# AI 设计下的文本视觉问答技术

晋赞霞[1]，覃京燕[1]，殷绪成[1,2]
（1.北京科技大学，北京 100083；2.北京科技大学顺德研究生院，佛山 528399）

**摘要：目的** 分析基于 AI 设计的文本视觉问答模型的有效性，旨在利用 AI 设计更好地指导当前 AI 模型的构建，提升模型效果和用户体验。**方法** 以传统文本视觉问答框架为基础，结合 AI 设计改进当前模型。具体包括加强基于场景设计原则的关系挖掘，根据不同理解层次需求的答案关键词预测，并对模型被投入应用所将面临的问题的分析。**结果** 基于 AI 设计完善模型可进一步提升模型效果；同时，通过 AI 设计对不同年龄认知差异的建模可指导回复生成，提升整体用户体验。**结论** 通过理论分析和实验对比，可以得出 AI 设计是 AI 技术投入到应用的一个重要步骤。基于 AI 设计对模型进行重构，可提高当前模型的效果，解决 AI 技术落地中将面临的用户体验问题，满足不同人群的需求。

**关键词：**AI 设计；AI；文本视觉问答；认知差异
**中图分类号：**TB472 **文献标识码：**A **文章编号：**1001-3563(2021)06-0007-06
**DOI：**10.19554/j.cnki.1001-3563.2021.06.002

## Text-based Visual Question Answering with AI Design

*JIN Zan-xia*[1], *QIN Jing-yan*[1], *YIN Xu-cheng*[1,2]
(1.University of Science and Technology Beijing, Beijing 100083, China; 2.Shunde Graduate School, University of Science and Technology Beijing, Foshan 528399, China)

**ABSTRACT:** It analyzes the effectiveness of the text visual question answering model based on AI design, aiming to better guide the construction of current artificial intelligence models with AI design, and improve model performance and user experience. It is based on the traditional text visual question answering framework, and the current model can be improved by combining AI design. Specifically, it includes strengthening relationship mining based on the principles of scenario design, predicting answer keywords according to the needs of different levels of understanding, and analyzing the problems that the model will face when it is put into application. Modifying the model based on AI design can further improve the performance of the model, and modeling the cognitive differences of different ages through AI design to guide response generation can improve the overall user experience. Through theoretical analysis and experimental comparison, it can be concluded that AI design is an important step in the application of AI technology. Reconstructing the model based on AI design can improve the performance of the current model, solve the user experience problems that will be faced in the implementation of AI technology, and meet the needs of different groups of people.

**KEY WORDS:** AI design; AI; text-based visual question answering; cognitive differences

随着人工智能（Artificial Intelligence，AI）的技术发展和深度学习的广泛应用，各界在自然语言处理领域和计算机视觉领域都取得了巨大的发展，如自然语言领域中的词性标注、机器翻译、自动问答等，以及计算机视觉领域中的物体检测、文本检测、人脸识别等。最近基于视觉和语言的多模态学习任务也引起

**收稿日期：**2020-12-08
**基金项目：**长江学者奖励项目（FRF-TP-18-010C1）；国家重大专项课题（2018YFB0704301）；北科大顺德研究生项目（BK19AE011）
**作者简介：**晋赞霞（1992—），女，山西人，北京科技大学博士生，主要研究方向为视觉与语言、多模态学习。
**通信作者：**覃京燕（1976—），女，四川人，博士，北京科技大学教授、博士生导师，主要研究方向为人工智能与创新设计、交互设计、信息设计。

了越来越多学者的关注，如图像字幕生成、视觉叙事、视觉问答等。虽然其中部分 AI 技术（如机器翻译、人脸识别等）已经得到了广泛的应用，但是还有一些 AI 技术（如文本视觉问答技术）距离实际应用还有很大差距。除了模型性能方面的影响，用户的使用体验也是一个很大的影响因素，因此只有结合 AI 设计更充分地对应用场景建模，才能满足不同的用户需求。

