中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2023)08-2253-23

论文引用格式:Zhang Y, Li Q, Shen H W, Zeng G Y, Zhou Y, Ma C, Zhang Y and Wang W P. 2023. Text-centric image analysis techniques: a crtical review. Journal of Image and Graphics, 28(08):2253-2275(张言, 李强, 申化文, 曾港艳, 周宇, 马灿, 张远, 王伟平. 2023. 以文字为中心的图像理解技术综述. 中国图象图形学报, 28(08):2253-2275)[DOI:10.11834/jig.220968]

以文字为中心的图像理解技术综述

张言[1,2],李强[1,2],申化文[1,2],曾港艳[3],周宇[1,2*],马灿[1,2],张远[3],王伟平[1,2]

1. 中国科学院信息工程研究所,北京　100093; 2. 中国科学院大学网络空间安全学院,北京　101408;
3. 中国传媒大学媒体融合与传播国家重点实验室,北京　100024

摘　要:文字广泛存在于各种文档图像和自然场景图像之中,蕴含着丰富且关键的语义信息。随着深度学习的发展,研究者不再满足于只获得图像中的文字内容,而更加关注图像中文字的理解,故以文字为中心的图像理解技术受到越来越多的关注。该技术旨在利用文字、视觉物体等多模态信息对文字图像进行充分理解,是计算机视觉和自然语言处理领域的一个交叉研究方向,具有十分重要的实际意义。本文主要对具有代表性的以文字为中心的图像理解任务进行综述,并按照理解认知程度,将以文字为中心的图像理解任务划分为两类,第1类仅要求模型具备抽取信息的能力,第2类不仅要求模型具备抽取信息的能力,而且要求模型具备一定的分析和推理能力。本文梳理了以文字为中心的图像理解任务所涉及的数据集、评价指标和经典方法,并进行对比分析,提出了相关工作中存在的问题和未来发展趋势,希望能够为后续相关研究提供参考。

关键词:文字图像理解;视觉信息抽取;场景文字图像检索;文档视觉回答;场景文字视觉问答;场景文字图像描述

Text-centric image analysis techniques: a crtical review

Zhang Yan[1,2], Li Qiang[1,2], Shen Huawen[1,2], Zeng Gangyan[3], Zhou Yu[1,2*], Ma Can[1,2], Zhang Yuan[3], Wang Weiping[1,2]

1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing　100093, China;
2. School of Cyber Security, University of Chinese Academy of Sciences, Beijing　101408, China;
3. State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing　100024, China

Abstract: Text can be as one of the key carriers for information transmission. Digital media-related text has been widely developing for such image aspects of document and scene contexts. To extract and analyze these text information-involved images automatically, Conventional researches are mainly focused on automatic text extraction techniques like scene text detection and recognition. However, text-centric images-based semantic information recognition or analysis as a downstream task of spotting text, remains a challenge due to the difficulty of fully leveraging multi-modal features from both vision and language. To this end, text-centric image understanding has been an emerging research topic and many related tasks have been proposed. For example, the visual information extraction technique is capable of extracting the specified content from the given image, which can be used to improve productivity in finance, social media, and other fields. In this paper, we introduce five representative text-centric image understanding tasks and conduct a systematic survey on them. According to the understanding level, these tasks can be broadly classified into two categories. The first category requires

收稿日期:2022-09-26;**修回日期**:2022-12-23;**预印本日期**:2022-12-30

*通信作者:周宇　zhouyu@iie.ac.cn

基金项目:中国科学院基础前沿科学研究计划从0到1原始创新项目(ZDBS-LY-7024)

Supported by: Key Research Program of Frontier Sciences, Chinese Academy of Sciences(CAS)(ZDBS-LY-7024)

the basic understanding ability to extract and distinguish information, such as visual information extraction and scene text retrieval. In contrast, besides the fundamental understanding ability, the second category is more concerned with high-level semantic understanding capabilities like information aggregation and logical reasoning. With the research progress in deep learning and multimodal learning, the second category has attracted considerable attention recently. For the second category, this survey mainly introduces document visual question answering, scene text visual question answering, and scene text image captioning tasks. Over the past few decades, the development of text-centric image understanding techniques has gone through several stages. Earlier approaches are based on heuristic rules and may only utilize unimodal features. Currently, deep learning methods have gained wide popularity and dominated this area. Meanwhile, multimodal features are valued and exploited to improve performance. To be more specific, traditional visual information extraction depends on pre-defined templates or specific rules. Traditional text retrieval task tends to represent words with pyramid histograms of character vectors and predict the matched image according to the representation distance. Expanded from the conventional visual question answering framework, earlier document visual question answering, and scene text visual question answering approaches simply add an optical character recognition branch to extract text information. As integrating knowledge from multimodal signals helps to better understand images, graph neural networks and Transformer-based frameworks are used to fuse multi-modal features recently. Furthermore, self-supervised pre-training schemes are applied to learn the alignment between different modalities, thus boosting model capabilities by a large margin. For each text-centric image understanding task, we summarize classical methods and further elaborate the pros and cons of them. In addition, we also discuss the potential problems and further research directions for the community. Firstly, due to the complexity of different modality features, such as mutative layout and diverse fonts, current deep learning architectures still fail to complete the interaction of multi-modal information efficiently. Secondly, existing text-centric image understanding methods are still limited in their reasoning abilities, involving counting, sorting, and arithmetic operations. For instance, in document visual question answering and scene text visual question answering tasks, current models have difficulty predicting accurate answers when they require to jointly reason over image layout, textual content, and visual art, etc. Finally, the current text-centric understanding tasks are often trained independently and the correlation between different tasks has not been effectively leveraged. We hope this survey can help researchers capture the latest progress in text-centric image understanding and inspire the new design of advanced models and algorithms.

Key words: text image understanding; visual information extraction; scene text retrieval; document visual question answering; scene text visual question answering; scene text image captioning

0 引 言

文字无处不在，是人类传递信息的主要载体之一，对人类理解周围世界起着至关重要的作用。随着数字化的加速推进，机器认知和理解图像中的文字已成为数字化进程中的关键一环，如何自动提取和分析以文字为中心的图像，吸引了学术界和工业界越来越多的关注。以往的研究主要集中在自动文字提取技术，如场景文字检测和识别（刘崇宇 等，2021）。然而，由于以文字为中心的图像中包含复杂的语义信息、视觉信息以及布局信息，对其进一步的理解和分析仍然是一个挑战。因此，以文字为中心的图像理解已经成为一个新兴的研究课题，并提出了许多相关任务。例如，场景文字视觉问答技术可以分析社交平台上用户发布的以文字为中心的图像，捕获特定的用户需求；视觉信息提取技术能够从给定图像中提取指定内容，从而提高金融、社交媒体和其他领域的生产率。

本文主要对具有代表性的以文字为中心的图像理解任务进行综述，并按照理解认知程度，将以文字为中心的图像理解任务划分为两类。第1类任务仅要求模型具备抽取信息的能力，如视觉信息提取（visual information extraction）和场景文字检索（scene text retrieval）；第2类任务不仅要求模型具备抽取信息的能力，而且要求模型具备一定的分析和推理能力，如文档视觉回答（document visual question answering）、场景文字视觉问答（scene text visual question answering）和场景文字图像描述（scene text image captioning）。

在过去十几年中，以文字为中心的图像理解技术经历了多个发展阶段，早期方法基于启发式规则，只利用单模态的特征或基于自定义的方式融合多模态特征。随着深度学习的蓬勃发展，大多数以文字为中心的图像理解方法利用深度学习模型充分融合多模态特征，以提高性能。具体来说，传统的视觉信息抽取依赖于固定的模板方法或特定的规则。传统的场景文字检索任务倾向于用字符金字塔直方图向量（pyramid histogram of character vector，PHOC）表示单词，并根据特征距离检索图像。Transformer 和图神经网络等网络框架相继提出并应用于多模态信息融合，有助于更好地理解以文字为中心的图像。此外，各种自监督预训练策略也被用于多模态特征之间的表示学习，从而大幅提高模型性能。

本文旨在对当前学术界以文字为中心的图像理解技术进行详细梳理，对各个理解任务的发展脉络与技术体系进行详细介绍，并对发展趋势进行展望，以期对相关工作起到启发与促进作用。

1 视觉信息抽取

视觉丰富图像（visually rich document）中包含丰富的文本，对图像理解起着至关重要的作用。随着互联网信息技术的飞速发展，许多企业需要利用视觉丰富图像中的关键文本信息，例如医院病例中的病人信息，餐厅小票中的顾客消费情况，火车票中的乘客信息等。企业需要电子化视觉丰富图像中的关键信息，以实现智能化管理。视觉丰富图像可分为数字化和非数字化两类。数字化视觉丰富图像如 word，excel，latex 等格式的文档图像，企业可直接使用或分析文档图像的信息；非数字化视觉丰富图像如 pdf 格式的文档、扫描文件、小票和医院病例等，企业无法直接分析或处理文档图像中的信息。而人工手动处理非数字化的文档图像费时费力，同时又面临数据泄露或标记错误等问题。因此，智能化非数字化视觉丰富图像迫在眉睫。在此情形之下，研究人员提出视觉信息抽取任务的概念。视觉信息抽取是指提取视觉信息丰富文档中，预先定义的关键词，例如发票中的抬头、号码、代码等。然而，视觉丰富图像存在布局结构多变且复杂、文本信息数据量大、非结构化数据量多等特点，使得视觉信息抽取任务具有一定的挑战性。

在早期研究中，研究人员基于特殊类型的视觉丰富图像结构，构建特定的模板或规则，提取文本信息（Cui 等，2021），但是该技术只针对个别文档图像或布局信息简单的文档图像，存在应用场景的限制，在视觉信息抽取任务中的效果十分有限。随着深度学习的快速发展，许多基于深度学习的视觉信息抽取方法已经出现，并且在准确性和性能上都远远超过了传统的基于模板和规则的方案。这些方法大多数将视觉信息抽取任务视做一个标记分类问题，预测检测识别后文档图像中的文字，属于哪类预先定义的关键词。根据表示视觉丰富图像的方法，本文将视觉信息抽取方法分为基于网格、基于图和基于序列表示的 3 类方法。

1.1 视觉信息抽取数据集

为了处理文档图像中的视觉信息抽取任务，提出了多个面向不同类型文档图像的数据集。FUNSD（form understanding in noisy scanned documents）（Jaume 等，2019）是最早提出的公开数据集，数据集中包含 199 张布局差异较大且有噪声的真实扫描表格文档，其中标注了 9 707 个实体和 31 485 个单词，并将实体划分为标题、问题、答案和其他 4 种类别，同时标注了 5 312 个可匹配的实体键值关系。FUNSD 旨在提取和理解嘈杂环境中不同类型文档的关键词，并构建结构化关键信息。SROIE（scanned receipts OCR and information extraction）（Huang 等，2019）提供了第 1 个标准化的 1 000 幅全扫描收据图像和注释的数据集。其中将实体划分为公司、日期、地址和全部金额 4 类。CORD（consolidated receipt dataset for post-OCR parsing）（Park 等，2019）收集了 11 000 幅印度尼西亚商店的收据图像，其中单词分为 8 类实例，包括存储、支付、菜单、小计和总计等。由于之前的数据集大多为英文标注的数据集，文字均为水平布局，且键值匹配比例较低，EPHOIE（examination paper head dataset for OCR and information extraction）（Wang 等，2021b）数据集针对此缺陷，从中国多所学校的真实学生答卷中抽取 1 494 张，将水平和任意形状的手写和打印文本充分标记。由于学生试卷中存在复杂的布局和嘈杂的环境背景，能够有效检验模型的鲁棒性和泛化性。EPHOIE 数据集中有 8 种类别的实体，分别为姓名、学校、考试编号、座位号、班级、学生学号、成绩和分

数。除实体标记外，EPHOIE同样也标注实体键值配对关系，实体键值配对关系比例（85.62%）高于FUNSD（74.69%）和SROIE（54.33%）。Xu等人（2022）考虑到多语言泛化的重要性，提出XFUND（Xu等，2022）多语言标注的数据集，包含7种语言的文档图像，具体为汉语、日语、西班牙语、法语、意大利语、德语和葡萄牙语。同时，XFUND数据集也标注了实体键值配对关系。以上提及数据集的训练集、测试集、实体类别和键值占比的统计数据如表1所示。

表1　视觉信息抽取数据集

Table 1　Visual information extraction dataset

数据集	训练集	测试集	实体类别	键值占比/%
FUNSD（Jaume等，2019）	149	50	4	74.69
SROIE（Huang等，2019）	626	347	4	54.33
CORD（Park等，2019）	800	100	30	-
EPHOIE（Wang等，2021b）	1 183	311	10	85.62
XFUND（Xu等，2021b）	745	350	4	-

