E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP3　　**文献标识码**:A　　**文章编号**:1006-8961(2023)06-1608-22

论文引用格式:Liu H F, Chen J J, Li L, Bao B K, Li Z C, Liu J Y and Nie L Q. 2023. Cross-modal representation learning and generation. Journal of Image and Graphics, 28(06):1608-1629(刘华峰, 陈静静, 李亮, 鲍秉坤, 李泽超, 刘家瑛, 聂礼强. 2023. 跨模态表征与生成技术. 中国图象图形学报, 28(06):1608-1629)[DOI:10.11834/jig.230035]

跨模态表征与生成技术

刘华峰[1],陈静静[2],李亮[3],鲍秉坤[4],李泽超[1],刘家瑛[5],聂礼强[6*]

1. 南京理工大学计算机科学与工程学院,南京　210094; 2. 复旦大学计算机科学技术学院,上海　200438;
3. 中国科学院计算技术研究所,北京　100190; 4. 南京邮电大学通信与信息工程学院,南京　230001;
5. 北京大学王选计算机研究所,北京　100871; 6. 哈尔滨工业大学(深圳)计算机科学与技术学院,深圳　518055

摘　要:多媒体数据持续呈现爆发式增长并显现出异源异构的特性,因此跨模态学习领域研究逐渐引起学术和工业界的关注。跨模态表征与生成是跨模态学习的两大核心基础问题。跨模态表征旨在利用多种模态之间的互补性剔除模态之间的冗余,从而获得更为有效的特征表示;跨模态生成则是基于模态之间的语义一致性,实现不同模态数据形式上的相互转换,有助于提高不同模态间的迁移能力。本文系统地分析了国际与国内近年来跨模态表征与生成领域的重要研究进展,包括传统跨模态表征学习、多模态大模型表示学习、图像到文本的跨模态转换和跨模态图像生成。其中,传统跨模态表征学习探讨了跨模态统一表征和跨模态协同表征,多模态大模型表示学习探讨了基于Transformer的模型研究,图像到文本的跨模态转换探讨了图像视频的语义描述、视频字幕语义分析和视觉问答等领域的发展,跨模态图像生成从不同模态信息的跨模态联合表示方法、图像的跨模态生成技术和基于预训练的特定域图像生成阐述了跨模态生成方面的进展。本文详细综述了上述各个子领域研究的挑战性,对比了国内外研究方面的进展情况,梳理了发展脉络和学术研究的前沿动态。最后,根据上述分析展望了跨模态表征与生成的发展趋势和突破口。

关键词:多媒体技术;跨模态学习;大模型;跨模态表征;跨模态生成;深度学习

Cross-modal representation learning and generation

Liu Huafeng[1], Chen Jingjing[2], Li Liang[3], Bao Bingkun[4], Li Zechao[1], Liu Jiaying[5], Nie Liqiang[6*]

1. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing　210094, China;
2. School of Computer Science, Fudan University, Shanghai　200438, China; 3. Institute of Computing Technology, Chinese Academy of Sciences, Beijing　100190, China; 4. College of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing　230001, China; 5. Wangxuan Institute of Computer Technology, Peking University, Beijing　100871, China; 6. School of Computer Science of Technology, Harbin Institute of Technology (Shenzhen), Shenzhen　518055, China

Abstract: Nowadays, with the booming of multimedia data, the character of multi-source and multi-modality of data has become a challenging problem in multimedia research. Its representation and generation can be as two key factors in cross-modal learning research. Cross-modal representation studies feature learning and information integration methods using

收稿日期:2023-01-15;**修回日期**:2023-02-17;**预印本日期**:2023-02-24

*通信作者:聂礼强　nieliqiang@gmail.com

基金项目:江苏省自然科学基金项目(BK20220936);中国博士后科学基金项目(2022M721626)