## 1 AI 设计下的文本视觉问答

文本视觉问答（Text Visual Question Answering，TextVQA）[1-2]是视觉推理的一个方向，即给定一张自然图片和与之相关的用自然语言描述的问题，该任务旨在通过理解图片场景中的视觉和文本信息，并使用自然语言回答该问题。文本视觉问答任务见图 1，根据输入的一张图片，提出“蓝色巴士去往哪里”的问题，计算机通过识别图片中的巴士、巴士颜色及其相关的文本等信息，推断出蓝色巴士前往“Swanage”。因此，文本视觉问答任务涉及到对问题的自然语言理解，对图片内容的视觉理解，对图片中文本的识别和语义理解，对涉及到的相关知识的表达，以及跨模态的信息融合与推理，这些都对机器智能提出了更高的挑战。目前自动文本视觉问答在对话系统、家政服务机器人、智能教育、视觉障碍人士的辅助工具等领域具有广泛的应用前景。

在现实生活中，将文本视觉问答系统最终落地到真实场景中的 AI 应用，仅仅实现基本的业务能力是完全不够的，还需要根据不同的需求进行相应的 AI 设计，才能更好地满足不同人群在不同场景中的需求。比如，在对话系统中，需要区分不同的人物主体，才能设计出相应的回答反馈层级。简单来说，如果是成年人对图片中的内容进行提问，那么返回的答案应该是准确的物体描述；如果是儿童对图片中的内容进行提问，那么返回的答案应该符合当前小孩的理解范围，应加入相应的描述帮助儿童理解图片内容。本文以问答系统为例，介绍文本视觉问答系统的实现和应用中可能面临的 AI 设计问题。

首先，自然场景中的文本通常是内容场景通过设计得到的最好的展示。设计是将信息转化为光学字符，而文本识别检测技术则是将光学字符转化为信息，从而理解场景内容。基于文本的视觉问答一般都是针对图片中的文本进行提问，因此图片的文本检测识别效果极大程度影响着后续问答模型的效果。在文本视觉问答中，常见的包含文本的载体包括指示牌、价签、说明、标志等，而这些文本通常是经过精心设计的，其中一些经常需要有通用的标准和设计形式，如指示牌、价签等，这些通用的标准和形式在文本检测识别时会相对容易一些。然而由于标志等的设计会加入丰富的创意，一般以图文结合的形式来表示相应的信息，在文本表示上有变形字体、曲形文本等多种设计方式，所以在文本检测和识别时存在一定难度。因此，如果可以根据不同的文本表现方式，加入一定的人工干预，也就是将设计原则规则化，比如将宽高比例、间隙、扭曲程度等数字化，并将其加入文本检测模型，可对文本检测准确性有一定提升，从而提高文本视觉问答模型的效果。

同时，设计师会通过设计隐式地告诉用户信息的重要程度与对应关系。比如一个瓶子上出现的最大的文本，用户会倾向于认为这个文本是该瓶子的品牌，而设计师也会通过一定的布局方式，告诉用户文本是如何对应的，比如菜单上的菜品和价格的对应关系。由此得到的启发是，用户也许可以根据设计思维先对页面进行分析，对场景进行区块划分，从而有针对性地理解场景中的内容。只有用户在对场景有所理解后，才能更好地回答用户关心的问题，才能将场景设计中想表达和承载的内容，翻译成用户能听到的自然语言，从而给出用户需要的答案。另外，图片拍摄也在一定程度上涉及到用户的艺术和设计思想，如果在拍摄当前场景时更能关注到用户所关注的问题，是否也能体现用户的意图，用户能否也利用该意图指导技术模型。正如当前所说的新媒体、数字媒体，多媒体、跨媒体等核心都是科技和艺术之间的关系。也就是说，在当代艺术观念语境下，光学媒介、电子媒介等技术手段成为了一种依托，完成了思维层面、精神层面的艺术表达。然而当前一些科学技术手段，比如文本视觉问答技术，其实是一种形式性的，将光学媒介传达的艺术观念和实体信息，反向挖掘反馈给人类的过程。因此，AI 设计和 AI 技术两者可以说是信息和艺术形式的正向和逆向转换方式。那么，其实 AI 设计和问答其实能体现一种对抗关系，一方面，用户通

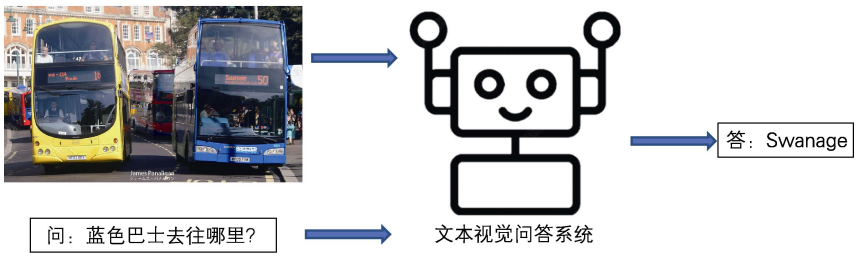


图 1 文本视觉问答任务

Fig.1 Text visual question answering task

过针对图片的问答可以得知，其是否关注了图片中设计师想要体现的关键信息，是否正确理解了当前图片的内容；另一方面，根据当前的图片场景设计，用户又可以得知，当前的场景是否是一个好的设计，是否顺利传达了用户所关注的信息内容。

