面向平面扫描图像的用户定制意图理解智能体. 中国图象图形学报, 30(01):0198-0211)[DOI:10.11834/jig.230903]

面向平面扫描图像的用户定制意图理解智能体

冯弋珂¹, 励雪巍^{2*}, 刘鹏伟³, 郭丰俊³, 龙腾³, 李玺²

1. 浙江大学软件学院, 宁波 315048; 2. 浙江大学计算机科学与技术学院, 杭州 310027;

3. 上海合合信息科技股份有限公司, 上海 200072

摘要: 目的 移动端应用中对平面扫描图像的用户意图理解是常见的现实需求, 传统方法主要是利用大量用户历史行为数据进行建模, 预测用户对新图像的意图, 但该应用场景面临着定制化问题、交互次数少等挑战, 限制了传统方法的效果。而近几年出现的智能体方法可以较好地应对这些挑战, 为定制意图理解任务提供了新的思路。基于此, 提出了一个面向平面扫描图像的用户定制意图理解智能体。方法 智能体由任务感知、任务规划、任务执行与反馈校正模块构成, 并针对方法面临的小样本增量问题以及计算资源有限、基准数据集不足等技术挑战, 首先提出了“分而治之”的域泛化方法, 将基任务与定制化任务的推理解耦, 使其互不影响。其次通过模板匹配进行意图理解, 以实现无需微调即可应对新的定制化任务的功能。然后通过自提升策略减少意图理解结果噪声, 提升域泛化的可靠性。此外还构建了平面扫描图像的定制意图理解基准数据集。结果 本文智能体在所提出的基准数据集上与其他7种方法进行了比较, 在平均交并比(mean intersection over union, mIoU)指标上, 智能体的mIoU达90.47%, 相比于性能第2的方法提高了15.60%, 总正确率提高了22.10%。同时进行了消融实验, 验证了智能体各部分的有效性。最后将智能体应用在公开票据数据集CORD(consolidated receipt dataset)上, 验证了智能体的泛化能力。结论 本文提出的智能体超越了前沿检测和分割模型在平面扫描图像的定制意图理解任务上的表现, 同时回避了对每个子任务微调模型的过程, 方法具有有效性和高效性。

关键词: 用户意图理解; 智能体; 小样本增量学习; 显著性检测; 交互式分割

User-customized intention understanding agent for planar scanned images

Feng Yike¹, Li Xuewei^{2*}, Liu Pengwei³, Guo Fengjun², Long Teng³, Li Xi²

1. College of Software Technology, Zhejiang University, Ningbo 315048, China; 2. College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China; 3. Shanghai Hehe Information Technology Co., Ltd., Shanghai 200072, China

Abstract: Objective In the era of mobile internet, mobile applications are developing rapidly and becoming increasingly common in society, thus becoming an indispensable part of daily life. The demands and expectations of users for mobile applications are also constantly increasing. In the development of mobile applications, user intention understanding is an important research field, which aims to provide more personalized and intelligent services for users by analyzing their behavior and needs. User intention understanding for image input in mobile applications is a common practical require-

收稿日期: 2024-01-04; 修回日期: 2024-08-19; 预印本日期: 2024-08-26

* 通信作者: 励雪巍 xueweili@zju.edu.cn

基金项目: 国家自然科学基金项目(62441602); 国家杰出青年科学基金项目(62225605); 国家自然科学基金项目(U20A20222); 浙江省重点研发计划资助(2023C03196); INTSIG 合作研究项目

Supported by: National Natural Science Foundation of China (62441602); National Science Foundation for Distinguished Young Scholars (62225605); National Natural Science Foundation of China (U20A20222); Zhejiang Key Research and Development Program (2023C03196); INTSIG Research Fund

ment. In image-related human-computer interaction, users often need to interact with mobile applications through touch clicks or gestures so that mobile applications can intelligently understand users' intentions. Traditional intention understanding methods mainly use a large amount of historical user behavior data to model and predict possible user intentions for new images. However, the application scenario is faced with some challenges such as the customization problem and few interactions, thus limiting the effectiveness of traditional methods. In recent years, with the development of autonomous artificial intelligence, the emergence of agent technology has provided new perspectives for user-customized intention understanding task. Agents can imitate and learn human thinking process, accurately understand users, reduce the burden of memory and operation, and help mobile applications better understand users' intentions and needs. Therefore, we propose to build a user-customized intention understanding agent for planar scanned images for user-customized intention understanding task. **Method** The user-customized intention understanding agent consists of task perception, task planning, task execution, and feedback correction modules. Task perception module extracts information from an input image and combines it with user-customized template information obtained from stored intention libraries to understand the intention of the input image. The task planning module utilizes the results of the perception module to plan and make decisions on input image. The task execution module follows the decision made in the planning module to execute the corresponding action and output the intention understanding result of the input image. If the user is not satisfied with the current result, then the user can provide feedback through feedback correction module, and then the agent will output a new result, use the corrected image as a new user-customized intention template image, and add it to the intention libraries. To address the technical challenges faced in building these modules such as few-shot incremental learning problem, limited computing resources, and insufficient benchmark datasets, we propose several new technological ideas. First, a "divide and conquer" domain generalization method is proposed to decouple the inferences of basic tasks and customized tasks so that they do not affect each other. Second, we achieve intention understanding through template matching to cope with new customized samples without fine-tuning, which solves the few-shot incremental learning problem. Then, we use a self-improvement strategy to reduce the noise of intention understanding results and improve the reliability of domain generalization. In addition, we construct a customized intention understanding benchmark for planar scanned images, providing a data foundation for this study and subsequent research. **Result** We conduct a series of ablation experiments on our dataset to clearly show the effectiveness of different parts of our agent. We compare our agent with seven state-of-the-art methods on our dataset, including traditional saliency detection models, model fine-tuning, and visual language models. The mean intersection over union of the agent on our dataset reaches 90.47%, which is 15.60% higher than that of the second best-performing method, and the total accuracy is improved by 22.10%. Finally, the agent is applied to a public dataset called consolidated receipt dataset to verify its generalization ability. **Conclusion** In this study, we propose a user-customized intention understanding agent for planar scanned images. Experimental results demonstrate that our proposed "divide and conquer" domain generalization method, self-improvement strategy, and object detection parameters all have a certain promoting effect on intention understanding. The agent outperforms the cutting-edge detection and segmentation models on the user intention understanding task without fine-tuning, proving the effectiveness and efficiency of our method.

Key words: user intention understanding; agent; few-shot incremental learning; saliency detection; interactive segmentation

0 引言

移动互联网时代,人们对平面图像和文档电子化、结构化的需求不断增长,光学字符识别(optical character recognition, OCR)等移动应用在社会中变得越来越普遍,已经成为人们日常生活中不可或缺的一部分。在OCR应用的发展过程中,为了实现对

平面图像的准确扫描,用户意图理解就是一个重要的研究领域,它旨在通过分析用户的行为,更加个性化和智能化地理解用户需求。过去对平面图像的用户意图理解方式主要是利用大量的历史行为数据进行建模,预测用户对新图像的可能意图,这样的传统方法是相对封闭和静态的。而近年来,随着人工智能(artificial intelligence, AI)不断向具有自主性的方向发展,各种新技术的出现为用户意图理解任务提