注："-"表示对应文献未给出键值占比 。

1.2　视觉信息抽取评价指标

视觉信息抽取包含实体识别（entity labeling）和实体链接（entity linking）两个子任务，可以评价视觉信息提取模型的性能。实体识别任务用预定义的标签对图像中的文本段或语义实体分类。实体链接对文本段或语义实体之间的关系分类，预测文本段或语义实体之间是否存在链接。这两个任务都采用F1分数（F1 score）作为评价指标。F1分数以传统的方式计算，即精确率（precision，P）和召回率（recall，R）之间的调和平均值。具体为

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{1}$$

1.3　基于网格表示的视觉信息抽取方法

基于网格表示的视觉信息抽取方法将图像表示为具有词嵌入的二维网格，这种表示方法既保留了二维特征图中的布局特征，又增添了编码后的文本信息。模型可以直接利用最先进的语义分割和实例分割模型，预测相同类别的实体目标，实现视觉信息抽取任务。

Katti等人（2018）提出Chargrid利用二维网格表示视觉信息丰富图像，预测二维网格中实体的类别，从而将视觉信息抽取任务转化为实例级别的语义分割任务。具体来说，通过word2vec或Glove等方法将文档中的字符编码为字符向量，用字符向量替代原图像中的视觉编码，从而将图像表示为稀疏的二维矩阵。Chargrid采用全卷积网络作为语义分割和实例分割的主体框架，二维字符网格作为输入，预测同类别语义掩码和文本框实例。BERTgrid（Denk和Reisswig，2019）受预训练语言模型BERT（bidirectional encoder representations from Transformers）（Devlin等，2019）的启发，利用BERT编码图像中的词级别的文本内容，优化二维网格中文本信息的表示。VisualWordGrid（Kerroumi等，2020）不仅关注图像中的文本信息，同时关注图像中的视觉信息，例如纹理信息、文字的颜色和字体等。其将二维网格表示与图像特征图结合，生成更强大的多模态二维网格表示。ViBERTgrid（Lin等，2021）结合BERTgrid和VisualWordGrid的优点，提出一种全新的多模态骨干网络，可以在二维特征图中同时编码视觉、文本和布局信息，生成更强大的二维网格表示，联合训练基于全卷积网络的语义分割模块和预训练语言模型BERT。

1.4　基于图表示的视觉信息抽取方法

基于图表示的视觉信息抽取方法将文档图像中的文本词或文本句表示为节点，将自定义或可学习的节点关系表示为边。基于图表示的方法可以保留局部信息，又可以探索图像中的全局信息和非序列化信息。

Qian等人（2019）、Liu等人（2019）、Carbonell等人（2020）提出利用先验知识构建邻接矩阵，使用图卷积神经网络或图注意力神经网络对视觉丰富的文档图像建模，既可以保留文本段中的局部信息，又可以探索各个文本段之间的全局和非顺序关系。通过文字识别（optical character recognition，OCR）系统获得文本段，包含每个文本段在图像中的位置和文本信息。具体来说，Liu等人（2019）将文档表示为全连接的有向图，即每个文本段构成一个节点，经过图卷积操作文本段的信息可以相互传递，通过调整图卷积堆积的层数，可以有效控制文本段的感受野。Qian等人（2019）通过文本段水平相邻或垂直相邻的位置关系，指定4种类型的边和相邻矩阵（从左到右、从右到左、从上到下、从下到上），从而构建预定义的图网络。Liu等人（2019）和Qian等人（2019）使

用图卷积在相邻节点之间迭代传递信息，得到节点嵌入表示。Carbonell等人(2020)计算各个词或文本段左上角之间的距离，选取k个距离最近的词或文本段，构建边和邻接矩阵，其中通过大量实验表明超参数为10时，节点之间信息传递效果最佳。图注意力神经网络通过计算节点之间注意力关系，决定图中节点重要信息的传递，得到丰富节点表示。获得文档的词嵌入后，Liu等人(2019)和Qian等人(2019)使用BiLSTM(bi-directional long short-term memory)(Graves和Schmidhuber，2005)和条件随机场(conditional random field)(Lafferty等，2001)联合全局信息，将视觉信息抽取任务转变为序列标签任务，预测词级别或句级别文本的类别。而Carbonell等人(2020)直接采用多层感知机将节点嵌入信息和相对节点嵌入信息作为输入，预测语义实体的类别和语义实体的链接。

Zhang等人(2020)关注到全连接、K近邻和预定义的网络图会出现网络聚合无效、节点信息冗余和忽视潜在的语义实体连接关系的情况，提出TRIE(end-to-end text reading and information extraction for document understanding)，将图学习模块与图卷积网络相结合，高效地捕获和学习节点之间的潜在关系，优化图神经网络上下文结构，得到更丰富的节点嵌入表示。同时，包含文档布局上下文的边界框也被建模到图嵌入表示中，图卷积模块可以得到非局部和非连续的特征。

Tang等人(2021)针对数字实体识别或模糊文本实体分类不佳的情况，提出一种全新的键值匹配模型MatchVIE(key-value matching model based on a graph neural network for VIE)，可以绕过各种语义的识别，只关注实体之间的强相关性。对于可以键值匹配的文本段，文本段的实体类别由文本段对应的键决定，对于无键值匹配的独立文本段，采用常规的BIOES(begin inside outside end single)(Sang和Veenstra，1999)标记识别模型区分实体类别。

1.5 基于序列表示的视觉信息抽取方法

视觉信息抽取旨在阅读和分析图像文档中的文本信息，与传统自然语言理解任务的学习目标大致相同。传统的自然语言理解任务应用Transformer框架(Vaswani等，2017)和大规模数据预训练技术，在模型精度和速度上取得极大进展和突破。因此，研究者旨在将文档图像序列化，构建基于Transformer框架的预训练语言模型，将视觉信息抽取作为预训练模型的下游任务。基于序列表示的视觉信息抽取方法涉及图像中多模态信息的提取和上下文多模态信息的融合，以及多个自监督任务的设计，仍具有较大挑战性。

1.5.1 多模态信息提取

视觉丰富图像抽取任务相比传统自然语言理解任务更为复杂，研究者除了需要关注文档图像中的文本信息，还需要关注丰富的布局信息和视觉信息，例如文本段之间复杂的逻辑关系和多样的文本字体颜色。通常，研究者通过OCR系统(Wang等，2021b)检测和识别文档图像中文字的内容和位置。之后，多模态信息提取模块将文本信息、视觉信息和布局信息分别映射为鲁棒的高级信息表示。

Li等人(2021a)、Dai等人(2019)、Lee等人(2022)关注于文本实体之间的关系，提取文档图像中文本信息和布局信息。然而，大多数基于序列表示视觉信息丰富图像的方法认为视觉信息与文本信息和布局信息一样重要，故提取视觉信息、文本信息和布局信息。

LayoutLM(pre-training of text and layout for document image understanding)(Xu等，2020)首次提出将视觉信息丰富图像表示为序列，利用基于Transformer的框架，联合学习文本信息和布局信息的预训练语言模型。LayoutLM提取词级别文本框的x轴和y轴绝对位置作为位置信息。之后，StructuralLM(structural pre-training for form understanding)(Li等，2021a)认为文本段之间的单词语义相近，提出不同粒度的绝对二维位置信息编码，例如文本段或词级别的文本信息。BROS(BERT relying on spatiality)(Dai等，2019)、Formnet(form document information extraction)(Lee等，2022)和LiLT(language-independent layout Transformer)(Wang等，2022a)关注到每个文本段之间的逻辑关系或语义关联，对每个词之间的相对位置关系编码。

对于从视觉丰富图像中提取视觉信息，本文总结了3种方法。1)Xu等人(2020)、Li等人(2021b)提出的利用目标检测的分支网络，例如RoIAlign(region of interest align)(He等，2017)，将感兴趣的区域作为RoIAlign的输入提取图像中的视觉信息。2)Xu等人(2021b)、Appalaraju等人(2021)、Gu等人(2022)提出的使用全卷积主干网络提取图像特征，

如 ResNet(deep residual learning for image recognition)(He 等,2016),无需利用区域监督和计算成本较大的目标检测模型,直接将图像中的网格作为输入,提取图像视觉信息。3)Huang 等人(2022)受到 ViT(vision transformer)(Dosovitskiy 等,2021)和 ViLT(vision-and-language Transformer)(Kim 等,2021)的启发,使用最简单的视觉线性嵌入,代替全卷积网络主干,直接对图像编码。

1.5.2 基于Transformer架构的多模态信息融合

基于序列表示的视觉信息抽取方法采用基于Transformer框架实现多模态信息融合和交互,大体可以分为3种融合和交互方法即单流多模态融合架构,联合多模态架构,双流多模态融合架构,如图1所示。

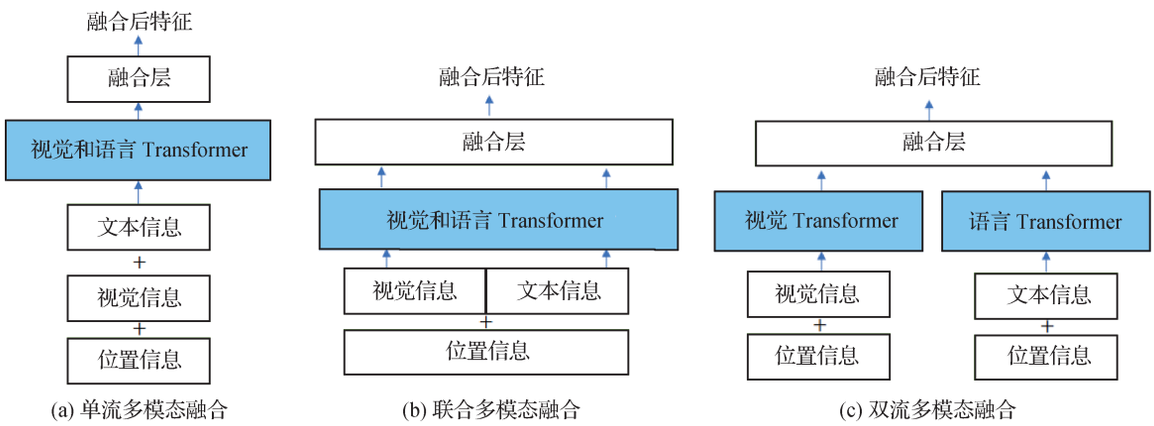

图1 基于Transformer架构的多模态特征融合交互图(Appalaraju 等,2021)

Fig. 1 Diagram of multi-modal features fusion based on Transformer architecture(Appalaraju et al, 2021)
((a)single-stream multi-modal;(b)joint multi-modal;(c)two-stream multi-modal)

1)LayoutLM(Xu 等,2020)、structuralLM(Li 等,2021b)、Docformer(Appalaraju 等,2021)、BROS(Hong 等,2022)、Formnet(Lee 等,2022)采用基于Transformer架构的单流多模态融合(single-stream)。LayoutLM(Xu 等,2020)和 StructuralLM(Li 等,2021b)将提取后的文本信息、视觉信息和布局信息直接相加。因为多模态信息属于多源异构的数据,导致信息无法有效交互。BROS(Hong 等,2022)、Formnet(Lee 等,2022)、Docformer(Appalaraju 等,2021)通过改变Transformer内部注意力机制计算的方式,促进多模态信息融合。具体来说,BROS(Dai 等,2019)提出使用文本框之间的相对空间位置编码文本框的布局特征和作为注意力矩阵的偏置,可以有效提升文本框感知其在图像中的空间位置。Formnet(Lee 等,2022)构建丰富注意力(rich attention),避免绝对位置嵌入和相对嵌入的缺陷,计算相对于布局网格上的 x 轴和 y 轴的标记对之间的顺序和对数距离,直接调整每个文本框的注意力分数。Docformer(Appalaraju 等,2021)提出空间特征和视觉特征作为残差连接,传递入Transformer的每一层中,考虑到空间和视觉依赖可能在不同层中有所不同,在每一层Transformer中,视觉特征和语义特征分别应用自注意力机制,并共享空间特征。

2)LayoutLMv2(Xu 等,2021c)、LayoutXLM(Xu 等,2021b)、StrucText(Li 等,2021c)、XYLayoutLM(Gu 等,2022)和 LayoutLMv3(Huang 等,2022)采用基于Transformer架构的联合多模态融合,可以区别对待视觉特征和文本特征。这些模型将视觉特征和文本特征连接成一个较长的序列,作为Transformer模型的输入,实现跨模态交互。同时,将空间感知的自注意力机制引入到模型体系中,以更好地建模文档布局。LayoutLMv2(Xu 等,2021b)首次在视觉信息抽取任务中应用基于Transformer架构的联合多模态融合,提供文本信息和视觉信息跨模态融合的可能性。LayoutLMv2(Xu 等,2021c)提出空间感知自注意力机制,编码文本框之间的相对关系,为上下文空间建模提供更广阔的视图。LayoutXLM(Xu 等,2021b)采用与 LayoutLMv2(Xu 等,2021c)相似的模型结构,关注视觉信息抽取任务的多语言问题。StrucText(Li 等,2021c)关联不同粒度的多模态特征,例如文本段级别或词级别。同时,StrucText使用Hadamard 点积处理不同级别和模态之间的特征,以