其次，在文本视觉问答答案关键词预测环节，AI 设计需要根据人群及使用场景，对问题进行不同理解层次的区分和定义，从而使系统有针对性地对当前的问题难度进行区分。具体体现在，系统需要判断当前问题是否是一个所见即所得的问题，即答案是否都在图片中；还是说问题的解答需要有一定的知识背景，需要进行相关推理才能得到答案。

最后，在答案生成环节，单纯地将答案关键词反馈给用户显然不是最好的结果，而当前一些答案生成模型，一般只关注生成回答的合理性和正确性，却没有考虑过不同年龄层的认知差异，如标准的回答语句并不能回答儿童提出的相关问题。因此这里将需要一些基于不同年龄认知差异的 AI 设计，对不同年龄之间的认知差异进行建模和定义，从而辅助文本视觉问答模型基于认知差异重构答案。其中还可能需要加入一定解释性语句来辅助儿童充分理解问题和答案。

## 2　文本视觉问答技术概述

文本视觉问答是视觉问答的一个子任务，该任务中答案通常与图片中的文本相关，需要对图片中的文本进行检测和识别，并对识别文本进行筛选，或者基于识别文本生成回答。人为环境中的文本内容传达了重要的高级语义信息，这些信息是显式的，并且在场景中无法以其他任何形式替代。由于需要执行大多数日常任务，例如购物、使用公共交通工具、在城市中定位、预约或查询商店是否开业，所以在人造环境中解释这些书面信息至关重要。然而在当前的大型图像数据集中，如 MS Common Objects[3]数据集中，约有 50%的图像存在文本内容，并且在城市环境中，该比例急剧上升。因此，设计利用这些显式文本提示的模型至关重要。

在基于场景文本的 VQA 中，ST-VQA（场景文本视觉问题回答）[1]和 TextVQA[2]是两个具有代表性的竞赛。根据这两个竞赛已发布的技术报告可知，当前大多数方法将 OCR 文本集成到现有的 VQA 模型中，以回答基于场景文本的问题。LoRRA[4]使用基于注意力机制的多模态融合方法，分别用来融合图像特征和问题特征，以及 OCR 特征和问题特征。TextVQA 2019 的获胜团队使用了与 LoRRA 相同的框架，并应用了多模态分解高阶池化（MFH）[5]进行了多模态融合。在 ST-VQA 竞赛中，获胜者提出的 VTA[4]方法使用了 BERT[6]对问题和文本进行编码，并使用了 VQA 中自底向上和自顶向下[7]的经典多模态融合方法。Clova AI OCR[4]方法采用了 MAC 网络[8]用来融合视觉特征和用 BERT 编码的问题特征，并使用了指针网络选择匹配的答案。Gao 等人[9]提出了一种多模态图神经网络（Multi-Modal Graph Neural Networks，MM-GNN），以获得图像中多模态内容的更好表示，从而回答所提问题。Hu 等人[10]基于多模态 Transformer[11]框架提出了 M4C 模型，该模型可以融合多种输入模态，并且能同时在不同模态内和不同模态间进行动态交互。Kant 等人[12]在 M4C 的基础上，提出了空间感知的多模态 Transformer（SA-M4C），使每个视觉实体都通过只关注相邻实体实现动态交互。

## 3　基于 AI 设计的文本视觉问答模型与方法

将在一个通用的文本视觉问答模型的基础上，介绍 AI 设计下的文本视觉问答模型框架，本文所设计的文本视觉问答模型包括 4 个重要模块：（1）文本理解模块——机器阅读理解模型；（2）多模态关系挖掘模块——OCR 文本和物体之间的关系挖掘；（3）答案关键词预测模块；（4）回复生成模块。不同的关键模块包括不同的 AI 设计内容：利用场景设计理念促进不同目标之间的关系挖掘；基于关于问题理解层次差异的 AI 设计指导答案关键词预测；根据关于不同年龄认知差异的 AI 设计辅助最终回复的生成。AI 设计下的文本视觉问答模型框架见图 2。