供了新的思路,如 AI 智能体(AI agent)技术。智能体已有在各个领域的探索和应用,如 WebGPT(web generative pre-trained Transformer)使用人工反馈进行浏览器辅助问答(Nakano 等,2021),ReAct(synergizing reasoning and acting in language models)在语言模型中协同推理和行动(Yao 等,2022),CAMEL(communitive agents for "mind" exploration of large language model society)是可以进行“角色扮演”的沟通式社会型智能体(Li 等,2023),Voyager是使用大语言模型的开放式具身智能体(Wang 等,2023a),张晨等人(2023)提出基于双层 DQN (deep Q-network)的多智能体路径规划方法。在用户意图理解领域,智能体能接收用户与机器的交互,如触摸点击或手势,并模仿和学习人类的思维过程,能更好地理解用户对平面图像的意图和需求,减轻用户的记忆负担与操作负担,因此智能体可能逐渐替代传统方法成为移动应用中平面图像意图理解任务的主流方法。

传统方法在移动应用中平面扫描图像的用户意图理解任务上效果不佳的原因是,移动应用面临着几个挑战,限制了已有方法的效果:1)定制化挑战。对于普通平面扫描图像的意图理解任务(以下称基任务),如图 1 顶部框所示,右侧图中高亮处表示图像的意图理解结果。对于基任务,传统方法已经能实现较好的性能。然而,由于用户需求的多样性与现实环境的复杂性,产生了定制意图理解任务。如图 1 底部框所示,存在多个潜在意图的图像称为歧义图像,需要用户进行定制化输入,手动选择目标框区域,移动应用才能完成意图理解,这样的任务称为用户定制意图理解任务,该任务具有异构、动态等特点,目前仍然是一个巨大的挑战。2)交互次数少的挑战。用户期望移动应用能智能理解识别意图,因此用户与应用的交互次数应尽量少,最好只通过一次交互就对所有同类图像实现意图理解。而交互次数少意味着移动应用获得的信息较少,如何利用少量信息学习到正确的意图是一个挑战。

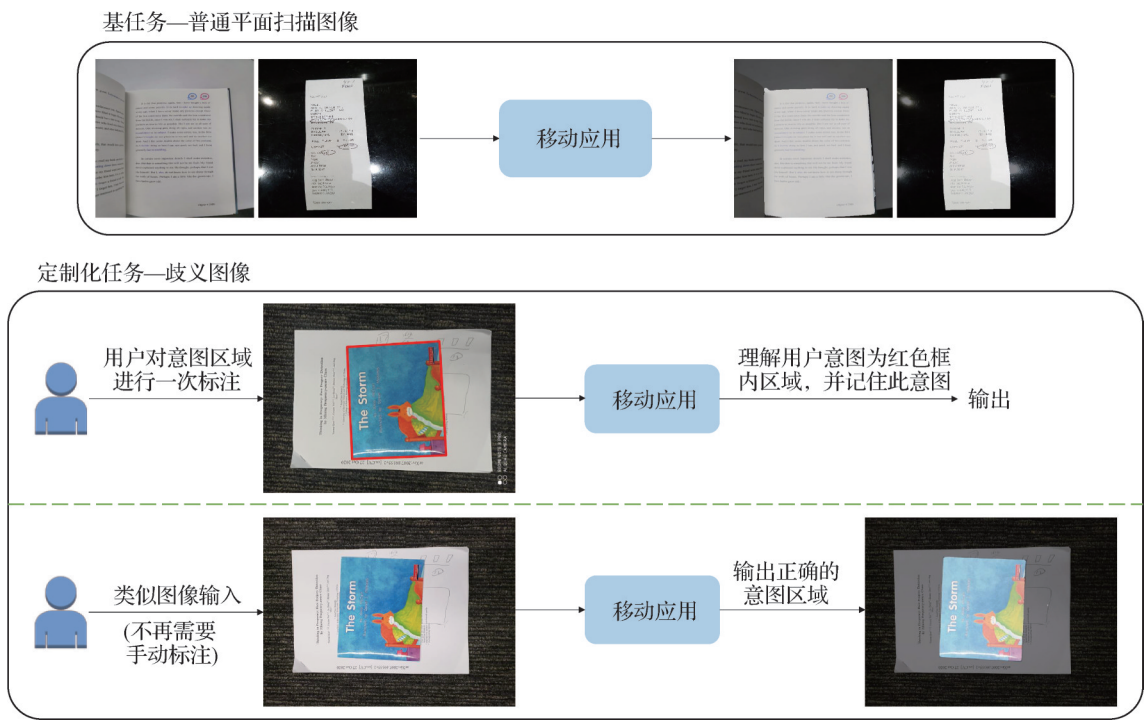


图 1 普通图像及歧义图像的意图理解

Fig. 1 Intention understanding of ordinary images and ambiguous images

基于应用面临的定制化、异构、可学习信息少等挑战,意图理解过程中建模是一个困难的问题,因此本文使用智能体方法实现准确灵活的意图理解,但面临几个技术上的挑战:1)学习范式是增量的;2)定

制化任务中定制样本的数量少;3)存储、计算等资源有限,算法必须是轻量级的;4)方法对不同类型图像需要具备很强的泛化能力;5)目前符合该场景的公开的基准数据集较少。因此本文提出了“分而治之”

的域泛化方法、模板匹配、自提升策略等新的技术思路来应对以上技术挑战,完成平面扫描图像的用户定制意图理解任务。

具体来说,首先,为了应对增量挑战并提高对不同类型图像的泛化能力,本文提出了一种“分而治之”的域泛化方法,将基任务与定制化任务的推理解耦,使其互不影响,在增强对新任务的适应能力的同时,避免对基任务的遗忘。其次,为了应对定制样本数量少的挑战,本文将意图理解定义为模板匹配问题,将不同的定制意图构建为不同模板,从而在决策时,判断样本属于某个定制化任务或属于基任务,然后加入模板或基任务的先验信息。这样的方式回避了对每个定制化任务进行网络微调,避免训练规模随定制化任务数增长而增大,从而极大地降低了网络的训练开销,解决了移动端用户对于开销较为敏感、存储和计算等资源有限的问题。随后,为进一步提升本文方法的鲁棒性,提出自提升策略减少预测噪声,提升域泛化的可靠性,提出反馈校正模块实现智能体的自我更新,提高意图理解能力。最后,从任务数据本身出发,为了应对可供使用和评测的相关基准数据集较少的挑战,提供了一个超过100组共计超过2000幅样本的数据集,为平面扫描图像的用户定制意图理解任务提供了一个较大的基准数据集,为本文以及后续的研究提供了数据基础。

针对平面扫描图像的用户定制意图理解任务,本文贡献点主要如下:1)将智能体方法引入到用户意图理解任务中,将智能体与环境的交互模式应用在用户与移动应用的交互上,构建了一个面向平面扫描图像的用户定制意图理解智能体,由任务感知、任务规划、任务执行与反馈校正模块构成;2)基于技术上面临的挑战,在智能体的构建过程中提出了“分而治之”的域泛化方法、模板匹配方法、自提升策略等技术点,实现准确高效的意图理解;3)构建了平面扫描图像的定制意图理解基准数据集,在该数据集上的实验证明了本文方法的可行性,并进行了落地的应用验证。

1 相关工作

1.1 智能体

智能体在人工智能领域中指能自主感知环境、做出决策然后采取相应行动的主体(Xi等,2023),

英文称 Agent。简单来说,它旨在让计算机能够理解用户的兴趣,并自主地代表用户采取行动(Green等,1997)。Xi等人(2023)将过去智能体的技术发展过程分为几个阶段:符号型智能体、反应型智能体、基于强化学习的智能体、基于迁移学习和元学习的智能体,而近两年大语言模型的出现将智能体的智能程度提升到了新的高度(Wei等,2022),智能体技术也发展到了基于大语言模型的智能体这一阶段。