实现高级特征的融合。XYLayoutLM(Gu 等，2022)采用与 LayoutLMv2(Xu 等，2021c)相似的模型结构，提出 Augmented XYCut 模块快速且有效地生成各个文本正确的阅读顺序，捕获和利用丰富的文档信息。此外，dilated conditional position encoding (Chu 等，2021)分支可以生成长度不同的位置嵌入向量，解决文档中文本信息和视觉信息编码后长度不同的问题。XYLayoutLM 在面对复杂的布局图像时，如图像中出现多表格或多列表的情况，有更强的鲁棒性。

3) SelfDoc (self-supervised document representation learning)(Li 等，2021b)和 LiLT(Wang 等，2022a)采用基于 Transformer 架构的双流多模态融合(two stream multi-modal)，每个模态都是一个单独的分支，允许每个模态分支使用任意模型，文本模块和视觉模块可以在各分支中进行信息交互。SelfDoc(Li 等，2021b)对于每个模态，使用一个单一模态编码器进行上下文学习，之后使用跨模态的编码器在两个单一编码器之后学习基于上下文的信息。LiLT(Wang 等，2022a)将文本和布局信息分别传入对应的基于 Transformer 架构的编码器获得丰富特征。Wang 等人(2022a)提出双向注意力互补机制(bi-directional attention complementation mechanism)可以实现文本与版面线索的跨模态交互。

1.5.3 预训练任务

Xu 等人(2020)、Appalaraju 等人(2021)、Li 等人(2021b)、Xu 等人(2021c)、Lee 等人(2022)以及 Wang 等人(2022a)均应用大规模数据预训练技术。大多数基于序列表示视觉信息丰富图像的方法大致设定 3 种预训练任务。受 BERT(Devlin 等，2019)中随机掩码语言(masked language modeling，MLM)预训练任务的启发，Xu 等人(2020)提出掩码视觉语言任务(masked visual language modeling，MVLM)，可以随机掩码文本信息或与之对应的视觉信息，但保留其二维位置信息，MVLM 利用跨模态信息改进了语言的模型学习，给定的布局嵌入还可以帮助模型更好地捕获句子间和句子内的关系，训练模型预测给定上下文的掩码标记。具体的掩码原则为掩码 15% 的标记(token)，其中 80% 被特殊标记 [MASK] 替换，10% 被从整个词汇表中抽样的随机标记替换，10% 保持不变。Wang 等人(2022a)提出关键点位置(key point location，KPL)预训练任务，KPL 随机掩码视觉信息，保持布局信息不变，掩码规则与 MVLM 一致。Xu 等人(2021c)提出匹配预训练任务，分别为文本视觉对齐(text-image alignment，TIA)和文本图像匹配(text-image matching，TIM)。TIA 可以帮助模型学习图像和文本框边界之间的空间位置对应关系。TIM 采用粗粒度跨模态对齐任务，帮助模型学习图像和文本内容之间的对应关系。

1.6 端到端的视觉信息提取

文本检测识别(text reading)(Wei 等，2022)和视觉信息抽取任务分别是文档情景下的两个子任务。Zhang 等人(2020)和 Wang 等人(2021b)认为文本检测识别和视觉信息抽取是两个强相关的任务，端到端的模型可以同时提升两个子任务的性能，文本检测模块可以促进视觉信息抽取任务的多模态信息融合，视觉信息抽取任务中对于关键信息的监督会反向传递到文本检测识别模块(Qiao 等，2021)。

Zhang 等人(2020)和 Wang 等人(2021b)提出视觉信息抽取的端到端模型，由文本检测、文本识别和信息抽取 3 个特定的分支组成。给定文档图像，文本检测分支通过 RoIAlign(He 等，2017)得到视觉特征，文本识别分支获得语义信息。TRIE(Zhang 等，2020)采用多头注意力机制将二维位置信息和语义特征融合，得到基于上下文的语义特征。在信息抽取模块，TRIE 提出可适应的多模态上下文融合(adaptive multi-modality context fusion)，计算视觉特征和基于上下文的语义特征的加权和，得到具有上下文的信息。VIES(robust visual information extraction system)(Wang 等，2021b)，对文本段语义信息进行多次一维卷积，得到更丰富的语义特征。自适应特征融合模块引入多层多头注意力机制，结合线性变换，分别丰富不同粒度的多模态向量。对多模态融合后的特征，TRIE(Zhang 等，2020)和 VIES(Wang 等，2021b)都将特征输入到标准的双向条件随机场(bidirectional LSTM conditional random field，BiLstm CRF)预测实体的类别。

1.7 视觉信息抽取方法性能对比分析

本节对比了现有视觉信息抽取方法在 FUNSD、CORD 和 SROIE 共 3 种数据集上，实体识别和实体链接任务的性能结果，如表 2 所示。由表 2 看出，基于网格表示和基于图表示的方法与基于序列表示的方法在性能上相比稍有逊色。本文认为，基于网格表示和基于图表示的方法有多种缺陷。首先，其无法有效结合 BERT(Devlin 等，2019)等自然语言预训练

表2　视觉信息提取方法F1分数对比分析

Table 2　F1 score comparison of visual information extraction task

/%

任务	方法	FUNSD	CORD	SROIE
实体识别	LayoutLM(Xu等,2020)	79.27	–	95.24
	TRIE(Zhang等,2020)	–	–	96.18
	PICK(Yu等,2020)	–	–	96.10
	SelfDoc(Li等,2021b)	83.36	–	–
	ViBERT(Lin等,2021)	–	–	96.40
	StrucText(Li等,2021c)	83.09	–	96.88
	StructuralLM(Li等,2021a)	85.14	96.08	–
	UDoc(Gu等,2022)	87.93	98.94*	–
	LayoutLMv2(Xu等,2021)	82.76	94.95	96.25
	DocFormer(Appalaraju等,2021)	83.34	96.33	–
	BROS(Hong等,2022)	84.52	97.28	96.62
	LiLT(Wang等,2022a)	88.41	96.07	–
	XYLayoutLM(Gu等,2022)	83.35	–	–
	LayoutLMv3(Huang等,2022)	90.29	96.56	–
	FormNet(Lee等,2022)	84.69	97.28	–
实体链接	StrucText(Li,2021c)	44.10	–	–
	LiLT(Wang,2022a)	62.76	–	–

注:*表示使用额外的数据集,"–"表示对应文献未给出实验结果。

模型,利用基于上下文的文本信息表示;其次,难以设计契合视觉信息抽取任务的预训练方法,从而大规模无标注的预训练数据得不到充分利用;最后,无法有效地融合多模态信息。

最早的基于序列表示的视觉信息抽取方法LayoutLM(Xu等,2020)认为,学术界提出的视觉信息抽取模型大多数是基于序列表示的。由表2可以看出,基于序列表示的视觉信息抽取方法LayoutLMv3(Huang等,2022)、BROS(Hong等,2022)、StrucText(Li等,2021c)分别在FUNSD、CORD和SROIE数据集上取得了最优结果,这是因为基于序列表示的视觉信息抽取方法可以应用Transformer框架(Vaswani等,2017)和大规模数据预训练技术(Zhang等,2022),从而在模型精度和速度上取得极大的进展和突破。

此外,基于序列表示的视觉信息抽取方法可以借助Transformer架构,增添轻量多模态编码器和设计不同的Transformer注意力机制的方式,有效地将图像中包含上下文内容的多模态信息融合。例如,LayoutLMv3提出文本—图像块对齐(word-patch alignment)的预训练任务,使用轻量的线性嵌入层编码图像块,预测文本对应图像块是否被遮挡,学习跨模态对齐。而BROS采用区域掩码策略(area-masking strategy)预训练任务,设计具有空间感知的注意力机制,编码二维空间文本的相对和绝对位置信息。

目前,尽管视觉信息抽取技术取得了很大的进步,但如何有效地融合多模态信息、如何提升在布局和视觉风格迥异文档图像中的泛化能力、如何解决视觉信息抽取任务中的few shot问题、如何避免文本检测和识别错误对视觉信息抽取任务精度上的影响等一直是这一领域亟待解决的问题。

2　场景文字检索

场景文字检索的目标是从图像库中定位和检索出给定查询文字指示的文字实例。早期的文字检索方法用于检索裁剪后的文档图像。例如,Almazán等人(2014)提出使用PHOC(pyramidal histogram of characters)向量表示单词,通过计算查询文本与单词图像之间的距离来实现检索。对于场景文字检索,

当前的方法大致包括两类，即按字符串检索方法和按描述检索方法。

2.1 场景文字检索数据集和评价指标

本文总结了针对按字符串检索方法提出的3种数据集，分别是街景文本数据集（street view text dataset，SVTQA）（Kai 等，2011）、场景文本检索数据集（scene text retrieval，STR）（Anand 等，2013）和文本检索数据集（cocotext retrieval dataset，COCO）（Veit 等，2016）。SVT 包含 349 幅街景图像，在测试集中有 427 个标记的查询文本。STR 包含 10 000 幅图像，50 个查询文字，由于多变的字体和图像角度，具有一定的挑战性。CTR 从 COCO-Text（Veit 等，2016）挑选 500 个查询文本和 7 196 幅图像组成数据集。通常按字符串检索的方法用平均精度（mean average precision，mAP）进行度量。

2.2 按字符串检索方法

为了处理自然场景图像，Mishra 等人（2013）首次提出场景文字检索任务，该任务需要模型返回所有包含查询文字的图像。该工作将场景文字检索分为检测和检索两个独立的子任务，通过图像文字实例与查询字符串的相似性度量对图像进行排序。Ghosh 等人（2015）采用选择搜索算法产生文字区域，再用支持向量机作为分类器来预测文字的 PHOC 向量。然而，这两种方法都依赖手工提取的特征，并且将文字检测和相似性度量分开处理，在精度和效率上受到限制。对此，Gómez 等人（2018）提出 Single Shot CNN（convolutional neural network）架构，能够同时预测文字实例的位置和其对应的 PHOC 向量，从而将检索任务转换为查询文字对所有图像的 CNN 输出向量的最近邻搜索问题。Mafla 等人（2021）进一步提高了检索速度，并提升了在训练时未见过的样本上的检索性能。Wang 等人（2021b）提出了一个更简洁的端到端框架，将文字检测和相似性度量预测联合优化。由此，模型只需根据相似度对检测到的文字实例进行排序即可实现场景文字检索。

2.3 按描述检索方法

随着跨模态检索、视觉定位（visual grounding）等多模态任务受到广泛关注，Rong 等人（2020）开始探索给定自然语言描述下的场景文字检索问题。假设在水果店的场景下，给定一句自然语言描述，例如“larger text above a pile of oranges”，模型需要从图像中找到该语句对应的文字实例，并输出检索结果“$5. 95/kg”。为了处理这个任务，首先提出了一个递归密集文字定位网络来获取文字实例，并通过记忆机制避免重复的文字检测。其次，将一个上下文推理的文字检索模型用于对文字实例及其上下文信息进行联合编码，由上下文兼容性评分函数对定位到的文字框进行排序。在后续工作中，Rong 等人（2022）进一步增加一个识别模块，从而实现端到端的文字定位、检索和识别，他们将类似的思想也用到了基于自然语言描述的场景文字分割任务上（Rong 等，2022）。

2.4 场景文字检索方法性能对比分析

现有方法在各个数据集不同任务上的性能比较结果如表 3 所示。Wang 等人（2021a）提出的端到端模型框架联合场景文本检测和多模态相似性学习共同优化，通过对检测后的文本实例相似度排序，检索出最相关的图像，并在 SVT、STR、CTR 等 3 个数据集中都达到了最先进的性能。现有方法对文本检测的速度和精度有较强的依赖性，如何提高现有方法的效率和避免文本检测精度的依赖是一个值得研究的课题。

表 3 场景文字图像检索方法性能对比分析

Table 3 Performance comparison of scene text retrieval task

方法	全类平均准确率/%			FPS/（帧/s）
	SVT	STR	CTR	
Mishra 等人（2013）	56.24	42.70	–	0.10
Gómez 等人（2018）	83.74	69.83	41.05	**43.50**
Wang 等人（2021a）	**91.69**	**80.16**	**69.87**	3.80

注：加粗字体表示各列最优结果，“–”表示对应文献未给出实验结果。

3 文档视觉回答

文档视觉回答（document visual question answering，Doc VQA）是一个高级语义理解任务，不仅要求模型具备抽取信息的能力，而且要求模型具备一定的分析和推理能力。文档视觉回答可以动态驱动文档分析识别算法（document analysis and recognition，DAR）有条件地解释文档图像，模型可以智能阅读，并响应特定信息的请求和回答视觉问题。具体来说，文档视觉回答可以利用较为成熟的文档分析识