Supported by: Natural Science Foundation of Jiangsu Province, China (BK20220936); China Postdoctoral Science Foundation Funded Project (2022M721626)

multi-modal data. To get more effective feature representation, multimodality-between mutual benefits are required to be strengthened. Cross-modal generation is focused on the knowledge transfer mechanism across modalities. The modals-between semantic consistency can be used to realize data-interchangeable profiles of different modals. It is beneficial to improve modalities-between migrating ability. The literature review in cross-modal representation and generation are critically analyzed on the aspect of 1) traditional cross-modal representation learning, 2) big model for cross-modal representation learning, 3) image-to-text cross-modal conversion, joint representation, and 4) cross-modal image generation. Traditional cross-modal representation has two categories: joint representation and coordinated representation. Joint representation can yield multiple single-modal information to the joint representation space when each of single-modal information is processed through the coordinated representations, and cross-modal representations can be learnt mutually in terms of similarity constraints. Deep neural networks (DNNs) based self-supervised learning ability are activated to deal with large-scale unlabeled data, especially for the Transformer-based methods. To enrich the supervised learning paradigm, the pre-trained large models can yield large-scale unlabeled data to learn training, and a downstream tasks-derived small amount of labeled data is used for model fine-tuning. The pre-trained model has better versatility and transfering ability compared to the trained model for specific tasks, and the fine-tuned model can be used to optimize downstream tasks as well. The developmentof cross-modal synthesis (a.k.a. image caption or video caption) methods have been summarized, including end-to-end, semantic-based, and stylize-based methods. In addition, current situation of cross-modal conversion between image and text has beenanalyzed, including image caption, video caption, and visual question answering. The cross-modal generation methods are summarized as well in relevance to the joint representation of cross-modal information, image generation, text-image cross-modal generation, and cross-modal generation based on pre-trained models. In recent years, generative adversarial networks (GANs) and denoising diffusion probabilistic models (DDPMs) have been faciliating in cross-modal generation tasks. Thanks to the strong adaptability and generation ability of DDPM models, cross-modal generation research can be developed and the constraints of vulnerable textures are optimized to a certain extent. The growth of GAN-based and DDPM-based methods are summarized and analyzed further.

Key words: multimedia technology; cross-modal learning; foundation model; cross-modal representation; cross-modal generation; deep learning

0 引　言

随着视频、图像和文本等不同类型媒体数据的日益增长，旨在挖掘、分析和处理多源异构数据的跨模态学习逐渐引起人们关注，而跨模态表征与生成作为跨模态学习的基本任务更是研究热点。跨模态表征旨在利用多种模态之间的互补性，剔除模态之间的冗余性，从而获得更为有效的特征表示；跨模态生成则是基于模态之间的语义一致性，实现不同模态数据形式上的相互转换，有助于提高不同模态间的迁移能力。

跨模态表征与生成的起源可追溯至20世纪30年代。1935年Hotelling提出一种用途广泛的统计学分析算法——典型相关性分析(canonical- correlation analysis，CCA)，并由Cooley和Lohnes推动了其发展。在跨模态表征中，CCA广泛地应用于度量两种模态信息之间的相关特征，并在计算中尽可能保持这种相关性。1998年，Blum和Mitchel在多模态数据处理领域提出了协同训练的学习方法，使得分类器可从不同数据源中学习到尽可能多的知识。

21世纪初，研究人员提出了多核学习方法。该方法可以为不同模态数据选取不同的核函数，而且可采用特定方法对不同核函数进行融合，能够较好地处理异构数据的分类和识别问题。考虑到多源异构数据在高层语义空间中具有较强的相关性，而底层的特征表示往往具有较大差异，因此，研究人员提出了共享子空间学习方法。该方法能够对多源数据的相关关系进行挖掘，得到跨模态特征的一致性表示。共享子空间学习方法的出现极大推动了跨模态表征与生成领域的发展。