一个智能体的常规决策流程为:感知、规划与行动。感知是指从环境中收集信息并从中提取相关知识的能力;规划是指为了某一目标而作出的决策过程;行动是指基于环境和规划做出的动作。智能体接收来自环境的输入信息,并通过一些算法、推理来处理这些信息,然后根据处理结果进行相应的动作。智能体可以在没有持续人类输入的情况下完成多种任务,其应用场景有很多,如在软件开发中支持编码和调试工作,擅长以自然语言作为输入处理任务;在自动驾驶中为汽车提供车外环境信息,提供决策指导和支持车辆控制等。

本文借鉴基于大语言模型的智能体思想,结合多模态模型来完成用户定制意图理解任务,通过智能体减轻用户的记忆与操作负担。

1.2 图像显著性检测

图像显著性检测,又称视觉显著性检测,分为基于传统方法的显著性检测与基于深度学习的显著性检测。

传统方法一般是利用图像的低层视觉特征,判断目标区域和它周围像素的差异,进而计算图像区域的显著性,如Itti等人(1998)提出的一种基于高斯金字塔融合图像颜色、亮度和方向特征的视觉注意力模型,以及后来陆续提出的GB(graph-based)算法(Harel等,2007)、SR(spectral residual)算法(Hou和Zhang,2007)和FT(frequency-tuned)算法(Achanta等,2009)等。

基于深度学习的显著性检测是利用高级视觉特征,通过学习先验知识来决定视觉注意区域。早期基于深度学习的显著性目标检测方法通常使用卷积神经网络(LeCun等,1989)在超像素层面提取特征,进行目标显著性的判断,得到稀疏的显著性检测结果图(Li和Yu,2015;He等,2015)。随着全卷积神经网络(Long等,2015)在像素级语义分割中取得的巨大成功,基于像素级别的密集检测方法(Liu和Han,

2016; Hou 等, 2017; Liu 等, 2018)随之产生,并逐渐成为显著性检测的主流方法。最典型的模型如U-Net模型(Ronneberger 等, 2015),它对简单编解码结构的网络模型作出改进,在关注高层语义特征的同时,也关注低层的空间细节信息,强化了显著性目标的边界。后来不少学者也在U-Net结构上陆续做出改进(Zhou 等, 2018; Huang 等, 2020; He 等, 2022),U-Net结构在基于深度学习的显著性检测方法中具有重要意义。除了全卷积神经网络外,不少研究也将注意力机制用在显著性检测中(何伟和潘晨, 2022)。

本文中,基任务通过训练显著性检测模型实现,定制化任务中也用到了显著性检测思想。

1.3 小样本学习

普通的机器学习方法需要借助大量的监督样本(Bearman 等, 2016)才能训练出具有泛化能力的模型并应用到新数据中,而如何利用极少的标注数据学习到具有一定泛化能力的模型,在标注较为稀缺的现实场景中显得非常重要,这也是小样本学习需要解决的问题。

小样本学习的方法大致可分为基于模型的方法和基于度量的方法。基于模型的方法旨在通过模型结构的设计快速在少量样本上更新参数,直接建立输入和预测值的映射函数,如Santoro等人(2016)提出的使用记忆增强的方法。基于度量的方法的一般目标是学习相似度度量,相似样本对获得较高的相似度得分,而不相似样本对则获得较低的相似度得分。相似度度量可以是简单的距离测量、复杂的网络或其他可行的模块或算法。几种典型的基于度量的小样本学习方法包括类相关性度量(Fink, 2004)、KernelBoost(Hertz 等, 2006)、Siamese Net(Koch 等, 2015)等。Siamese Net是首个将深度神经网络用于小样本学习的工作。

本文借鉴度量学习的思想,从模板匹配的角度来解决小样本问题,将样本相似性定义为样本与模板图像之间的特征相似度,从而简单高效地解决小样本问题,避免模型训练带来的高昂代价和过拟合等问题。

2 本文方法

2.1 方法框架简介

用户在移动应用中上传平面图像进行扫描时,

用户意图区域形状通常近似于矩形,因此可以通过目标检测先获得意图区域的矩形边界框,再通过分割来进行细粒度调整。在此基础上,考虑到应用挑战及技术挑战,故本文构建了面向平面扫描图像的用户定制意图理解智能体,在应对新的定制化任务过程中无需进行网络参数的微调,具体设计如图2所示。

智能体包括任务感知、任务规划、任务执行以及反馈校正4个模块。任务感知从输入图像中提取信息,结合从存储的意图库中获得的用户定制模板信息,对输入图像进行意图理解;任务规划利用感知过程的结果进行规划,做出对输入图像的决策;任务执行按上一步的决策执行相应的动作,输出对图像的意图理解结果;若用户不满意当前输出结果,则用户可以进行反馈校正,随后智能体会输出新的结果并将校正的图像作为新的用户定制意图模板加入意图库中。

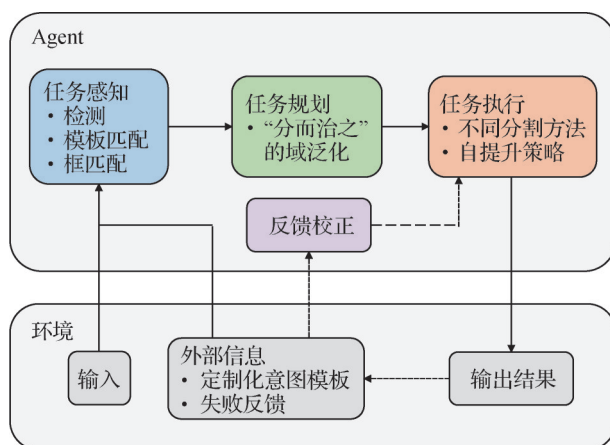


图2 面向平面扫描图像的用户定制意图理解智能体

Fig. 2 User-customized intention understanding agent for planar scanned images

面向平面扫描图像的用户定制意图理解智能体的感知—规划—执行的具体流程如图3所示。给定基域数据集 $I^B = \{(X_i^B, L_i^B)\}_{i=1}^{N^B}$ (上标“B”表示与基域(base)相关),其中, N^B 表示基域图像数量,图像 X_i^B 的标签为 L_i^B ,基域图像是存在单个意图的普通平面扫描图像,基任务是对单个意图的理解与识别。给定歧义数据集,即本文的目标域数据集 $I^T = \{(X_{ij}^T, L_{ij}^T)\}_{j=1}^{N_i^T}\}_{i=1}^C$ (上标“T”表示与目标域(target)相关),其中, C 表示目标域图像类别数, N_i^T 表示第 i

个类别的图像数量, $\mathbf{X}_{ij}^T$ 表示第 i 类的第 j 幅图像, 其标签为 L_{ij}^T , 目标域图像是存在多个潜在意图的平面扫描图像, 定制化任务(以下称目标任务)是在多个潜在意图中识别出用户真实意图区域。利用训练得到的基域分割模型 $F^B(X; \theta^B)$ 对目标域数据集 I^T 进行推理, θ^B 表示模型参数, 产生基础像素级掩膜 M^B ; 利用检测、模板匹配、框匹配和分割流程对目标域数据集 I^T 进行推理, 产生目标像素级掩膜 M^T ; 再将 M^B 和 M^T 通过自提升策略得到最终的高质量像素级掩膜 M^F 。同时, 将智能体应用于基域数据集 I^B 时得到的掩膜质量不低于直接使用基域分割模型得到的掩