别算法，自动从文档图像中提取信息，将文档图像转换为数字化的文本内容，例如文档字符识别（Qiao等，2020）、布局分析和表格提取等，之后再充分利用提取后的文本、布局和样式等信息，抽取或推理开放式答案。本文按照理解认知程度和不同文档的类型，将文档视觉回答划分为3类，分别是单文档视觉回答、文档集合视觉回答（document collection VQA，DocCVQA）（Tito等，2021）和信息图形视觉回答（infographics VQA）（Mathew等，2022）。

3.1 文档视觉回答数据集

Mathew等人（2021）提出单文档视觉回答任务（single document visual question answering，DocVQA），根据给定的单幅文档图像和提出的问题，可以直接从文档图像的内容中抽取出合理且准确的答案。DocVQA是第一个关于文档图像的视觉回答任务数据集，要求模型具备最基本的信息抽取能力，其主要由商业业务文档组成，涉及图表、表格、复杂布局的文本句和文本段落，共计12 767幅文档图像和与之对应的50 000个问题，问题的平均长度为8.12个单词，独立的问题占70.72%。

Tito等人（2021）提出文档集合视觉问答（document collection question answering，DocCVQA），旨在理解包含丰富上下文的多幅文档图像集合，对相同模板且内容不同的文档图像集合提出问题。模型不仅需要回答给定问题的答案，还需要检索出包含推理答案信息的文档图像编号。具体来说，DocCVQA数据集包括14 362幅文档图像，由于文档集合中的图像模板较为类似，DocCVQA数据集只设定20个问题，答案来源于多个不同的文档。

Mathew等人（2022）提出信息图形视觉回答数据集（infographics VQA），该任务减少了由大量文本构成文档图像的数量，增加了需要考虑图形、布局等视觉信息才能获得答案的问题数量。此外，信息图形视觉回答不仅要求模型具备抽取信息的能力，还要求模型具备通过图形信息、手写文本或复杂的布局信息中分析、推断出正确答案的能力。信息图形视觉回答数据集包含5 485幅信息图形回答图像和30 035对问答对。为了分析模型的能力，信息图形视觉回答数据集还收集了除问题和答案外的额外信息作为分析依据，包括答案的线索是否来自于文档图像中的单个部分或多个部分，是否来自文本、表格、图形、地图或其他视觉信息和布局信息，答案是否可以直接抽取或需要模型推理。对于需要推理才能获得答案的问题，该数据集统计了需要计数、分类或运算操作获得答案的数量。

3.2 文档视觉回答评价指标

文档视觉回答借鉴Biten等人（2021）提出的平均归一化编辑相似性（average normalized levenshtein similarity，ANLS）对模型进行评估。给定一个问题Q，答案列表为G（长度为M）和模型预测的答案列表P（长度为N），先执行匈牙利匹配算法得到真值—预测答案对列表U。然后计算所有真值—预测值答案对的NLS（normalized levenshtein similarity）分数之和，再除以两个列表长度的最大值即可得到ANLS指标值。具体为

$$\begin{aligned} U &= \psi(NLS(G,P)) \\ ANLS &= \frac{1}{\max(M,N)}\sum_{z=1}^{k} NLS(u_z) \end{aligned} \tag{2}$$

式中，u_z为真值和预测值答案对。

文档集合视觉问答参考检索任务（Liu，2009）的评价指标为平均精度（retrieval MAP），评价模型给出答案线索的能力。

3.3 单文档视觉回答

Mathew等人（2021）提出单文档视觉回答任务（single document visual question answering，DocVQA），根据给定的单幅文档图像和提出的问题，给出合理且准确的正确答案。同时，Mathew等人（2021）提出两种基线模型，分别是传统的问答模型BERT（Devlin等，2019）和场景文字视觉回答模型M4C（multimodal multi-copy mesh）（Hu等，2020）。具体来说，通过文档分析和识别算法抽取图像中的文本内容，根据从左到右和从上到下的排序方式，将文本内容连接为一段序列，作为两种基线模型的输入。基于BERT（Devlin等，2019）的基线模型使用SquAD（stanford question answering dataset）数据集（Rajpurkar等，2016）进行预训练，随后在单文档视觉回答文档微调。基于M4C（Hu等，2020）的基线方法将问题、图像中的文本合并到一个空间中，在该空间中应用Transformer堆叠，允许每个实体参与模态间和模态内的特征融合，并根据动态指针迭代图像中识别的文字或固定大小的字典，由解码器生成答案。但由于该基线方法仅针对场景文字中的问答问题，并不适用于单文档视觉回答任务。

单文档视觉回答答案的证据文本段或文本语料

库，由于使用大规模未标注的数据预训练和微调语言模型获得极大成功，研究者在解决单文档视觉回答任务时也采用大规模的文本语料库预训练模型。Li 等人（2021b）、Xu 等人（2021c）和 Huang 等人（2022）提出基于 Transformer 编码器的预训练模型，设计特定的预训练任务，通过预测文档图像的部分范围作为答案微调模型。Li 等人（2021a）提出 StructuralLM 预训练模型，关注到单元级别的语义信息和布局信息对文档理解的重要性，并通过两种预训练任务，即掩码视觉语言任务（masked visual-language modeling）和单元级别位置预测任务（cell position classification），充分融合单元级别的语义信息和布局信息。Xu 等人（2021c）认为图像中文本的视觉信息与语义信息和位置信息同等重要，提出 LayoutLMv2 模型并引入文本和图像对齐（text-image alignment）预训练任务，预测行级别的文本是否在图像中被遮挡，帮助模型学习行级别文本的视觉信息和位置信息之间的空间对应关系。 LayoutLMv3（Huang 等，2022）将图像切割数块，直接采用原始的图像块作为视觉信息，不依赖卷积神经网络或 Faster R-CNN（region convolutional neural network）（Ren 等，2017）主干来提取视觉特征，节省模型参数和降低计算复杂度。

TILT（Powalski 等，2021）采用端到端的 Transformer 编码器—解码器架构模型，可以根据文本段之间的相对位置信息，生成注意力矩阵偏差（attention bias），更好地表示文本段的布局信息，解码器将融合后的多模态信息作为输入，直接生成给定问题的答案。

现有方法在 DocVQA 数据集上的性能比较结果如表 4 所示。由表 4 可以看出，抽取式模型的性能均低于生成式模型，生成式模型 TILT（Powalski 等，2021）达到最优性能。抽取式的模型一般采用 Transformer 编码器架构，将问题和文档图像的文本内容作为输入，根据文本图像中文本内容对应的输出部分，直接预测答案的起始和终止位置，而生成式模型采用端到端的 Transformer 编码器—解码器架构模型，直接生成问题的答案，与抽取式模型相比更具有鲁棒性和泛化能力。

Doc VQA 数据集问题的答案全部可以直接从文档图像中抽取获得，但业界对于文档视觉回答任务的要求远高于此，需要模型具有生成答案的能力和数学推理的能力，因此学术界相继提出 DocCVQA 数据集和 Infographics VQA 数据集，期待研究者可以解决当前文档视觉回答方法的现有缺陷。

表 4　DocVQA 数据集上现有方法性能
平均归一化编辑相似性性能对比

Table 4　Average normalized Levenshtein similarity comparison of existing methods on DocVQA dataset

/%

方法	TestANLS
Xu 等人（2021c）	78.08
Li 等人（2021a）	86.10
TILT（Powalski 等，2021）	**87.05**
Huang 等人（2022）	78.76

注：加粗字体表示最优结果。

3.4　文档集合视觉回答

Tito 等人（2021）提出 DocCVQA 数据集，提供了两种基线方法，即 baseline BERT（Devlin 等，2019）和 baseline database 基线数据库。它们都是根据包含给定问题的置信度对文档进行排序，然后从排序最高的文档中获取答案。不同点在于，baseline BERT 首先根据文本定位问题中的特定单词检索文档，然后使用预训练文档视觉问答模型，从检索到的文档中提取答案。baseline database 则是提取集合中所有文档的信息，映射成结构化的键值对数据，然后检索与问题相关的文档和答案。

3.5　信息图形视觉回答

Mathew 等人（2021）提出 Infographics VQA 数据集，提供 3 种基线方法，即 LayoutLM（Xu 等，2020）、M4C（Hu 等，2020）和 Human。为了适用信息图形视觉回答任务，LayoutLM 更改输入设置，在输出端加入文本段预测。由于 M4C 模型是针对场景文字视觉回答模型设计的，性能在信息图形视觉回答任务中表现不佳。Human 方法为人类手动标记的性能，可以看做是信息图形视觉回答性能的上限。此外，单文档视觉回答任务的方法 TILT（Powalski 等，2021）可以直接用于信息图形视觉回答，因为其具有一定的生成答案的能力，而不是仅抽取文档内容作为答案，在信息图形视觉问答任务上有不错的表现。然而，现有信息图形视觉回答方法仍然存在不足，例如在面对计数、简单运算和需要从图像中的多个线索推导时，无法给出精确的答案。

4 场景文字视觉回答

传统的视觉问答问题虽然能够有效根据图像内容回答给定的问题，但是却忽略了图像中的一个重要组成部分——场景文字。Singh等人(2018)首次提出了场景文字视觉问答(TextVQA)任务，它需要视觉问答(VQA)模型阅读图像中的文字，并且通过对场景文字和其他视觉内容进行推理来回答给定的问题。以场景文字为中心的视觉问答具有十分重要的实践意义，如帮助视障用户与周围环境交互，检查商店是否营业，在城市中查找一个地方等。

本节首先介绍场景文字视觉问答相关的数据集和评价指标，然后对已有的经典方法进行介绍。现有的大多数场景文字视觉问答方法都包含3个主要模块，即特征提取模块、多模态融合模块和答案预测模块。本文依次按照这3个模块对已有方法进行详细阐述。此外，由于多模态预训练和端到端模型优越的性能表现，最近一些工作还探索了场景文字问答的预训练模型以及端到端的场景文字问答模型。

4.1 场景文字视觉回答数据集

为了处理场景文字视觉问答任务，提出了多个面向不同场景的场景文字视觉问答数据集。主要包括TextVQA(text visual question answering)(Singh等，2019a)、ST-VQA(scene text visual question answering)(Biten等，2019)、OCR-VQA(optical character recognition visual question answering)(Mishra等，2019)、TextKVQA(knowledge-enabled visual question answering)(Singh等，2019b)和STE-VQA(scene text evidence visual question answering)(Wang等，2020b)。

TextVQA(Singh等，2019b)是最早提出的面向场景文字视觉问答的数据集。TextVQA数据集包含了来自Open Images v3(Krasin等，2016)数据集的28 408幅图像以及与之对应的45 336个问题。每个问题都需要阅读并推理图像中的文字信息才能够准确回答。TextVQA的平均问题长度高于传统的视觉问答数据集，给场景文字视觉问答任务提出了更大挑战。

与TextVQA数据集不同，ST-VQA(Biten等，2019)的数据来源于6个不同数据集以便减少数据偏见，增加问题的多样性，而且该数据集更关注于可以直接使用部分图像中的场景文字作为答案的问题。

OCR-VQA(Mishra等，2019)数据集包括207 572幅书籍封面的图像以及与之对应的1 002 146个问答对，问题主要涉及到书本标题、作者姓名和书本类型等。

Text-KVQA(Singh等，2019b)是第一个将场景文字视觉问答与外部知识进行联合的数据集，主要涉及商业场景、电影海报以及书籍封面等与现实场景密切相关的内容，包含25.7万幅图像、130万对问答对和相关的外部知识库。

STE-VQA(Wang等，2020b)数据集是一个同时包含中英文的场景文字视觉问答数据集，特别标注了每个问题需要关注的图像区域位置作为推理证据，从而减弱训练数据中的虚假相关性带来的影响。

4.2 场景文字视觉回答评价指标

与传统的视觉问答一样，准确率(accuracy，Acc)也可用于评估预测的答案与真实值之间的匹配程度。由于每个问题可以对应多个答案标签，所以准确率需要通过这些答案的软投票来计算。具体为

$$Acc = \max\left(\frac{n}{3}, 1\right) \tag{3}$$

式中，n是与预测答案相匹配的答案标签的数量。如果有3个或更多的匹配项，则认为预测的答案是正确的。

然而，准确率度量不能区分OCR错误和推理错误。例如，对于问题“瓶子里有什么牌子的苏打水?”，模型直接复制文字识别结果，输出“EPSI”而不是“PEPSI”。在这种情况下，即使该模型能够准确地进行推理，但该模型也不会得到奖励。因此，平均归一化编辑相似性(average normalized levenshtein similarity，ANLS)(Biten等，2019)度量是从预测答案和真实答案之间的编辑距离来衡量的，其定义为

$$ANLS = \frac{1}{N}\sum_{i=0}^{N}\left(\max_{j} S(a_{ij}, a_{qi})\right) \tag{4}$$

$$S(a_{ij}, a_{qi}) = \begin{cases} 1 - NL(a_{ij}, a_{qi}) & NL(a_{ij}, a_{qi}) < \tau \\ 0 & NL(a_{ij}, a_{qi}) \geqslant \tau \end{cases}$$