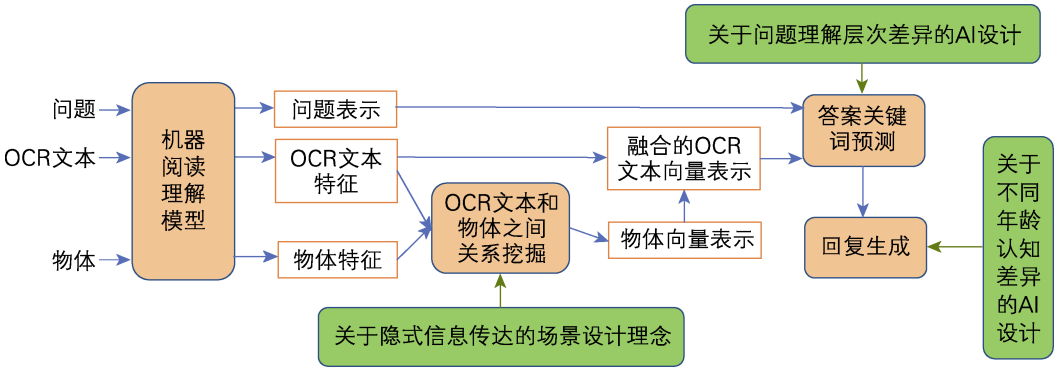


图 2　AI 设计下的文本视觉问答模型框架

Fig.2　Text visual question answering model frame based on AI design

### 3.1 关于隐式信息传达的场景设计理念

考虑到场景设计过程中，设计师会通过设计隐式地告诉用户信息的重要程度与对应关系。因此通过位置关系的挖掘，可以更充分地理解当前场景传达的隐式消息。因此用户在当前的文本视觉问答模型中，加入了基于位置的注意力机制，从而可以辅助模型学习到更好的向量表示。

### 3.2 关于问题理解层次差异的 AI 设计

基于同样的场景图片，回答不同的问题时可能需要不同的理解程度。有的问题是所见即所得的问题，比如“红色牌子上写的什么字”，那么模型只需要读取牌子上的文字，即可回答问题。然而有的问题可能需要进一步推理才能得到相应答案，比如“图片中是否有快餐店”，而从图中能直接获取的信息包括一个具有 Subway 标识的建筑，那么模型需要具有一定常识和推理能力，即知道 Subway 是一家快餐店的品牌，才能推理得到当前场景中有快餐店的结论。因此当前的文本视觉问答模型的答案关键词预测模块包含 2 个组件：答案匹配——用来回答所见即所得的问题；答案推理——用来回答需要根据当前文本进行推理，才能得到正确答案的问题。

### 3.3 关于不同年龄认知差异的 AI 设计

针对同样的场景，不同年龄层的理解能力存在一定的差异性，因此需要提出关于不同年龄认知差异的 AI 设计。也就是说，同样的问题，对于模型来说，只获取到回答问题的关键词还远远不够，如果要真正解决实际场景中的问答问题，还需要对用户认知层次进行建模，生成符合当前用户理解能力的回答。比如针对“图中的蓝色牌子是什么标识”的问题，获得的答案关键词是“Exit”，那么向视力障碍的成年人可能仅提供“Exit 标识”的回答就足以传达信息，而对于儿童来说，还需要解释“Exit 标识”代表的意思，即出口标识，用来指引出去的路线。

## 4 实验对比

### 4.1 实验目的

提出一个基于 AI 设计的文本视觉问答框架，内容包括 3 个部分：利用场景设计理念，促进不同目标之间的关系挖掘；基于问题理解层次差异的 AI 设计，指导答案关键词预测；根据不同年龄认知差异的 AI 设计，辅助最终回复的生成。为了验证该模型及各个模块的有效性，分别在当前权威的文本视觉问答数据集 ST-VQA 和 TextVQA 数据集上进行了对比实验。首先，为了证明每个 AI 设计模块的重要性，针对 3 个研究内容，分别设计了对应的消融实验来进行对比分析。接着，为了验证该模型的整体性能，分别在 ST-VQA 和 TextVQA 数据集上设计相应的实验，与权威方法进行对比分析。最后，通过分析说明 AI 设计在文本视觉问答具体应用中具有重要的指导作用，有利于 AI 的发展。