膜质量。

面向平面扫描图像的用户定制意图理解智能体通过“分而治之”的域泛化、小样本角度的模板匹配、自提升策略和反馈校正模块来实现本文目标。“分而治之”的域泛化保证基任务与新任务互不影响且基任务性能不下降; 模板匹配简单高效地解决了任务存在的小样本问题, 实现在域泛化过程中网络参数无需微调; 自提升策略实现了基域和目标域知识的融合, 减少了预测结果的噪声; 反馈校正模块通过与用户的交互实现智能体自我更新完善, 提升智能体意图理解能力; 最终实现了对平面扫描图像的用户定制意图理解。

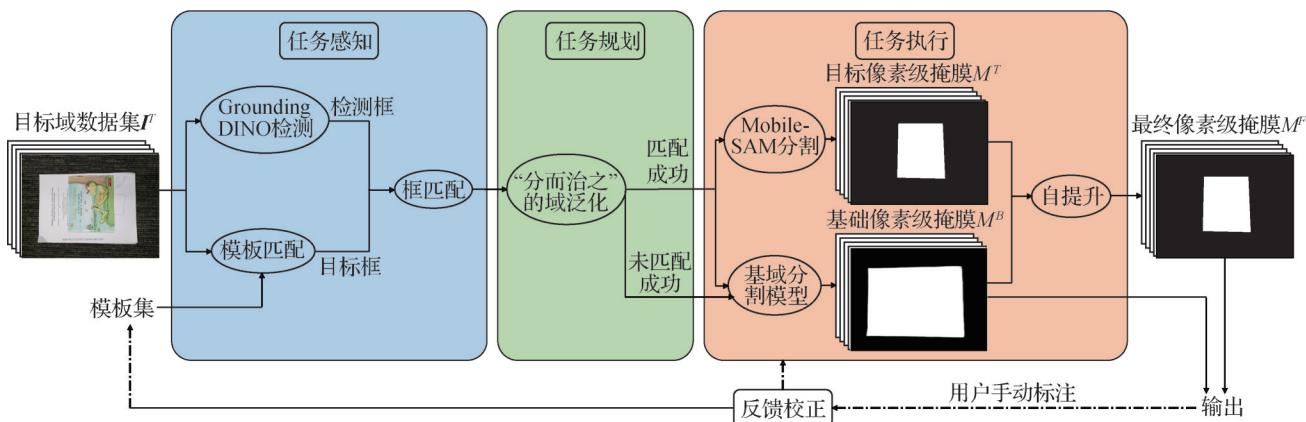


图3 用户定制意图理解智能体的运行流程

Fig. 3 Process of the user-customized intention understanding agent

2.2 任务感知

任务感知阶段的方法技术主要包括目标检测、模板匹配与框匹配, 目的是在面对小样本增量的挑战时理解任务, 即理解用户的意图区域。

目标域数据集 $I^T = \left\{ \left\{ \left(\mathbf{X}_{ij}^T, L_{ij}^T \right) \right\}_{j=1}^{N_i^T} \right\}_{i=1}^C$, 在实际场

景中, 用户对第1类到第 C 类的样本具有不同的意图, 因此每类样本是相对独立的, 且会不断出现新的 $C+1, C+2, \dots$ 类样本, 存在增量问题。另外, 大部分用户对标定样本的耐心有限, 对于同类样本用户希望只标注一次, 存在小样本问题。基于上述小样本增量问题的挑战, 本文将用户对图像的不同意图构建为不同的模板, 将意图理解定义为模板匹配问题。具体步骤为:

1) 构建用户意图库。从目标域数据集的每一类图像中取一幅加入模板图像集 $\mathbf{P} = \{p_i\}_{i=1}^C$, 共 C 个

模板, 再将每个模板进行尺度不变特征变换(scale-invariant feature transform, SIFT)后的描述子分别存储在文本文件里, 一共 C 个文本文件。包含 C 个模板图像、标签及 SIFT 文件的模板集即构成意图库。

2) 当输入一幅新图像时, 先进行检测。用视觉语言大模型 Grounding DINO(Liu 等, 2023)对输入图像的所有可能意图进行检测, 获得检测框集合 $\mathbf{box}_{det} = [det_0, det_1, \dots, det_k]$, 其中, $det_i, i = 0, \dots, k$ 表示每一个检测框的坐标。

3) 模板匹配。计算输入图像的 SIFT 特征描述子, 并分别计算此描述子与所有模板图像的描述子间的距离, 作为输入图像与每个模板间的相似度, 将相似度最高的一个模板作为输入图像对应的模板。从模板标签中获得意图区域的4个角点坐标, 对其进行变换: 如果输入图像与模板图像的角度不同, 将模板图像进行对应旋转, 使二者角度相同; 如果输入图像与模板图像的分辨率不同, 将模板图像进行对

应缩放,使二者尺度相同。模板图像变换后,标签位置也随之变换,变换后的位置信息就近似于输入图像中用户目标意图位置,以下称目标框,表示为 $\mathbf{box}_{\text{target}}$, 其一般形式为 $\mathbf{box}_{\text{target}} = [\text{external}, \text{inner}]$, 其中, external 表示目标外框坐标, inner 表示目标内框坐标。

4) 框匹配。将检测框与目标框利用匈牙利算法 (Kuhn, 1955) 进行匹配。用框之间的交并比 (intersection over union, IoU) 表示框之间的距离, 将检测框与目标框两两计算, 得到一个 $k+1$ 行 2 列的距离矩阵 $\mathbf{iou_matrix}$, 具体为

$$\mathbf{iou_matrix}[i, j] = \text{IoU}(\mathbf{box}_{\text{det}}[i], \mathbf{box}_{\text{target}}[j])$$

$$i = 0, 1, \dots, k; j = 0, 1 \quad (1)$$

再利用距离矩阵进行匈牙利匹配 (以下表示为 *HungarianMatching*), 即在检测框中寻找与目标外框和目标内框距离最小的两个框的组合, 得到初步的框匹配矩阵 $\mathbf{match_matrix}_{\text{temp}}$, 表示为

$$\mathbf{match_matrix}_{\text{temp}} = \text{HungarianMatching}(-\mathbf{iou_matrix}) \quad (2)$$

框匹配矩阵的一般形式为 $\begin{bmatrix} a & 0 \\ b & 1 \end{bmatrix}$, 每一行表示一个匹配对, 第 1 列表示检测框的索引, 第 2 列表示目标框的索引, 即 det_a 是与目标外框 external 位置最接近的检测框坐标, det_b 是与目标内框 inner 位置最接近的检测框坐标。在此基础上, 设定 IoU 阈值, 避免匹配的两个框之间 IoU 过低, 得到最终的框匹配矩阵 $\mathbf{match_matrix}$, 表示为

$$\mathbf{match_matrix}[0] = \mathbf{match_matrix}_{\text{temp}}[0]$$

$$\text{IoU}(\text{det}_a, \text{external}) > \beta \quad (3)$$

$$\mathbf{match_matrix}[1] = \mathbf{match_matrix}_{\text{temp}}[1]$$

$$\text{IoU}(\text{det}_b, \text{inner}) > \beta \quad (4)$$

式中, $\mathbf{match_matrix}[0]$ 和 $\mathbf{match_matrix}[1]$ 分别表示矩阵的第 1 行和第 2 行。 β 为设定的 IoU 阈值。只有当每对匹配框之间的 $\text{IoU} > \beta$ 时, 才认为是真正匹配成功。