式中，N是问题的总数，M是每个问题的真实答案的总数。$a_{ij}\left(0 \leqslant i \leqslant N, 0 \leqslant j \leqslant M\right)$是真实答案，$a_{qi}$是第$i$个问题$q_i$的预测答案，$NL$是归一化的编辑相似性。阈值$\tau = 0.5$，用于确定模型是否存在推理错误或识别错误。

4.3 特征提取

特征提取模块会将问题文字特征、图像中的物体特征和场景文字特征提取出来并映射到一个公共特征空间。具体来说，多数模型会利用 LSTM（bi-directional long short-term memory）（Gómez 等，2021）、BERT（Devlin 等，2019）或 GloVe（global vectors for word representation）（Pennington 等，2014）等预训练语言模型提取问题的特征。对于图像中的视觉物体特征的提取，主要分为基于网格的特征表示和基于区域的特征表示。其中，基于网格的特征表示是用预训练的卷积神经网络提取视觉物体的特征；基于区域的特征表示是运用目标检测算法提取视觉物体的特征。此外，视觉 Transformer 也因其优越的性能和无需额外的物体检测器而得到广泛应用。

对于场景文字特征的提取，多数工作选择使用光学字符识别（OCR）方法检测和识别场景文字（Qin 等，2021），然后用一些自然语言处理方法对识别结果进行丰富的特征表示。例如，LoRRA（look，read，reason & answer）（Singh 等，2019a）将每个文字字段映射为一个 300 维的 FastText（Joulin 等，2017）向量。在此基础上，M4C（Hu 等，2020）又额外加入了一个 604 维的 PHOC（pyramidal histogram of characters）（Almazán 等，2014）向量作为语义向量的补充，并利用预训练的 Faster R-CNN（Ren 等，2017）提取场景文字的外观视觉特征和边框位置特征。Sharma 和 Jalal（2022）将预训练的高斯混合模型（Gaussian mixture model，GMM）应用于 PCA（principal component analysis）处理后的 PHOC 特征上，构造 Fisher 向量描述符作为场景文字语义信息的补充。由于 FastText 和 PHOC 等语义特征向量是直接对场景文字识别结果进行表征的，因此会受到文字识别结果的影响。为了解决这个问题，SMA（structured multimodal attention）（Gao 等，2020a）和 SSBaseline（simple strong baseline）（Zhu 等，2021）模型引入了一个 RecogCNN（recognition convolutional neural network）语义特征，它是直接从文字识别网络的中间层进行提取的，所以受到识别结果影响较小。进一步地，BOV（beyond optical character recognition visual question answering）（Zeng 等，2021）设计了一个文字相关的视觉语义映射网络，直接根据文字区域的视觉特征获取文字的语义向量，从而缓解由文字识别错误造成的错误累积，增强模型的鲁棒性。

4.4 多模态推理融合

针对上述特征提取模块获取的多模态特征，多模态融合推理模块需要进行模态内和模态间信息的融合与交互，并进行一定的推理，这对于场景文字视觉问答任务非常关键。多模态融合推理模块根据所采用的网络框架，主要可分为基于注意力机制的方法、基于 Transformer 的方法和基于图神经网络的方法。

4.4.1 基于注意力机制的方法

受传统视觉问答模型 Pythia（Singh 等，2018）的启发，Singh 等人（2019a）提出 TextVQA 数据集的一个基准模型 LoRRA，它完全依靠注意力机制来实现模态间交互。LoRRA 模型包含 3 个组成部分，即 VQA 组件、OCR 组件和回答模块。其中，VQA 组件和 OCR 组件根据问题文字编码分别对图像视觉特征和图像场景文字特征进行注意力加权，两个模块的组合输出用于预测答案。ST-VQA（Biten 等，2019）采用堆叠注意力网络 SAN（stacked attention network）（Yang 等，2016）作为基线，使用问题的语义表示作为查询来探测图像中与答案相关的区域。Gómez 等人（2021）和 Wu 等人（2022）提出先将多个模态的特征组合成融合特征，然后用问题文字作为注意力查询，计算在融合特征上的注意力权重，从而实现对图像和场景文字两个模态进行联合推理。该注意力权重可以理解为图像的某个空间位置包含回答给定问题的答案文本的概率，增强了模型的可解释性。之前的许多方法都是简单地将从图像中提取的特征信息直接添加到模型中，而 RUArt（reading，understanding and answering the related text）（Jin 等，2020）使用机器阅读理解方法 SDNet（Zhu 等，2019）来获取场景文字和视觉物体的上下文关系，提升了模型对特征信息的理解能力。此外，为了更好地挖掘场景文字和视觉物体之间的关联关系，RUArt 还提出两种注意力机制分别建模文字和物体之间的语义关系和位置关系，提升了模型的推理能力。

4.4.2 基于 Transformer 的方法

Transformer（Vaswani 等，2017）运用自注意力机制来建模全局的上下文信息，能够更有效地处理长距离依赖关系，由于其优越的性能在计算机视觉（computer vision，CV）、自然语言处理（natural language processing，NLP）和多模态领域都取得了巨大成功。在场景文字视觉问答领域，Transformer 结构

首先在M4C(Hu等,2020)中得到应用。M4C将提取的问题特征、视觉物体特征和场景文字特征编码映射到一个公共空间,使用堆叠的多头注意力机制实现各个模态内和模态之间的自由交互。借助于Transformer强大的特征表示能力,M4C较之前方法取得了很大的性能提升,并引出了一系列以M4C作为基础模型的改进工作。

考虑到M4C的同构处理方式没有对不同模态的输入信息加以区分,可能学习到冗余知识,SA-M4C(spatial aware multimodal multi-copy mesh)(Kant等,2020)构建了一个空间图结构表征视觉实体(包括场景文字和视觉物体)之间的空间关系,使用此关系辅助多模态Transformer的自注意力层学习。CRN(cascade reasoning network)(Liu等,2020)在使用Transformer之前,用一个渐进式注意力模块逐步地融合多模态信息,并设计了一个多模态推理图网络动态地建模不同实体间的关系。Zhu等人(2021)认为简单的注意力机制能达到与复杂Transformer结构可比甚至更佳的性能,设计了一种更简单有效的基线方法,先采用普通注意力机制聚合多模态信息,再将其送入基于Transformer的解码器,显著减小了计算复杂度。Zhang和Yang(2021)认为以往的场景文字视觉问答方法倾向于使用复杂的图结构或手工特征建立视觉实体间的关系,不能有效捕获相对位置信息,且丢失了原始图像的全局特征。为了解决这些问题,他们将连续的相对位置信息显式地并入Transformer的计算过程中,而且设计entity-aligned mesh将场景文字和物体的特征直接与图像特征整合。LOGOS(localize, group, and select)(Lu等,2021)同样基于Transformer网络建模,提出一个定位、聚类和选择的3步串行策略。它与M4C的区别是训练了一个visual grounding模型,能够根据问题提示定位到应该关注的图像区域,并使用聚类的方法将有关联的场景文字内容进行组合,从而提升预测答案的完整性。特别地,LaTr(layout aware Transformer)(Biten等,2021)的Transformer架构基于强大的文本到文本传输变压器(T5)模型,并在多个数据集上实现了最先进的性能。

4.4.3 基于图神经网络的方法

由于图神经网络在关系建模上的优势,一些方法提出利用图神经网络来处理场景文字视觉问答任务中复杂的多模态交互问题。MM-GNN(multi-modal graph neural network)(Gao等,2020b)提出一个基于多模态图神经网络的模型,可以捕获图像中各种模态的关系来推理出未见过的场景文字的含义。具体来说,MM-GNN首先用3个图结构分别表示视觉物体信息、文字语言信息以及数字型文字的数值信息,然后引入3个图网络聚合器,引导不同模态的知识从一个图传递到另一个图中。SMA(structured multimodal attention)(Gao等,2020b)提出一个结构化的多模态注意力模型,主要分为3个步骤。1)采用问题自注意力模块将问题文字映射为6种特征,用于指导后续图网络的节点和边;2)基于物体和文字节点构建一个关系图网络,并使用问题引导的图注意力模块更新此图网络;3)采用全局—局部注意模块迭代产生变长答案。TIG(text-instance graph)(Li等,2022)也构建了类似的图网络,主要建模物体和文字节点间的重叠关系,并动态更新图网络来扩充图节点的感知空间,从而能够解析复杂的语义关系。

4.5 答案预测模块

为了回答与场景文字有关的问题,目前场景文字视觉问答模型的答案空间通常包含两部分,即由训练集中高频答案词组成的固定词典和由场景文字组成的动态词典。早期的LoRRA(Singh等,2019a)方法将答案预测建模为一个单步分类问题,只能输出单个词作为答案,性能十分有限。为了输出不定长答案,M4C(Hu等,2020)提出了一个动态指针网络,以自回归迭代解码的方式生成答案。在此基础上,LaAP-Net(localization aware answer prediction network)(Han等,2020)认为现有场景文字视觉问答模型预测答案时缺乏证据支持,从而提出一个位置感知的答案预测网络。该网络不仅能生成问题的答案,还预测一个边界框作为答案预测的证据。Wu等人(2022)在答案预测模块额外添加了一个注意力图损失,目的是拉近在答案空间上预测解码向量与真实解码向量的分布距离。然而,上述方法都是基于预先获取的文字识别结果进行推理和答案预测,如果文字识别结果出现错误,“复制”的答案词也有可能出现错误。因此,BOV(beyond optical character recognition visual question answering)(Zeng等,2021)在定位到答案区域后,设计了一个基于上下文的答案修正模块,利用问题、物体等其他场景上下文信息修正答案词的识别错误。此LaTr(Biten等,

2021)利用了强大的T5模型的生成解码,可以生成固定词汇表之外的单词,还增强了对OCR错误的鲁棒性。

4.6 基于预训练表示学习的场景文字视觉回答方法

随着预训练模型范式的蓬勃发展,越来越多的研究将其运用到多模态领域以取得优越效果。TAP(text-aware pre-training)(Yang等,2021)首次将多模态预训练方法用于场景文字视觉问答任务,在M4C框架上实现了性能的大幅提升。具体地,TAP沿用了两个经典的视觉—语言预训练任务,即掩码语言建模和图像文字匹配,并设计了一个新的相对位置预测任务来学习图像中文字与物体的空间关系。

4.7 基于端到端阅读推理的场景文字视觉问答方法

大多数场景文字视觉问答框架将OCR模块视为“黑盒”,无法与下游任务共同优化,Singh等人(2021)提出一个新的场景文字数据集TextOCR,其提供了对TextVQA数据集的OCR标注,并且联合了TextOCR训练的Mask TextSpotter v3(Liao等,2020)识别网络和M4C(Hu等,2020)网络,创建了第一个端到端的场景文字视觉问答框架PixelM4C。由此,场景文字的检测、识别结果和对应特征可以在训练时实时提取,有效提升模型效率。

4.8 场景文字视觉问答方法性能对比分析

表5展示了目前主要的场景文字视觉问答方法在TextVQA(Singh等,2019a)和ST-VQA(Biten等,2019)两个数据集上的性能对比结果。由表5中结果可知,基于普通注意力机制的方法(LoRRA等)因为需要人工设计不同模态之间的交互方式和顺序,性能有限。相对而言,以M4C(Hu等,2020)为代表的基于Transformer的方法能够同时实现各个模态内