### 4.2 实验数据集

当前文本视觉问答任务的数据集包括：ST-VQA 和 TextVQA 数据集。ST-VQA 数据集包含来自场景文本的不同标准数据集的图像，例如 COCO-Text[3]、VizWiz[13]、ICDAR 2013[14]、ICDAR 2015[15]和 IIIT 场景文本检索[16]数据集，以及来自通用数据集的图像，例如 ImageNet[17]和 Visual Genome[18]，其中每个选定的图像至少包含两个文本实例。ST-VQA 包含约 23 000 张图像，每张图像最多包含 3 个问答对，并分为训练集（约 19 000 张图像和 26 000 个问答对）和测试集（每个任务约 3 000 张图像和 4 000 个问答对）。TextVQA 的训练和验证集是从 Open Images v3[19]数据集的训练集中收集的，而测试集是从 Open Images 的测试集中收集的。TextVQA 包含约 28 000 张图像，每张图像最多包含 2 个问答对，并分为训练集（约 22 000 张图像和 34 000 个问答对）、验证集（约 3 000 张图像和 5 000 个问答对）和测试集（约 3 000 张图像和 5 000 个问答对）。

### 4.3 评价方法

ST-VQA 的评价指标为 ANLS（平均标准化 Levenshtein 相似度）[12]：

$$ANLS=\frac{1}{N}\sum_{i=0}^{N}(\max_{j}s(a_{ij},o_{q_i})) \tag{1}$$

$$s(a_{ij},o_{q_i})=\begin{cases}1-NL(a_{ij},o_{q_i}), & if\ \ NL(a_{ij},o_{q_i})<\tau\\ 0, & if\ \ NL(a_{ij},o_{q_i})\geqslant\tau\end{cases} \tag{2}$$

其中：$N$ 是问题总数，$M$ 是每个问题的 $GT$ 答案总数，$a_{ij}$ 是标准答案，$o_{q_i}$ 是第 $i$ 个问题 $q_i$ 模型得到的答案，$NL(a_{ij},o_{q_i})$ 是字符串 $a_{ij}$ 和 $o_{q_i}$ 之间的归一化 Levenshtein 距离，且 $\tau$ =0.5。

TextVQA 的评价标准为：

$$Acc(ans)=\min\left\{\frac{\#humans\ that\ said\ ans}{3},1\right\} \tag{3}$$

### 4.4 实验结果

基于 ST-VQA 数据集的关系挖掘和答案推理效果评估见表 1，通过加入对场景设计的考虑，挖掘不

**表 1 基于 ST-VQA 数据集的关系挖掘和答案推理效果评估**

**Tab.1 Relationship mining and answer inference evaluation based on ST-VQA data sets**

| 模型 | 结果（ANLS） |
|---|---|
| 基础模型 | 0.287 7 |
| +关系挖掘 | 0.293 1 |
| ++答案推理 | 0.313 3 |

表 2　基于不同年龄认知差异的回复生成

**Tab.2　Response generation based on cognitive differences at different ages**

| 图片 | 问题 | 相应关键词 | 年龄层 | 生成的回复 |
|---|---|---|---|---|
|  | 图中是什么标志 | EXIT | 儿童 | 出口标识，用来指引出去的路线 |
| | | | 青少年 | 出口标志 |
| | | | 成年人 | EXIT 标志 |

表 3　基于 ST-VQA 数据集的各模型效果

**Tab.3　Model effects based on ST-VQA dataset**

| 模型 | 结果（ANLS） |
|---|---|
| USTB-TQA | 0.170 |
| Clova AI OCR | 0.215 |
| QAQ | 0.256 |
| MM-GNN | 0.207 |
| VTA | 0.282 |
| M4C | 0.462 |
| Ours | 0.313 |

表 4　基于 TextVQA 数据集的各模型效果

**Tab.4　Model effects based on TextVQA dataset**

| 模型 | 结果（Acc/%） |
|---|---|
| Image Only | 5.88 |
| Pythia | 14.01 |
| LoRRA | 27.63 |
| Schwail | 30.54 |
| MM-GNN | 31.10 |
| DCD ZJU (DCD) | 31.44 |
| MSFT VTI | 32.46 |
| M4C | 39.10 |
| Ours | 33.54 |