2.3 任务规划

根据任务感知阶段对任务的理解结果, 任务规划阶段采用“分而治之”的域泛化方法对输入图像进行决策。

由于基域和目标域图像特点不同且差异较大, 因此很难用同样的方法同时在基任务和目标任务上

实现较好的性能。针对这个问题, 本文提出“分而治之”的域泛化方法, 即将不同任务的推理解耦, 以使它们互不影响, 在提高新任务性能的同时避免对基任务的遗忘。“分而治之”的域泛化具体实现如下。

在 2.2 节任务感知阶段的框匹配步骤后, 若匹配矩阵为 $\begin{bmatrix} a & 0 \\ b & 1 \end{bmatrix}$ 的形式, 则选择与用户意图对应的 det_a 或 det_b 作为匹配成功的框。另外, 由于设定了 IoU 阈值, 故存在两种特殊情况: 1) 框匹配矩阵为空: 表示没有达到 IoU 阈值的匹配对, 说明输入图像的实际意图区域与模板中的意图区域完全不匹配, 故认为此图像不属于目标域, 判定其为基域图像。2) 框匹配矩阵只有一行: 表示某个匹配对由于 IoU 过低被过滤掉了, 导致只有一个成功的匹配对。假设目标区域是内部意图区域, 而只成功匹配到了外框, 为增强鲁棒性, 此时就直接选择检测框中面积最大的框作为匹配成功的框。

通过以上对框匹配是否成功的判断以及对框的选择, 就可做出决策, 将基域与目标域的图像区分开, 在后续的执行阶段采用不同的方法进行处理, 即可保证基任务和目标任务互不影响。

2.4 任务执行

依据任务规划阶段做出的决策, 任务执行阶段对输入图像采取相应的行动: 若输入图像属于基域, 利用训练得到的基域分割模型 $F^B(X; \theta^B)$ 对图像进行分割即可完成, 不需要再进行其他步骤, 避免冗余和耗时; 若输入图像属于目标域, 利用任务规划阶段中匹配成功的检测框作为提示框 (prompt), 用 Mobile-SAM 模型 (Zhang 等, 2023) 对图像进行分割提高细粒度, 得到掩膜 M^T , 再与图像经过基域分割, 模型 $F^B(X; \theta^B)$ 得到的掩膜 M^B 一起执行图 3 中自提升策略, 即可获得最终掩膜 M^F 。

自提升策略的目标是为了减少意图理解结果的噪声, 提高域泛化的可靠性。本文智能体是由大模型和传统算法组合而成, 由于目标域数据集没有经过微调, 因此将大模型直接应用在数据集上会出现噪声, 影响目标任务的效果。为此, 本文提出自提升策略, 通过双域融合减少图像噪声, 提高掩膜质量。自提升策略如图 3 橙色框右侧所示, 输入图像经过分割后产生两种掩膜, 即目标域数据集 $I^T = \left\{ \left\{ (X_{ij}^T, L_{ij}^T, M_{ij}^B, M_{ij}^T) \right\}_{j=1}^{N_i^T} \right\}_{i=1}^C$, 其中, M_{ij}^B 表示 $F^B(X; \theta^B)$

对图像 X_y^T 的像素级分割结果, M_y^T 表示 Mobile-SAM 对图像 X_y^T 的像素级分割结果。 M_y^B 的特点是其为一个连通区域,但往往与目标意图区域相比范围过大; M_y^T 的特点是它能比较准确地描绘出目标意图区域的轮廓,但其边界和内部往往存在噪点,且可能出现分割出的区域不连通的情况。因此,可将 M_y^T 和基域知识 M_y^B 结合起来,简单有效的方法是取它们的交集,即逐一对比选择相同的像素点,得到融合掩膜 $M_y^A = M_y^B \cap M_y^T$,减少了背景部分的噪声,同时对前景部分也不会产生影响。然后,在 M_y^A 中获得层级轮廓,取最大轮廓作为目标轮廓,将其内部所有像素都视为前景部分,以解决内部可能出现的中空等问题,去除前景部分的噪声。经过以上处理后得到最终像素级掩膜 M_y^F ,即为目标域图像 X_y^T 的意图理解结果。

如上所述,自提升策略包括融合和后处理两个步骤。在自提升策略的作用下,目标域图像的掩膜质量相比使用单一基域分割模型时有一定提高,自提升策略对提升目标任务的性能起到了一定的促进作用。

2.5 反馈校正模块

当用户认为智能体输出结果有误时,反馈校正模块发挥作用。如图3中紫色部分及虚线所示,用户在图像上对意图区域进行手动标注,智能体会以此标注作为提示框,再次通过任务执行阶段进行分割与自提升后,输出新的意图理解结果。同时,智能体会按2.2节中构建模板的方式学习该图像及用户标注,构建一个新的定制意图模板,加入意图库中,以便后续使用,从而意图库的大小及模板图像特征不断得到更新。

反馈校正模块通过与用户的交互不断学习新知识,实现智能体的自我更新完善,提升智能体的用户定制意图理解能力。

3 实验分析

3.1 数据集与评价指标

1)目标域数据集。即歧义数据集,是本文提出的用于平面扫描图像定制意图理解的基准数据集,该数据集的图像为具有多个潜在意图的平面扫描图像。数据集共2 241幅图像,分为112组增量子数据,每组子数据含20幅图像左右,均是对相同对象在不同角度或距离下拍摄得到的图像,拍摄对象包

括书本、文档、纸张、插画和卡证等。不同组子数据间图像的潜在意图区域个数及分布各不相同。图4展示了一部分具有代表性的图像,其中每幅图像属于不同的组别。另外,数据集中每幅图像均有一个对应的标注文件,内含标注状态、图像尺寸、旋转角度和多个潜在意图区域的角点坐标等。基于图像的多样性和标签的完整性,该数据集适用范围较广,可以作为多种视觉任务(如目标检测、语义分割和实例分割等)的基准数据集。

2)模板集。即意图库,在目标域数据集每组子数据中取一幅图像作为模板图像,共112幅图像,按2.2节中的方法构建模板集。

3)测试集。在目标域数据集中除去模板图像、无标注图像和标注不全的图像,其余作为测试集,测试集共含2 031幅图像。在测试集上对智能体的精度进行评估。

4)评价方式。本文采用以上测试集进行评价,评价前根据图像标签(目标区域的角点坐标)生成标签对应的二值化掩膜,即将4个角点连线后形成的四边形目标框内部填充为1,外部填充为0。评价时根据图像的预测结果及图像对应的标签掩膜进行精度评估。评价指标有4种:平均交并比(mean intersection over union, mIoU)、平均像素准确率(mean pixel accuracy, mPA)、总正确率(total accuracy, TA)和组内正确率(accuracy of each group, AoG)。

mIoU 是图像预测结果与标签掩膜的交集和并集的比值,mPA 是每个类被正确分类像素数的比例的平均值,计算为

$$mIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{TP}{TP + FN + FP} \quad (5)$$

$$mPA = \frac{1}{k+1} \sum_{i=0}^k \frac{TP}{TP + FP} \quad (6)$$

式中, k 表示类别数(不含背景类), TP (true positive)为将正类预测为正类的像素点个数, FN (false negative)为将正类预测为负类的像素点个数, FP (false positive)为将负类预测为正类的像素点个数。