表5 场景文字视觉回答方法性能对比分析

Table 5 Performance comparison of scene text VQA

方法	OCR模块	额外数据	TextVQA数据集		ST-VQA数据集		
			ValAcc/%	TestAcc/%	ValAcc/%	ValANLS	TestANLS
Singh等人(2019a)	Rosetta-ml	No	26.56	27.63	–	–	–
Gómez等人(2021)	Rosetta-ml	No	26.07			–	–
Wu等人(2022)	Mask R-CNN+CRNN	No	28.42	28.9	–	–	0.282
MM-GNN(Gao等,2020b)	Rosetta-ml	No	31.44	31.1	–	–	0.207
M4C(Hu等,2020)	Rosetta-en	No	39.4	39.01	38.05	0.472	0.462
RUArt(Jin等,2020)	Rosetta-en	No	–	40.45	–	–	0.481
SMA(Gao等,2021)	Rosetta-en	No	40.39	40.86	–	–	0.486
CRN(Liu等,2020)	Rosetta-en	No	40.39	40.96	–	–	0.483
LaAP-Net(Han等,2020)	Rosetta-en	No	40.68	40.54	39.74	0.497	0.485
BOV(Zeng等,2021)	Rosetta-en	No	40.9	41.23	40.18	0.5	0.472
Sharma和Jalal (2022)	Rosetta-en	No	41.23	41.03	40.98	0.478	0.463
TIG(Li等,2022)	Rosetta-en	Yes	41.09	40.77	–	–	–
PixelM4C(Singh等,2021)	MaskTextSpotterv3	Yes	42.12	–	–	–	–
Zhang和Yang(2021)	Google-OCR	No	42.8	43.41	41.1	–	0.508
SA-M4C(Kant等,2020)	Google-OCR	No	43.9	–	42.23	0.512	0.504
SSBaseline(Zhu等,2020)	SBD-Trans	Yes	45.53	45.66	–	–	0.55
LOGOS(Lu等,2021)	Microsoft-OCR	Yes	<u>51.53</u>	<u>51.08</u>	<u>48.63</u>	<u>0.581</u>	<u>0.579</u>
TAP(Yang等,2021)	Microsoft-OCR	Yes	54.71	53.97	50.83	0.598	0.597
Latr(Biten等,2020)	Amazon-OCR	Yes	**61.05**	**61.6**	**61.64**	**0.702**	**0.696**

注:加粗字体表示各列最优结果,“–”表示对应文献未给出实验结果,下划线字体表示使用了额外的训练数据集。

和模态间的自由交互，性能更优，框架也更简洁。但是与M4C相比，大多数在M4C上改进的工作，诸如SMA（structured multimodal attention）（Gao等，2020）、CRN（cascade reasoning network）（Liu 等，2020）、LAAP-Net（localization aware answer prediction network）（Han等，2020）、RUArt（Jin等，2020）等，所带来的改进并不显著。目前性能较好的模型都受益于更强大的OCR模块或额外的训练数据。基于强大的预训练语言模型t5和对扫描文档的布局感知预训练方案LATR（Biten等，2021）取得了显著的效果。在未来，如何提升模型推理能力、提高模型对OCR结果的鲁棒性、更充分地利用视觉模态信息仍是值得探索的课题。

5 场景文字图像描述

图像描述是多模态学习中研究最广泛的课题之一，并在近几年取得了巨大发展。然而传统的图像描述方法并没有意识到场景文字的重要性，因此它们不能对图像进行准确全面的描述。为了解决这一问题，Sidorov等人（2020）提出了基于文本的图像描述任务并创建了专用的数据集TextCaps。给定一个包含场景文字的图像，该任务的目标是在考虑图像视觉信息和场景文字的情况下生成一个合理的描述。本节介绍该任务的常用数据集TextCaps，根据生成图像描述的类型，将以场景文字为中心的图像描述方法分为面向准确性、多样性和可控性的场景文字图像描述方法。

5.1 场景文字图像描述数据集

Facebook AI Research于2020年提出TextCaps（Sidorov等，2020）数据集，是最常用的基于场景文本的图像描述数据集。它与TextVQA（Singh等，2019a）数据集共享图像数据，该数据集包含28 408幅图像和与之对应的145 329个描述句子，平均每个描述语句包含12.4个单词。

5.2 场景文字图像描述评价指标

与传统的图像描述方法一样，以场景文字为中心的图像描述方法将生成的图像描述与手动标注的描述进行比较。采用的常见评估指标，包括BLEU（method for automatic evaluation of machine translation）（Papineni等，2002）、METEOR（metric for evaluation of translation with explicit ordering）（Agarwal 和Lavie，2008）、ROUGE-L（recall-oriented understudy for gisting evaluation）（Lin，2004）、SPICE（semantic propositional image caption evaluation）（Anderson等，2016）和CIDEr（Vedantam等，2015）。在这些指标中，CIDEr是与人类评价最接近的，因为它考虑了每个单词的TF-IDF（term frequency - inverse document frequency）分数，更多地关注场景文本信息。

对于面向多样性的TextCaps系统，性能可以在分词（token）级别或语义级别上进行评估。对于分词级度量，Div-n（n = 1，2）（Li等，2016）用来衡量单元组（unigrams）或双元组（bigrams）占描述单词总数的比例。对于语义级的度量，SelfCIDEr（Wang和Chan，2019）使用CIDEr作为潜在语义分析的核心函数，这与人类的多样性判断结果较好地契合。

5.3 面向准确性的场景文字图像描述方法

面向准确性的场景文字图像描述方法目的是生成与人工标注更加接近的图像描述。场景文字图像描述的基线方法M4C-Captioner（Sidorov等，2020）在沿用M4C的基础结构的同时，去除了M4C中的问题编码输入。通过实验证明，M4C-Captioner在TextCaps数据集上的性能显著优于没有考虑场景文字的图像描述方法，这验证了场景文字对于图像描述的重要性。MMA-SR（multimodal attention captioner with OCR spatial relationship）（Wang等，2020a）和LSTM-R（Wang等，2021c）使用带有注意力机制的LSTM网络来建模上下文信息，并显式地考虑了场景文字之间的空间几何关系，取得了较好的性能。在M4C-Captioner（Sidorov等，2020）的基础上，CNMT（confidence-aware non-repetitive multimodal Transformers）（Wang等，2021d）提出了一个更好的场景文字阅读模块，并且将场景文字的识别置信度作为其重要性的衡量标准来选取更重要的场景文字，显著提高了模型的性能，而且该模型还通过在生成模块中利用重复掩码来抑制预测中的冗余单词。此外，经典的场景文字视觉问答方法SSbaseline（Zhu等，2021）和TAP（text-aware pre-training）（Yang等，2021）也迁移到了场景文字图像描述领域，取得了显著的性能提升。

5.4 面向多样性的场景文字图像描述方法

由于一幅图像往往包含复杂的文本和视觉信息，很难得到全面而准确的描述，一些工作开始进行多样性描述的探索。Anchor-Captioner（Xu等，2021a）试图生成多个描述来准确详细地描述图像的不同部

分。进一步地，MAGIC（multimodal relational graph adversarial inference）（Zhang 和 Yang，2021）在选定中心目标的基础上，自适应地构建多个多模态关系图学习对齐的特征表示，然后在不使用成对图像—文本的情况下，通过级联生成对抗网络生成多样化的场景文字图像描述。

5.5 面向可控性的场景文字图像描述方法

针对目前的方法不能根据各种信息需求生成个性化的图像描述，Hu 等人（2021）提出了问题可控的场景文字图像描述（Qc-TextCap）任务。对应的 GQAM（geometry and question aware model）模型首先使用视觉编码器根据空间关系对物体和场景文字信息进行融合，然后使用问题引导的编码器选择与给定问题最相关的视觉特征。通过这种方式，可以生成兼具信息量和多样性的描述，更具实用性。

5.6 场景文字图像描述方法性能对比分析

现有方法在 TextCaps 数据集上的性能比较结果如表 6 所示。LSTM-R（Wang 等，2021c）和 TAP（Yang 等，2021）模型在 5 个指标上取得了最优的结果。这说明显式地建模文字之间的几何关系对于预测有意义的图像描述是很有必要的。TAP 中设计的文字感知预训练技术也适用于 TextCaps 任务，并有助于提高性能。在多样性度量 SelfCIDEr 方面，Xu 等人（2021a）提出的 Anchor-Captioner 模型显著优于 M4C-Captioner（Sidorov 等，2020）基线方法。虽然 MAGIC（Zhang 和 Yang，2021）模型的性能落后于 Anchor-Captioner，但 MAGIC 不需要成对的图像—文本数据，在模型泛化性方面更具优势。

现有场景文字图像描述方法主要借鉴传统图像描述技术，采取相似的框架和性能指标。然而，场景文字作为一种高层语义信息，对全局图像理解和细致的图像描述生成应该有更广泛的作用。如何综合考虑准确性、连贯性和合理性，设计出更适用于场景文字图像描述的性能指标，对场景文字图像描述任务的发展具有重要意义。另外，如何提高模型对 OCR 结果的鲁棒性、如何更充分地利用视觉模态信息、如何利用外部知识辅助生成图像描述仍是未来值得探索的课题。

表 6 场景文字图像描述方法性能对比分析

Table 6 Performance comparison of scene text image captioning

/%

方法	BLEU-4	METEOR	ROUGE	SPICE	CIDEr	SelfCIDEr
M4C-Captioner（Sidorov 等，2020）	23.3	22.0	46.2	15.6	89.6	49.4
MMA-SR（Wang 等，2020a）	24.6	23.0	47.3	16.2	98.0	–
SSBaseline（Zhu 等，2020）	24.9	22.7	47.2	15.7	98.8	–
CNMT（Wang 等，2021d）	24.8	23.0	47.1	16.3	101.7	–
LSTM-R（Wang 等，2021c）	**27.9**	23.7	**49.1**	16.6	**109.3**	–
TAP（Yang 等，2021）	25.8	**23.8**	47.9	**17.1**	109.2	–
PixelM4C-Captioner（Singh 等，2021）	23.3	21.3	45.7	14.6	85.3	–
Anchor-Captioner（Xu 等，2021a）	24.7	22.5	47.1	15.9	95.5	**57.6**
MAGIC（Zhang 和 Yang，2021）	22.2	20.5	42.3	13.8	76.6	49.6

注：加粗字体表示各列最优结果，“–”表示对应文献未给出实验结果。

6 结 语

以文字为中心的图像理解需要对自然场景图像或视觉丰富的文档图像进行信息挖掘、理解和分析，是人工智能从感知走向认知过程的重要技术。本文对近年来以文字为中心的理解技术进行综述，并按照理解认知程度，将以文字为中心的理解任务划分成两类进行梳理，分析了各任务各类方法的优点和局限性。希望通过本篇综述能够促进以文字为中心的图像理解技术的进步，启发模型和算法的新设计。

第 1 类以文字为中心的图像理解任务仅要求模型具备抽取信息的能力，包含视觉信息抽取和场景文字检索任务。大多数视觉信息抽取方法建模为标

记分类任务，预先定义文档图像中文字的类别，利用图神经网络或Transformer框架，充分融合检测识别后的多模态信息，预测文档图像中文字的类别。目前基于OCR的视觉信息抽取方法已取得很大进步，但OCR模块可能存在计算成本高、对多语言文档图像的泛化性能差以及由于检测识别错误导致累积误差传播等问题。此外，实际场景中实体类别数量通常较大，现有方法将信息抽取任务视做分类任务，需要大量的人工标记。如何通过端到端的视觉信息抽取模型，获得可供机器直接使用或分析的结构化信息是目前视觉信息抽取任务的重点。现有的场景文字检索可以从图像库中定位和检索出给定查询文字指示的文字实例，但尚未充分利用图像中的多模态信息。

第2类以文字为中心的图像理解任务不仅要求模型具备抽取信息的能力，而且要求模型具备一定的分析和推理能力，如文档视觉回答、场景文字视觉回答和场景文字图像描述。该类任务的通常方法首先对输入的多模态信息（问题的文本特征、视觉物体特征以及图像中的文本特征等）进行特征提取，然后将它们映射到一个公共特征空间，进行模态内和模态间的信息融合交互，最后模型通过由训练集中高频答案词组成的固定词典和图像中文字组成的动态词典回答问题或生成图像描述。目前，该方法在面对可以从图像中找到答案的问题和训练集中出现过的图像描述时性能较好，但是对涉及复杂的线索推理、常识判断时仍有局限性，而且比较依赖前端文字检测识别的结果。

总体而言，在计算机视觉、自然语言处理技术的推动下，以文字为中心的图像理解取得了较大进步，但仍然面临严峻挑战。如何有效利用多模态信息、如何减小文字感知精度对文字理解性能的影响、如何更好地结合外部知识进行推理，是目前以文字为中心的图像理解领域的主要难题。另外，为了促进以文字为中心的图像理解技术向工业界的落地应用，提高处理效率、提升模型鲁棒性以及扩宽应用领域也是未来的发展方向。

参考文献（References）

Agarwal A and Lavie A. 2008. METEOR, M-BLEU and M-TER: evaluation metrics for high-correlation with human rankings of machine translation output//Proceedings of the 3rd Workshop on Statistical Machine Translation. Columbus, USA: Association for Computational Linguistics: 115-118 [DOI: 10.3115/1626394.1626406]

Anand M, Karteek A and Jawahar C V. 2013. Image retrieval using textual cues//Proceedings of 2013 ICCV.[s.l.]: [s.n.]