同目标之间的关系，可以将基础模型的性能从 0.287 7 提升至 0.293 1。同时，如果再基于问题理解层次差异的 AI 设计，修正答案关键词预测模块，除了能针对简单问题匹配答案以外，若再加入答案推理的组件，还能使最终结果从 0.293 1 提升至 0.313 3。由此可见，在设计文本视觉问答系统时，结合相应的 AI 设计可以使模型效果提升，并且更容易落地，普适性更好。最后针对不同年龄认知差异进行建模，并辅助回复生成，也能使 AI 产品更有效、更容易落地。然而因为当前数据集不支持不同年龄层的答案差异的评估，因此这里只简单地举例说明相应的生成答案，不进行量化比较。基于不同年龄认知差异的回复生成见表 2，该模型可以根据不同的年龄层生成不同的回复，从而提升整体的用户体验。

分别将当前方法在 ST-VQA 和 TextVQA 数据集上与其他方法进行了对比，可见用户的方法在不同的数据集上都有很好的效果，这为后续的 AI 应用提供了坚实的基础。基于 ST-VQA 数据集的各模型效果见表 3，基于 TextVQA 数据集的各模型效果见表 4。

## 5　结语

提出了一种 AI 设计下的文本视觉问答框架，并进行了实验分析和效果对比，旨在通过 AI 设计更好地指导当前 AI 模型的设计，提升模型效果和用户体验。以该设计和应用为基础，促进更多研究者对 AI 设计和 AI 技术的进一步融合的研究热情，这将更有利于更多的 AI 技术更好地投入应用，从而解决用户更多的日常生活需求，加快 AI 时代的到来。

**参考文献：**

[1] BITEN A F, TITO R, MAFLA A, et al. Scene Text Visual Question Answering[C]. Seoul: Proceedings of the IEEE International Conference on Computer Vision, 2019.

[2] SINGH A, NATARAJAN V, SHAH M, et al. Towards Vqa Models that Can Read[C]. Long Beach: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[3] VEIT A, MATERA T, NEUMANN L, et al. Coco-text: Dataset and Benchmark for Text Detection and Recognition in Natural Images[J]. 2016.

[4] BITEN A F, TITO R, MAFLA A, et al. Icdar 2019 Competition on Scene Text Visual Question Answering[C]. Sydney: 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019.

[5] YU Z, YU J, XIANG C, et al. Beyond Bilinear: Generalized Multimodal Factorized High-order Pooling for Visual Question Answering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(12): 5947-5959.

[6] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. Minneapolis: Proceedings

of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

[7] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering[C]. Salt Lake: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[8] HUDSON D A, MANNING C D. Compositional Attention Networks for Machine Reasoning[C]. Vancouver: Proceedings of the 6th International Conference on Learning Representations, 2018.

[9] GAO D, LI K, WANG R, et al. Multi-Modal Graph Neural Network for Joint Reasoning on Vision and Scene Text[C]. Seattle: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[10] HU R, SINGH A, DARRELL T, et al. Iterative Answer Prediction with Pointer-augmented Multimodal Transformers for Textvqa[C]. Seattle: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need[C]. Long Beach: Advances in Neural Information Processing Systems, 2017.

[12] KANT Y, BATRA D, ANDERSON P, et al. Spatially Aware Multimodal Transformers for TextVQA[C]. Glasgow: Proceedings of the European Conference on Computer Vision, 2020.

[13] GURARI D, LI Q, STANGL A J, et al. Vizwiz Grand Challenge: Answering Visual Questions from Blind People[C]. Salt Lake City: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[14] KARATZAS D, SHAFAIT F, UCHIDA S, et al. ICDAR 2013 Robust Reading Competition[C]. Washington: 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013.

[15] KARATZAS D, GOMEZ-BIGORDA L, NICOLAOU A, et al. ICDAR 2015 Competition on Robust Reading[C]. Nancy: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2015.

[16] MISHRA A, ALAHARI K, JAWAHAR C V. Image Retrieval Using Textual Cues[C]. Sydney: Proceedings of the IEEE International Conference on Computer Vision, 2013.

[17] DENG J, DONG W, SOCHER R, et al. Imagenet: a Large-scale Hierarchical Image Database[C]. Miami: 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[18] KRISHNA R, ZHU Y, GROTH O, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73.

[19] KRASIN I, DUERIG T, ALLDRIN N, et al. Openimages: a Public Dataset for Large-scale Multi-label and Multi-class Image Classification[EB/OL]. (2017-01-12) [2020-12-01]. https://github. com/openimages, 2017.