TA 是测试集中正确分割的图像占比,此处“正确分割的图像”定义为 $mIoU > r$ 的图像,具体为

$$TA = \frac{S_{(mIoU > r)}}{N} \quad (7)$$

式中, r 为设定的 mIoU 阈值, S 表示正确分割的图像数, N 表示测试集图像数。

AoG_i 是第*i*组图像中正确分割的图像数占测试图像数的比例,计算为

$$AoG_i = \frac{S_{i(mIoU>r)}}{N_i}, \quad i = 1, 2, \dots, 112 \quad (8)$$

式中, S_i 表示组内正确分割的图像数, N_i 表示该组的测试图像数。

3.2 实验细节与结果

本文实验在 24 GB 显存的 NVIDIA A10 上进行。在基域数据集上对 U²-Net 模型(Qin 等,2020)进行训练得到基域分割模型。本文智能体中的检测和分割流程通过预训练好的 Grounding DINO 模型(Liu 等,2023)和 Mobile-SAM 模型(Zhang 等,2023)完成,模型无需训练即可直接使用。检测参数包括文本提示词“paper. illustration.”,框阈值 box_threshold = 0.05,文本阈值 text_threshold = 0.2。构建意图库时先将模板图像尺寸调整为 400 × 300 像素,再使用 OpenCV 中 SIFT 相关库函数计算图像的 SIFT 特征描述子存到文本文件中。对新的输入图像进行模板匹配时,同样将它调整到尺寸为 400 × 300 像素后计算其特征描述子,再将输入图像的特征与每幅模板图像的特征进行*k*近邻匹配,设置*k*为 2,得到所有匹配数,设置距离阈值为 0.75,计算每对匹配点间的距离,小于此阈值的定义为良好匹配对,良好匹配对数量与所有匹配数的比值即为输入图像与某幅模板图像的相似性,并将其记录下来。计算完输入图像与所有模板图像的相似性后,选择相似性最大的一幅模板图像用于后续过程。框匹配时 IoU 阈值设为 0.3,两框之间的 IoU 大于此阈值才认为真正匹配成功。Mobile-SAM 分割时以匹配成功的检测框作为提示框。对得到的分割结果进行自提升时,先以按位与的方式将结果与基域分割模型的结果进行融合,再使用 OpenCV 中 findContours 函数进行后处理,轮廓检索模式设置为 RETR_TREE,轮廓表达模式设置为 CHAIN_APPROX_SIMPLE,获得所有轮廓后,取最大轮廓并将其内部所有像素点的值都设置为 1(即前景)。

计算总正确率 TA 和组内正确率 AoG 时,将*r*设为 90%,并规定若某组的 AoG > 60%,则认为该组的测试结果是达标的,即:若一组图像内有超过 60%的测试结果的 mIoU 大于 90%,则该组达标。

实验结果如下:本文智能体在目标域数据集的测试集上的 mIoU 为 90.47%,mPA 为 94.33%,TA 为

68.78%,AoG > 60% 的组数为 75 组。

3.3 消融实验

为深入探索面向平面扫描图像的用户定制意图理解智能体,本文针对“分而治之”域泛化方法的应用、目标检测参数框阈值以及自提升策略进行了消融实验。

3.3.1 “分而治之”域泛化方法的应用

为探索是否解耦以及解耦阶段对最终性能的影响,本文进行了以下实验:1)不采用“分而治之”方法,对基域和目标域数据集使用完全相同的执行过程;2)在模板匹配阶段后使用“分而治之”方法,并设定不同阈值来进行图像过滤。最终智能体在基域数据集和目标域数据集上的性能如表 1 所示。

表 1 “分而治之”应用于不同阶段的实验结果
Table 1 Experimental results of “divide and conquer” in different stages

		/%	
“分而治之”应用阶段	阈值	mIoU	
		目标域数据集	基域数据集
无	—	90.47	48.72
	0.10	86.88	82.05
	0.14	82.42	88.71
模板匹配	0.15	81.10	90.57
	0.20	74.93	93.34
	0.30	65.80	95.31
模板匹配+框匹配	—	90.47	95.69

注:加粗字体表示各列最优结果。“—”表示没有设置。

首先,在表 1 中,相较于不使用“分而治之”方法(第 1 行),使用“分而治之”的推理能有效地解耦基域和目标域图像,避免在增量过程中出现灾难性遗忘导致基任务性能下降,“分而治之”使得基域数据集的性能大幅提高(第 2—7 行)。其次,当“分而治之”用于模板匹配阶段后,随着阈值的增加,目标域数据集的性能会逐步衰减(第 2—6 行),原因是部分目标域图像被错误地过滤了,没有正确执行它原本对应的流程。最后,在模板匹配和框匹配阶段后采用“分而治之”的方法,能获得更多的图像特征,从而更好地进行解耦,在基域和目标域上都实现更好的性能(第 7 行)。

3.3.2 框阈值

设置实验探索目标检测参数中框阈值 `box_threshold` 对智能体在目标域数据集上性能的影响。`box_threshold` 是 Grounding DINO 模型检测时设置的框阈值,当某个候选框与设置的文本提示词的最高相似度高于此阈值时,该候选框就会被选择,作为检测框输出。不同 `box_threshold` 下的实验结果如表 2 所示。由表 2 可知,当检测参数 `box_threshold` 逐渐降低,测试集的 mIoU 会逐渐增大。原因是 `box_threshold` 越小,满足条件的检测框就越多,从而检测框里有更大概率包含与需要的目标框位置最接近的框,即有更大概率能获得目标框,最终以此框作为提示框进行分割时性能就会更好。

表 2 检测参数 (`box_threshold`) 对最终性能的影响
Table 2 Impact of the detection parameter
(`box_threshold`) on final performance

/%		
box_threshold	测试集 mIoU	测试集 TA
0.3	86.90	57.61
0.2	88.57	62.63
0.1	90.24	67.60
0.05	90.47	68.78

注:加粗字体表示各列最优结果。

3.3.3 自提升策略

设置实验探索不同自提升策略对智能体在目标域数据集上性能的影响,得到如表 3 所示的结果。

表 3 不同自提升策略的实验结果

Table 3 Experimental results of different self-improvement strategies

/%		
自提升策略	测试集 mIoU	测试集 TA
无	89.28	64.40
融合	89.43	64.65
后处理	90.42	68.44
融合+后处理	90.47	68.78

注:加粗字体表示各列最优结果。

由表 3 可知,单独采用融合(第 2 行)或单独采用后处理(第 3 行)时,测试集的性能相比于不使用自提升策略(第 1 行)都有一定的提高,而同时采用融合与后处理的自提升策略(第 4 行)时,测试集性能有进一步的提升,原因是融合过程通过结合基域知识,减少了背景部分的噪声,后处理过程主要减少了前景部分的噪声,这两个部分是相互补充完善的。本文采用的自提升策略对提高目标任务的性能起到了促进作用。

3.4 对比实验与可视化结果

针对平面扫描图像的用户定制意图理解任务,本文构建的智能体与前沿的目标检测和分割模型在目标域数据集上的性能比较如表 4 所示。

本文方法共与 7 种方法进行了比较,其中 U²-Net (Qin 等, 2020)是常规的显著性检测模型,它只能对普通平面扫描图像进行意图理解,不适用于歧义图像的定制意图理解;U²-Net 调整权重是在已训练得

表 4 不同模型在本文提出的数据集上的测试性能对比

Table 4 Performance comparison of different models on proposed dataset

/%				
方法模型	是否需要训练	提示类型	测试集 mIoU	测试集 TA
U ² -Net	是	\	59.72	3.94
U ² -Net 调整权重	是	\	72.82	7.98
SAM	否	点	56.13	21.22
Mobile-SAM	否	点	59.19	24.03
SegGPT	否	一幅图像及标注	74.87	46.68
CLIPSeg	否	文本	52.20	1.48
Grounded-SAM	否	文本	72.60	33.87
本文智能体	否	文本	90.47	68.78