Almazán J, Gordo A, Fornés A and Valveny E. 2014. Word spotting and recognition with embedded attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36 (12): 2552-2566 [DOI: 10.1109/TPAMI.2014.2339814]

Anderson P, Fernando B, Johnson M and Gould S. 2016. SPICE: semantic propositional image caption evaluation//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 382-398 [DOI: 10.1007/978-3-319-46454-1_24]

Appalaraju S, Jasani B, Kota B U, Xie Y S and Manmatha R. 2021. DocFormer: end-to-end transformer for document understanding//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 973-983 [DOI: 10.1109/ICCV48922.2021.00103]

Biten A F, Litman R, Xie Y S, Appalaraju S and Manmatha R. 2021. LaTr: layout-aware transformer for scene-text VQA [EB/OL]. [2022-09-10]. https://arxiv.org/pdf/2112.12494.pdf

Biten A F, Tito R, Mafla A, Gomez L, Rusiñol M, Jawahar C V, Valveny E and Karatzas D. 2019. Scene text visual question answering//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4290-4300 [DOI: 10.1109/ICCV.2019.00439]

Carbonell M, Riba P, Villegas M, Fornés A and Lladós J. 2020. Named entity recognition and relation extraction with graph neural networks in semi structured documents//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE: 9622-9627 [DOI: 10.1109/ICPR48806.2021.9412669]

Chu X X, Tian Z, Zhang B, Wang X L and Shen C H. 2021. Conditional positional encodings for vision transformers [EB/OL]. [2022-09-10]. https://arxiv.org/pdf/2102.10882.pdf

Cui L, Xu Y H, Lv T C and Wei F R. 2021. Document AI: benchmarks, models and applications [EB/OL]. [2022-09-10]. https://arxiv.org/pdf/2111.08609.pdf

Dai Z H, Yang Z L, Yang Y M, Carbonell J, Le Q and Salakhutdinov R. 2019. Transformer-XL: attentive language models beyond a fixed-length context//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics: 2978-2988 [DOI: 10.18653/v1/P19-1285]

Denk T I and Reisswig C. 2019. BERTgrid: contextualized embedding for 2D document representation and understanding [EB/OL]. [2022-09-10]. https://arxiv.org/pdf/1909.04948.pdf

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understand-

ing//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota, USA: Association for Computational Linguistics: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale [EB/OL]. [2022-09-10]. https://arxiv.org/pdf/2010.11929.pdf

Gao C Y, Zhu Q, Wang P, Li H, Liu Y L, van den Hengel A and Wu Q. 2020a. Structured multimodal attentions for TextVQA [EB/OL]. [2022-09-10]. https://arxiv.org/pdf/2006.00753.pdf

Gao D F, Li K, Wang R P, Shan S G and Chen X L. 2020b. Multi-modal graph neural network for joint reasoning on vision and scene text//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12743-12753 [DOI: 10.1109/CVPR42600.2020.01276]

Ghosh S K, Gómez L, Karatzas D and Valveny E. 2015. Efficient indexing for query by string text retrieval//Proceedings of the 13th International Conference on Document Analysis and Recognition. Tunis, Tunisia: IEEE: 1236-1240 [DOI: 10.1109/ICDAR.2015.7333961]

Gómez L, Biten A F, Tito R, Mafla A, Rusiñol M, Valveny E and Karatzas D. 2021. Multimodal grid features and cell pointers for scene text visual question answering. Pattern Recognition Letters, 150: 242-249 [DOI: 10.1016/j.patrec.2021.06.026]

Gómez L, Mafla A, Rusiñol M and Karatzas D. 2018. Single shot scene text retrieval//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 728-744 [DOI: 10.1007/978-3-030-01264-9_43]

Graves A and Schmidhuber J. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5/6): 602-610 [DOI: 10.1016/j.neunet.2005.06.042]

Gu Z X, Meng C H, Wang K, Lan J, Wang W Q, Gu M and Zhang L Q. 2022. XYLayoutLM: towards layout-aware multimodal networks for visually-rich document understanding [EB/OL]. [2022-09-10]. https://arxiv.org/pdf/2203.06947.pdf

Han W, Huang H T and Han T. 2020. Finding the evidence: localization-aware answer prediction for text visual question answering//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain: International Committee on Computational Linguistics: 3118-3131 [DOI: 10.18653/v1/2020.coling-main.278]

He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2980-2988 [DOI: 10.1109/ICCV.2017.322]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hu A W, Chen S Z and Jin Q. 2021. Question-controlled text-aware image captioning//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: ACM: 3097-3105 [DOI: 10.1145/3474085.3475452]

Hu R H, Singh A, Darrell T and Rohrbach M. 2020. Iterative answer prediction with pointer-augmented multimodal Transformers for TextVQA//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9989-9999 [DOI: 10.1109/CVPR42600.2020.01001]

Huang Y P, Lv T C, Cui L, Lu Y T and Wei F R. 2022. LayoutLMv3: pre-training for document AI with unified text and image masking [EB/OL]. [2022-09-10]. https://arxiv.org/pdf/2204.08387.pdf

Huang Z, Chen K, He J H, Bai X, Karatzas D, Lu S J and Jawahar C V. 2019. ICDAR2019 competition on scanned receipt OCR and information extraction//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia: IEEE: 1516-1520 [DOI: 10.1109/ICDAR.2019.00244]

Jaume G, Ekenel H K and Thiran J P. 2019. FUNSD: a dataset for form understanding in noisy scanned documents//Proceedings of 2019 International Conference on Document Analysis and Recognition Workshops. Sydney, Australia: IEEE: 1-6 [DOI: 10.1109/ICDARW.2019.10029]

Jin Z X, Wu H R, Yang C, Zhou F, Qin J Y, Xiao L and Yin X C. 2020. RUArt: a novel text-centered solution for text-based visual question answering [EB/OL]. [2022-09-10]. https://arxiv.org/pdf/2010.12917.pdf

Joulin A, Grave E, Bojanowski P and Mikolov T. 2017. Bag of tricks for efficient text classification//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Valencia, Spain: Association for Computational Linguistics: 427-431 [DOI: 10.18653/v1/E17-2068]

Kai W, Boris B and Serge J B. 2011. End-toend scene text recognition//Proceedings of 2011 ICCV. [s.l.]: [s.n.]

Kant Y, Batra D, Anderson P, Schwing A, Parikh D, Lu J S and Agrawal H. 2020. Spatially aware multimodal Transformers for TextVQA//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 715-732 [DOI: 10.1007/978-3-030-58545-7_41]

Katti A R, Reisswig C, Guder C, Brarda S, Bickel S, Höhne J and Faddoul J B. 2018. Chargrid: towards understanding 2d documents//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics: 4459-4469 [DOI: 10.18653/v1/D18-1476]

Kim W, Son B and Kim I. 2021. ViLT: vision-and-language transformer

without convolution or region supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR: 5583-5594

Krasin I, Duerig T, Alldrin N, Veit A, Abu-El-Haija S, Belongie S, Cai D, Feng Z Y, Ferrari V and Gomes V. 2016. Openimages: a public dataset for large-scale multi-label and multi-class image classification.

Lafferty J D, McCallum A and Pereira F C N. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data//Proceedings of the 18th International Conference on Machine Learning. Williams College, USA: Morgan Kaufmann: 282-289

Lee C Y, Li C L, Dozat T, Perot V, Su G L, Hua N, Ainslie J, Wang R S, Fujii Y and Pfister T. 2022. FormNet: structural encoding beyond sequential modeling in form document information extraction//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics: 3735-3754 [DOI: 10.18653/v1/2022.acl-long.260]

Li C L, Bi B, Yan M, Wang W, Huang S F, Huang F and Si L. 2021a. StructuralLM: Structural pre-training for form understanding//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Virtual: Association for Computational Linguistics: 6309-6318 [DOI: 10.18653/v1/2021.acl-long.493]

Li J W, Galley M, Brockett C, Gao J F and Dolan B. 2016. A diversity-promoting objective function for neural conversation models//Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA: The Association for Computational Linguistics: 110-119 [DOI: 10.18653/v1/N16-1014]

Li P Z, Gu J X, Kuen J, Morariu V I, Zhao H D, Jain R, Manjunatha V and Liu H F. 2021b. SelfDoc: self-supervised document representation learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5648-5656 [DOI: 10.1109/CVPR46437.2021.00560]

Li X P, Wu B, Song J K, Gao L L, Zeng P P and Gan C. 2022. Text-instance graph: exploring the relational semantics for text-based visual question answering. Pattern Recognition, 124: #108455 [DOI: 10.1016/j.patcog.2021.108455]

Li Y L, Qian Y X, Yu Y C, Qin X M, Zhang C Q, Liu Y, Yao K, Han J Y, Liu J T and Ding E R. 2021c. StrucTexT: structured text understanding with multi-modal transformers//Proceedings of the 29th ACM International Conference on Multimedia. Virtual, China: ACM: 1912-1920 [DOI: 10.1145/3474085.3475345]

Liao M H, Pang G, Huang J, Hassner T and Bai X. 2020. Mask TextSpotter v3: segmentation proposal network for robust scene text spotting//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 706-722 [DOI: 10.1007/978-3-030-58621-8_41]

Lin C Y. 2004. ROUGE: a package for automatic evaluation of summaries//Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics: 74-81

Lin W H, Gao Q F, Sun L, Zhong Z Y, Hu K, Ren Q and Huo Q. 2021. VIBERTgrid: a jointly trained multi-modal 2D document representation for key information extraction from documents//Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer: 548-563 [DOI: 10.1007/978-3-030-86549-8_35]

Liu C Y, Chen X X, Luo C J, Jin L W, Xue Y and Liu Y L. 2021. Deep learning methods for scene text detection and recognition. Journal of Image and Graphics, 26(6): 1330-1367 (刘崇宇, 陈晓雪, 罗灿杰, 金连文, 薛洋, 刘禹良. 2021. 自然场景文本检测与识别的深度学习方法. 中国图象图形学报, 26(6): 1330-1367) [DOI: 10.11834/jig.210044]

Liu F, Xu G H, Wu Q, Du Q, Jia W and Tan M K. 2020. Cascade reasoning network for text-based visual question answering//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 4060-4069 [DOI: 10.1145/3394171.3413924]

Liu T Y. 2009. Learning to rank for information retrieval. Foundations and Trends in Information Retrieval, 3(3): 225-331 [DOI: 10.1561/1500000016]

Liu X J, Gao F Y, Zhang Q and Zhao H S. 2019. Graph convolution for multimodal information extraction from visually rich documents//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers). Minneapolis, Minnesota: Association for Computational Linguistics: 32-39 [DOI: 10.18653/v1/N19-2005]

Lu X P, Fan Z, Wang Y S, Oh J and Rosé C P. 2021. Localize, group, and select: boosting text-VQA by scene text modeling//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 2631-2639 [DOI: 10.1109/ICCVW54120.2021.00297]

Mafla A, Tito R, Dey S, Gómez L, Rusiñol M, Valveny E and Karatzas D. 2021. Real-time lexicon-free scene text retrieval. Pattern Recognition, 110: #107656 [DOI: 10.1016/j.patcog.2020.107656]

Mathew M, Bagal V, Tito R, Karatzas D, Valveny E and Jawahar C V. 2022. InfographicVQA//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2582-2591 [DOI: 10.1109/WACV51458.2022.00264]

Mathew M, Karatzas D and Jawahar C V. 2021. DocVQA: a dataset for VQA on document images//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2199-2208 [DOI: 10.1109/WACV48630.2021.00225]

Mishra A, Alahari K and Jawahar C V. 2013. Image retrieval using textual cues//Proceedings of 2013 IEEE International Conference on

Computer Vision. Sydney, Australia: IEEE: 3040-3047 [DOI: 10.1109/ICCV.2013.378]

Mishra A, Shekhar S, Singh A K and Chakraborty A. 2019. OCR-VQA: visual question answering by reading text in images//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia: IEEE: 947-952 [DOI: 10.1109/ICDAR.2019.00156]

Papineni K, Roukos S, Ward T and Zhu W J. 2002. BLEU: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, USA: ACL: 311-318 [DOI: 10.3115/1073083.1073135]

Pennington J, Socher R and Manning C. 2014. GloVe: global vectors for word representation//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: ACL: 1532-1543 [DOI: 10.3115/v1/D14-1162]

Powalski R, Borchmann Ł, Jurkiewicz D, Dwojak T, Pietruszka M and Pałka G. 2021. Going full-TILT boogie on document understanding with text-image-layout transformer//Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer: 732-747 [DOI: 10.1007/978-3-030-86331-9_47]

Qian Y J, Santus E, Jin Z J, Guo J and Barzilay R. 2019. GraphIE: a graph-based framework for information extraction//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics: 751-761 [DOI: 10.18653/v1/N19-1082]

Qiao Z, Zhou Y, Wei J, Wang W, Zhang Y, Jiang N, Wang H B and Wang W P. 2021. PIMNet: a parallel, iterative and mimicking network for scene text recognition//Proceedings of the 29th ACM International Conference on Multimedia. Virtual, China: ACM: 2046-2055 [DOI: 10.1145/3474085.3475238]

Qiao Z, Zhou Y, Yang D B, Zhou Y C and Wang W P. 2020. SEED: semantics enhanced encoder-decoder framework for scene text recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 13525-13534 [DOI: 10.1109/CVPR42600.2020.01354]