注:加粗字体表示各列最优结果,“\”表示没有提示。

到的模型基础上,输入歧义图像及其定制意图区域掩膜进行继续训练,由于定制样本数量少,继续训练很难改变模型已有权重,故这个方法的效果也不理想;SegGPT(Wang等,2023b)是无需训练即可使用的分割模型,SAM(segment anything model)(Kirillov等,2023)、Mobile-SAM(Zhang等,2023)、CLIPSeg(Lüddecke和Ecker,2022)和Grounded-SAM(IDEA-Research,2023)是无需训练即可使用的视觉语言模型,可根据点、框和文本等多种提示类型直接进行推理,但它们的推理结果同样是不精确的。特别地,本文方法相较于同样使用文本作为提示词的前沿的Grounded-SAM模型,性能有明显提升,在测试集上的mIoU高出17.87%,测试集中正确分割的图像占

比高出一倍。本文方法的mIoU在所有方法中唯一超过了90%,总正确率TA在所有方法中唯一超过了60%。综上,表4结果说明相较于现今可用于平面扫描图像的用户意图理解方法,本文提出的智能体在规避大规模复杂的训练模式的同时,在测试集上取得了更优的性能。

图4是本文智能体在目标域数据集上的可视化结果,图中红色框表示用户目标意图区域,图像高亮处表示意图理解结果。图4展现了本文智能体在图像中存在单个意图、两个及多个潜在意图的情况下,意图理解结果都与用户目标意图区域十分接近,说明智能体在图像存在意图干扰时也能理解识别出真实意图区域,在用户定制意图理解任务上具有较好的性能。



图4 智能体在本文提出的数据集上的可视化结果
Fig. 4 Visualization results of the agent on proposed dataset

3.5 公开数据集实验

由于面向平面扫描图像的用户定制意图理解任务是图像处理领域的一个新型任务,除了本文提出的新的基准数据集外,目前还没有适用于该任务的既有公开基准数据集。因此,为了验证本文智能体的泛化能力,将其在一个公开票据数据集上进行实验,并与先进方法进行对比。

CORD(consolidated receipt dataset)数据集(Park等,2019)是一个票据数据集,包含1 000幅来自杂货店的票据图像,主要用于文档智能领域的任务。CORD数据集标签中具有感兴趣区域(region of interest, ROI)关键字,标注了图像中票据区域的位置,可用于票据理解任务,类似于本文的用户意图理解任务,故可使用该数据集评估本文智能体的性能,

且该数据集图像具有倾斜、光照、畸变等干扰因素,也有助于验证智能体的鲁棒性。

CORD数据集中部分图像标签不全,故先进行数据清洗,只保留标签中具有ROI关键字的图像,共346幅,从中选择100幅图像构建模板集,剩余246幅图像作为测试集。

针对CORD数据集上的票据理解任务,本文构建的智能体与前沿的目标检测和分割模型的性能比较如表5所示。在CORD数据集上,本文方法共与4种方法进行了比较,均是无需训练,可直接对图像进行推理的方法。当输入一幅新图像进行推理时,对于SAM和Mobile-SAM模型,使用的提示是图像的中点,对于CLIPSeg、Grounded-SAM模型以及本文智能体,使用的提示是文本提示词“receipt”。表5显示

本文智能体在 CORD 测试集上的 mIoU 达到 86.52%, 性能高于其他 4 种方法, 证明了智能体的泛化能力和鲁棒性。

表 5 不同模型在 CORD 数据集上的测试性能对比
Table 5 Performance comparison of different models on CORD dataset

方法模型	提示类型	测试集 mIoU /%
SAM	点	60.80
Mobile-SAM	点	66.64
CLIPSeg	文本	79.50
Grounded-SAM	文本	85.83
本文智能体	文本	86.52

注: 加粗字体表示最优结果。

4 结 论

对于移动端应用中平面扫描图像的用户意图理解任务, 传统方法无法解决其中的定制化和小样本增量等问题, 适用范围有限。随着人工智能技术的发展, 意图理解已经变得越来越智能化与自主化, 传统方法逐渐不能满足用户定制的、多样的需求。针对这一场景, 本文提出了面向平面扫描图像的用户定制意图理解智能体, 用“分而治之”的域泛化方法将不同任务解耦, 提升了任务域泛化能力; 从模板匹配角度解决了小样本增量问题, 回避了网络微调, 降低了训练开销; 通过自提升策略减少了预测结果的噪声; 通过反馈校正模块与用户进行交互, 实现智能体的自我更新, 提升定制意图理解能力。

针对该任务, 本文还提出了一个平面扫描图像定制意图理解基准数据集, 并在数据集上进行了消融实验与对比实验。以与 SegGPT 方法的对比为例, 本文智能体同样无需训练, 但 mIoU 相比提升了 15.60%, TA 相比提升了 22.10%。与最新的 Grounded-SAM 方法相比, mIoU 提升了 17.87%, TA 提升了一倍。在所有对比的方法中, 本文方法的 mIoU 唯一达到了 90%, TA 唯一达到了 60%。实验证明了本文智能体的有效性, 实现了对用户定制化的、精准的意图理解。另外, 在公开数据集 CORD 上也进行了对比实验, 证明了智能体的泛化能力和鲁

棒性。

然而, 实验结果中也存在某些类别的意图理解精度较低的现象, 原因在于内部意图和外部意图位置非常接近时, Grounding DINO 检测效果和框匹配效果都会受到较大的负面影响。未来可以通过改进检测时的提示词设计和改进框匹配算法来优化。

本文将智能体概念引入到用户意图理解领域, 并将智能体与环境的交互模式应用在用户与移动应用的交互上, 对于平面扫描图像的用户定制意图理解任务提供了新的方向。

参考文献 (References)

- Achanta R, Hemami S, Estrada F and Sussstrunk S. 2009. Frequency-tuned salient region detection//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 1597-1604 [DOI: 10.1109/CVPR.2009.5206596]
- Bearman A, Russakovsky O, Ferrari V and Li F F. 2016. What's the point: semantic segmentation with point supervision//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 549-565 [DOI: 10.1007/978-3-319-46478-7_34]
- Fink M. 2004. Object classification from a single example utilizing class relevance metrics//Proceedings of the 17th International Conference on Neural Information Processing Systems. Vancouver British, Canada: MIT Press: 449-456
- Green S, Hurst L, Nangle B, Cunningham P, Somers F and Evans R. 1997. Software agents: a review. Tech. Rep. TCS-CS-1997-06. Department of Computer Science, Trinity College Dublin
- Harel J, Koch C and Perona P. 2007. Graph-based visual saliency//Proceedings of the 19th International Conference on Neural Information Processing Systems. Canada: MIT Press: 545-552
- He S F, Lau R W H, Liu W X, Huang Z and Yang Q X. 2015. Super-CNN: a superpixelwise convolutional neural network for salient object detection. International Journal of Computer Vision, 115(3): 330-344 [DOI: 10.1007/s11263-015-0822-0]
- He W and Pan C. 2022. The salient object detection based on attention-guided network. Journal of Image and Graphics, 27(4): 1176-1190 (何伟, 潘晨. 2022. 注意力引导网络的显著性目标检测. 中国图象图形学报, 27(4): 1176-1190) [DOI: 10.11834/jig.200658]
- He X, Zhou Y, Zhao J Q, Zhang D, Yao R and Xue Y. 2022. Swin Transformer embedding UNet for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 60: #4408715 [DOI: 10.1109/TGRS.2022.3144165]
- Hertz T, Hillel A B and Weinshall D. 2006. Learning a kernel function for classification with small training samples//Proceedings of the