Qin X G, Zhou Y, Guo Y H, Wu D Y, Tian Z H, Jiang N, Wang H B and Wang W P. 2021. Mask is all you need: rethinking mask R-CNN for dense and arbitrary-shaped scene text detection//Proceedings of the 29th ACM International Conference on Multimedia. Virtual, China: ACM: 414-423 [DOI: 10.1145/3474085.3475178]

Rajpurkar P, Zhang J, Lopyrev K and Liang P. 2016. SQuAD: 100, 000+ questions for machine comprehension of text//Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA: The Association for Computational Linguistics: 2383-2392 [DOI: 10.18653/v1/D16-1264]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Rong X J, Yi C C and Tian Y L. 2020. Unambiguous scene text segmentation with referring expression comprehension. IEEE Transactions on Image Processing, 29: 591-601 [DOI: 10.1109/TIP.2019.2930176]

Rong X J, Yi C C and Tian Y L. 2022. Unambiguous text localization, retrieval, and recognition for cluttered scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3): 1638-1652 [DOI: 10.1109/TPAMI.2020.3018491]

Sang E F T K and Veenstra J. 1999. Representing text chunks//Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics. Bergen, Norway: The Association for Computer Linguistics: 173-179 [DOI: 10.3115/977035.977059]

Sharma H and Jalal A S. 2022. Improving visual question answering by combining scene-text information. Multimedia Tools and Applications, 81(9): 12177-12208 [DOI: 10.1007/s11042-022-12317-0]

Sidorov O, Hu R H, Rohrbach M and Singh A. 2020. TextCaps: a dataset for image captioning with reading comprehension//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 742-758 [DOI: 10.1007/978-3-030-58536-5_44]

Singh A, Natarajan V, Jiang Y, Chen X, Shah M, Rohrbach M, Batra D and Parikh D. 2018. Pythia —— a platform for vision and language research//SysML Workshop, NeurIPS. Montréal, Canada: MIT Press

Singh A, Natarajan V, Shah M, Jiang Y, Chen X L, Batra D, Parikh D and Rohrbach M. 2019a. Towards VQA models that can read//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8309-8318 [DOI: 10.1109/CVPR.2019.00851]

Singh A, Pang G, Toh M, Huang J, Galuba W and Hassner T. 2021. TextOCR: towards large-scale end-to-end reasoning for arbitrary-shaped scene text//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8798-8808 [DOI: 10.1109/CVPR46437.2021.00869]

Singh A K, Mishra A, Shekhar S and Chakraborty A. 2019b. From strings to things: knowledge-enabled VQA model that can read and reason//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4601-4611 [DOI: 10.1109/ICCV.2019.00470]

Tang G Z, Xie L L, Jin L W, Wang J P, Chen J D, Xu Z, Wang Q Y, Wu Y Q and Li H. 2021. MatchVIE: exploiting match relevancy between entities for visual information extraction//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada: IJCAI.org: 1039-1045 [DOI: 10.24963/ijcai.2021/144]

Tito R, Karatzas D and Valveny E. 2021. Document collection visual question answering//Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer: 778-792 [DOI: 10.1007/978-3-030-86331-9_50]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Vedantam R, Zitnick C L and Parikh D. 2015. CIDEr: consensus-based image description evaluation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 4566-4575 [DOI: 10.1109/CVPR.2015.7299087]

Wang H, Bai X, Yang M K, Zhu S G, Wang J and Liu W Y. 2021a. Scene text retrieval via joint text detection and similarity learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4556-4565 [DOI: 10.1109/CVPR46437.2021.00453]

Wang J, Tang J H and Luo J B. 2020a. Multimodal attention with image text spatial relationship for OCR-based image captioning//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 4337-4345 [DOI: 10.1145/3394171.3413753]

Wang J, Tang J H, Yang M K, Bai X and Luo J B. 2021c. Improving OCR-based image captioning by incorporating geometrical relationship//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1306-1315 [DOI: 10.1109/CVPR46437.2021.00136]

Wang J P, Jin L W and Ding K. 2022a. LiLT: a simple yet effective language-independent layout Transformer for structured document understanding//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics: 7747-7757 [DOI: 10.18653/v1/2022.acl-long.534]

Wang J P, Liu C Y, Jin L W, Tang G Z, Zhang J X, Zhang S T, Wang Q Y, Wu Y Q and Cai M X. 2021b. Towards robust visual information extraction in real world: new dataset and novel solution. Proceedings of the AAAI Conference on Artificial Intelligence, 35(4): 2738-2745 [DOI: 10.1609/aaai.v35i4.16378]

Wang Q Z and Chan A B. 2019. Describing like humans: on diversity in image captioning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4190-4198 [DOI: 10.1109/CVPR.2019.00432]

Wang W, Zhou Y, Lv J H, Wu D Y, Zhao G Q, Jiang N and Wang W P. 2022b. TPSNet: reverse thinking of thin plate splines for arbitrary shape scene text representation//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 5014-5025 [DOI: 10.1145/3503161.3547882]

Wang X Y, Liu Y L, Shen C H, Ng C C, Luo C J, Jin L W, Chan C S, van den Hengel A and Wang L W. 2020b. On the general value of evidence, and bilingual scene-text visual question answering//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10123-10132 [DOI: 10.1109/CVPR42600.2020.01014]

Wang Z K, Bao R D, Wu Q and Liu S. 2021d. Confidence-aware non-repetitive multimodal Transformers for TextCaps. Proceedings of the AAAI Conference on Artificial Intelligence, 35(4): 2835-2843 [DOI: 10.1609/aaai.v35i4.16389]

Wei J, Zhang Y, Zhou Y, Zeng G Y, Qiao Z, Guo Y H, Wu H Y, Wang H B and Wang W P. 2022. TextBlock: towards scene text spotting without fine-grained detection//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 5892-5902 [DOI: 10.1145/3503161.3548051]

Wu J J, Du J, Wang F R, Yang C, Jiang X Z, Hu J S, Yin B, Zhang J S and Dai L R. 2022. A multimodal attention fusion network with a dynamic vocabulary for TextVQA. Pattern Recognition, 122: #108214 [DOI: 10.1016/j.patcog.2021.108214]

Xu G H, Niu S C, Tan M K, Luo Y C, Du Q and Wu Q. 2021a. Towards accurate text-based image captioning with content diversity exploration//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12632-12641 [DOI: 10.1109/CVPR46437.2021.01245]

Xu Y, Xu Y H, Lyu T C, Cui L, Wei F R, Wang G X, Lu Y J, Florêncio D, Zhang C, Che W X, Zhang M and Zhou L D. 2021c. LayoutLMv2: multi-modal pre-training for visually-rich document understanding//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics: 2579-2591 [DOI: 10.18653/v1/2021.acl-long.201]

Xu Y H, Li M H, Cui L, Huang S H, Wei F R and Zhou M. 2020. LayoutLM: pre-training of text and layout for document image understanding//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Virtual, USA: ACM: 1192-1200 [DOI: 10.1145/3394486.3403172]

Xu Y H, Lv T C, Cui L, Wang G X, Lu Y J, Florêncio D, Zhang C and Wei F R. 2021b. LayoutXLM: multimodal pre-training for multilingual visually-rich document understanding [EB/OL]. [2022-09-10]. https://arxiv.org/pdf/2104.08836.pdf

Xu Y H, Lyu T C, Cui L, Wang G X, Lu Y J, Florêncio D, Zhang C and Wei F R. 2022. XFUND: a benchmark dataset for multilingual visually rich form understanding//Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics: 3214-3224 [DOI: 10.18653/v1/2022.findings-acl.253]

Yang Z C, He X D, Gao J F, Deng L and Smola A. 2016. Stacked attention networks for image question answering//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las

Vegas，USA：IEEE：21-29［DOI：10.1109/CVPR.2016.10］

Yang Z Y，Lu Y J，Wang J F，Yin X，Florêncio D，Wang L J，Zhang C，Zhang L and Luo J B. 2021. TAP：text-aware pre-training for text-VQA and text-caption//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：8747-8757［DOI：10.1109/CVPR46437.2021.00864］

Zeng G Y，Zhang Y，Zhou Y and Yang X M. 2021. Beyond OCR + VQA：involving OCR into the flow for robust and accurate Text-VQA//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event，China：ACM：376-385［DOI：10.1145/3474085.3475606］

Zhang P，Xu Y L，Cheng Z Z，Pu S L，Lu J，Qiao L，Niu Y and Wu F. 2020. TRIE：end-to-end text reading and information extraction for document understanding//Proceedings of the 28th ACM International Conference on Multimedia. Seattle，USA：ACM：1413-1422［DOI：10.1145/3394171.3413900］

Zhang W Q，Shi H C，Guo J N，Zhang S Y，Cai Q P，Li J C，Luo S H and Zhuang Y T. 2022. MAGIC：multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning［EB/OL］.［2022-09-10］. https://arxiv.org/pdf/2112.06558.pdf

Zhang X Y and Yang Q. 2021. Position-augmented Transformers with entity-aligned mesh for TextVQA//Proceedings of the 29th ACM International Conference on Multimedia. Virtual，China：ACM：2519-2528［DOI：10.1145/3474085.3475425］

Zhu C G，Zeng M and Huang X D. 2019. SDNet：contextualized attention-based deep network for conversational question answering［EB/OL］.［2022-09-10］. https://arxiv.org/pdf/1812.03593.pdf

Zhu Q，Gao C Y，Wang P and Wu Q. 2021. Simple is not easy：a simple strong baseline for TextVQA and TextCaps. Proceedings of the AAAI Conference on Artificial Intelligence，35(4)：3608-3615［DOI：10.1609/aaai.v35i4.16476］

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//Workshop on Document Intelligence at NeurIPS 2019. Vancouver，Canada：［s.n.］

Kerroumi M，Sayem O and Shabou A. 2020. VisualWordGrid：information extraction from scanned documents using a multimodal approach［EB/OL］.［2022-09-10］. http://arxiv.org/pdf/2010.02358.pdf

Hong T，Kim D，Ji M，Hwang W，Nam D and Park S. 2021. BROS：a pre-trained language model focusing on text and layout for better key information extraction from documents//Proceedings of AAAI Conference on Artificial Intelligence. Vancouver，Canada：AAAI：10767-10775

Zhang P，Xu Y，Cheng Z，Pu S，Lu J，Qiao L，Niu Y and Wu F. 2020. TRIE：end-to-end text reading and information extraction for document understanding//Proceedings of the 28th ACM International Conference on Multimedia. Seattle，USA：ACM

Yu W，Lu N，Qi X，Gong P and Xiao R. 2020. PICK：processing key information extraction from documents using improved graph learning-convolutional networks//Proceedings of the 25th International Conference on Pattern Recognition（ICPR）. Milan，Italy：IEEE：4363-4370

Gu J，Kuen J，Morariu V I，Zhao H，Barmpalios N，Jain R，Nenkova A and Sun T. 2022. Unified pretraining framework for document understanding［EB/OL］.［2022-09-10］. https://arxiv.org/pdf/2204.10939.pdf

Veit A，Matera T，Neumann L，Matas J and Belongie S J. 2016. COCO-Text：dataset and benchmark for text detection and recognition in natural images［EB/OL］.［2022-09-10］. https://arxiv.org/pdf/1601.07140.pdf

Zhu Q，Gao C，Wang P and Wu Q. 2020. Simple is not easy：a simple strong baseline for textvqa and textcaps //Proceedings of the AAAI Conference on Artificial Intelligence. New York，USA：AAAI

作者简介

张言，男，博士研究生，主要研究方向为计算机视觉深度学习、视觉信息抽取和文档图像理解。

E-mail：zhangyan223@mails.ucas.edu.cn

周宇，通信作者，男，副研究员，博士生导师，主要研究方向为计算机视觉、深度学习、人工智能安全、场景文字理解、视觉目标检测、自监督学习及增量学习。

E-mail：zhouyu@iie.ac.cn

李强，男，硕士研究生，主要研究方向为计算机视觉深度学习、场景文字视觉描述。E-mail：liqiang@iie.ac.cn

申化文，男，硕士研究生，主要研究方向为计算机视觉、表格理解和文档图像理解。E-mail：shenhuawen@iie.ac.cn

曾港艳，女，博士研究生，主要研究方向为计算机视觉、文档图像理解和多模态学习。E-mail：zgy1997@cuc.edu.cn

马灿，男，正高级工程师，博士生导师，主要研究方向为网络空间大数据管理与分析、社交网络分析。

E-mail：macan@iie.ac.cn

张远，女，教授，博士生导师，主要研究方向为多媒体信号处理和人工智能。E-mail：yzhang@cuc.edu.cn

王伟平，男，研究员，博士生导师，主要研究方向为大数据安全和人工智能。E-mail：wangweiping@iie.ac.cn