- 23rd International Conference on Machine Learning. Pittsburgh, USA: ACM: 401-408 [DOI: 10.1145/1143844.1143895]
- Hou Q B, Cheng M M, Hu X W, Borji A, Tu Z W and Torr P. 2017. Deeply supervised salient object detection with short connections//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5300-5309 [DOI: 10.1109/CVPR.2017.563]
- Hou X D and Zhang L Q. 2007. Saliency detection: a spectral residual approach//Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE: 1-8 [DOI: 10.1109/CVPR.2007.383267]
- Huang H M, Lin L F, Tong R F, Hu H J, Zhang Q W, Iwamoto Y, Han X H, Chen Y W and Wu J. 2020. UNet 3+: a full-scale connected UNet for medical image segmentation//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 1055-1059 [DOI: 10.1109/icassp40776.2020.9053405]
- IDEA-Research. 2023. Grounded-segment-anything [EB/OL]. [2024-01-04].
<https://github.com/IDEA-Research/Grounded-Segment-Anything>
- Itti L, Koch C and Niebur E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11): 1254-1259 [DOI: 10.1109/34.730558]
- Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, Xiao T T, Whitehead S, Berg A C, Lo W Y, Dollár P and Girshick R. 2023. Segment anything [EB/OL]. [2024-01-04].
<https://arxiv.org/pdf/2304.02643.pdf>
- Koch G, Zemel R and Salakhutdinov R. 2015. Siamese neural networks for one-shot image recognition//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: [s.n.]
- Kuhn H W. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1/2): 83-97 [DOI: 10.1002/nav.3800020109]
- LeCun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W and Jackel L D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4): 541-551 [DOI: 10.1162/neco.1989.1.4.541]
- Li G B and Yu Y Z. 2015. Visual saliency based on multiscale deep features//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 5455-5463 [DOI: 10.1109/CVPR.2015.7299184]
- Li G H, Al Kader Hammoud H A, Itani H, Khizbullin D and Ghanem B. 2023. CAMEL: communicative agents for “mind” exploration of large language model society [EB/OL]. [2024-01-04].
<https://arxiv.org/pdf/2303.17760.pdf>
- Liu N and Han J W. 2016. DHSNet: deep hierarchical saliency network for salient object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 678-686 [DOI: 10.1109/CVPR.2016.80]
- Liu N, Han J W and Yang M H. 2018. PiCANeT: learning pixel-wise contextual attention for saliency detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3089-3098 [DOI: 10.1109/CVPR.2018.00326]
- Liu S L, Zeng Z Y, Ren T H, Li F, Zhang H, Yang J, Jiang Q, Li C Y, Yang J W, Su H, Zhu J and Zhang L. 2023. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection [EB/OL]. [2024-01-04].
<https://arxiv.org/pdf/2303.05499.pdf>
- Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]
- Lüddecke T and Ecker A. 2022. Image segmentation using text and image prompts//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 7076-7086 [DOI: 10.1109/CVPR52688.2022.00695]
- Nakano R, Hilton J, Balaji S, Wu J, Ouyang L, Kim C, Hesse C, Jain S, Kosaraju V, Saunders W, Xu J, Cobbe K, Eloundou T, Krueger G, Button K, Knight M, Chess B and Schulman J. 2021. WebGPT: browser-assisted question-answering with human feedback [EB/OL]. [2024-01-04]. <https://arxiv.org/pdf/2112.09332.pdf>
- Park S, Shin S, Lee B, Lee J, Surh J, Seo M and Lee H. 2019. CORD: a consolidated receipt dataset for post-OCR parsing//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: [s.n.]
- Qin X B, Zhang Z C, Huang C Y, Dehghan M, Zaiane O R and Jagersand M. 2020. U²-Net: going deeper with nested U-structure for salient object detection. *Pattern Recognition*, 106: #107404 [DOI: 10.1016/j.patcog.2020.107404]
- Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]
- Santoro A, Bartunov S, Botvinick M, Wierstra D and Lillicrap T. 2016. Meta-learning with memory-augmented neural networks//Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR.org: 1842-1850
- Wang G Z, Xie Y Q, Jiang Y F, Mandlekar A, Xiao C W, Zhu Y K, Fan L X and Anandkumar A. 2023a. Voyager: an open-ended embodied agent with large language models [EB/OL]. [2024-01-04]. <https://arxiv.org/pdf/2305.16291.pdf>
- Wang X L, Zhang X S, Cao Y, Wang W, Shen C H and Huang T J. 2023b. SegGPT: segmenting everything in context [EB/OL]. [2024-01-04]. <https://arxiv.org/pdf/2304.03284.pdf>
- Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S,

- Yogatama D, Bosma M, Zhou D, Metzler D, Chi E H, Hashimoto T, Vinyals O, Liang P, Dean J and Fedus W. 2022. Emergent abilities of large language models [EB/OL]. [2024-01-04]. <https://arxiv.org/pdf/2206.07682.pdf>
- Xi Z H, Chen W X, Guo X, He W, Ding Y W, Hong B Y, Zhang M, Wang J Z, Jin S J, Zhou E Y, Zheng R, Fan X R, Wang X, Xiong L M, Zhou Y H, Wang W R, Jiang C H, Zou Y C, Liu X Y, Yin Z Y, Dou S H, Weng R X, Cheng W S, Zhang Q, Qin W J, Zheng Y Y, Qiu X P, Huang X J and Gui T. 2023. The rise and potential of large language model based agents: a survey [EB/OL]. [2024-01-04]. <https://arxiv.org/pdf/2309.07864.pdf>
- Yao S Y, Zhao J, Yu D, Du N, Shafran I, Narasimhan K and Cao Y. 2022. ReAct: synergizing reasoning and acting in language models [EB/OL]. [2024-01-04]. <https://arxiv.org/pdf/2210.03629.pdf>
- Zhang C, Jiang W Y, Chen S Y, Zhou W and Yan F T. 2023. Multi-agent path planning based on improved double DQN. *Journal of Image and Graphics*, 28(7): 2167-2181 (张晨, 蒋文英, 陈思源, 周文, 闫丰亭. 2023. 基于双层DQN的多智能体路径规划. *中国图象图形学报*, 28(7): 2167-2181) [DOI: 10.11834/jig.211239]
- Zhang C N, Han D S, Qiao Y, Kim J U, Bae S H, Lee S and Hong C S. 2023. Faster segment anything: towards lightweight SAM for mobile applications [EB/OL]. [2024-01-04]. <https://arxiv.org/pdf/2306.14289.pdf>
- Zhou Z W, Siddiquee M M R, Tajbakhsh N and Liang J M. 2018. UNet++: a nested U-Net architecture for medical image segmentation//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018. Granada, Spain: Springer: 3-11 [DOI: 10.1007/978-3-030-00889-5_1]

作者简介

冯弋珂,女,硕士研究生,主要研究方向为计算机视觉和智能体。E-mail: wuhen836973250@163.com

励雪巍,通信作者,男,博士研究生,主要研究方向为计算机视觉、全景图像理解、知识蒸馏和场景图生成。

E-mail: xueweili@zju.edu.cn

刘鹏伟,男,硕士,主要研究方向为计算机视觉、图像处理和深度学习。E-mail: pengwei_liu@intsig.net

郭丰俊,男,高级工程师,主要研究方向为图像处理、文档图像分析与理解。E-mail: fengjun_guo@intsig.net

龙腾,男,博士,主要研究方向为计算机视觉和机器学习。

E-mail: mike_long@intsig.net

李玺,男,教授,主要研究方向为计算机视觉、模式识别和深度学习。E-mail: xilizju@zju.edu.cn