Ngiam等人(2011)构建了以玻尔兹曼机为基本单元的深度学习模型，实现了对视频和音频等数据的联合表征，将跨模态表征与生成推至快速发展时期。Goodfellow等人(2014)提出了生成对抗网络(generative adversarial network，GAN)，其由互为博弈

的生成器和判别器构成，通过对抗训练不断进行迭代优化。至此，GAN成为跨模态生成的主流方法。同时，伴随着算力和数据规模的提升，多模态预训练模型凭借强大的跨模态表征能力成为研究主导，突破了已有模型结构的精度限制。

得益于深度学习技术的发展和硬件算力的不断提升，以DALL-E为代表的自回归模型问世，通过离散化图像和使用Transformer模型在千万级图文对数据上进行自回归学习，使得生成图像在真实性和语义一致性上有了飞跃式提升。在随后一年中，扩散模型的出现更是带起了一股人工智能艺术创作的热潮，其不仅能够控制迭代次数和生成时间，还能增加生成图像的多样性，为跨模态生成带来了新的发展机遇。

随着当今智能化与信息化时代的来临，跨模态数据呈现爆炸式增长。跨模态学习技术在各行各业蓬勃发展，是智慧城市、智慧家居等领域的核心技术，也是智能交通、智慧医疗等行业应用不可或缺的关键模块。2021年，工信部等部门联合发布《物联网新型基础设施建设三年行动计划（2021—2023年）》，提出要加快多模态生物识别、互联互通和空中下载等技术与家电、照明、门锁和家庭网关等产品的融合应用，首次将跨模态列为物联网新基建在民生消费领域的重点任务。跨模态学习技术符合国家科技发展规划，有助于促进产业转型与升级，推动信息产业化步伐。

模态是指特定类型的信息或信息存储的表示形式，例如文本、图像、音频和视频。跨模态内容通常是多个耦合模态的共同体，相关研究的技术基础是跨模态表征与生成。首先，海量跨模态数据广泛存在，模态间的关联关系复杂多样，精准的跨模态表征是有效使用跨模态数据的关键。另外，跨模态内容之间存在较大的语义鸿沟，为增强跨模态深度语义关联理解，从复杂跨模态内容中获取全面、深入的语义信息，开展跨模态生成研究是必要的。

随着人工智能技术的快速发展，跨模态学习成为重要的研究方向，跨模态表征与生成的发展呈现百花齐放、各有千秋的趋势。在跨模态表征中，预训练模型基于强大的表征能力，成为突破现有模型结构精度局限的有效手段；自监督学习通过挖掘无标签数据自身的表征特性，为缓解大模型预训练和跨模态标注数据稀缺之间的矛盾提供了突破点；多模态融合表征通过挖掘不同模态信息之间的冗余性和互补性，为多模态信息寻找好的特征空间。在跨模态生成方面，生成对抗网络通过对抗训练迭代优化生成器和判别器，推动了跨模态生成任务的发展；自回归模型通过开展自回归学习，使生成图像在真实性和语义一致性上有了飞跃式提升；扩散模型通过多步映射不断将与目标数据大小一致的噪声转换为图像，不仅可以控制迭代次数和生成时间，还能增加生成图像的多样性，将跨模态生成的研究推向了全新的维度。

面向跨模态领域的发展需求，本文从跨模态表征和生成两个基础关键技术着手进行研究，归纳其发展现状和最新前沿动态，分析展望跨模态表征与生成的发展趋势和突破口，对推动相关技术进步及跨模态领域的持续发展起到积极的作用。

1 国际研究现状

1.1 传统跨模态表征学习

跨模态表征学习是跨模态机器学习中的一个关键研究方向。表征学习旨在去除原始数据中冗余的部分，提取出有效信息并产生对应的特征。相比单模态表征学习，跨模态表征学习面临更多的挑战，例如模态之间的信息融合、多模态噪声处理以及不同模态信息互补等。传统的跨模态表征学习的相关研究主要可以分为统一表征和协同表征两种类型（Baltrušaitis等，2019）。统一表征将多个单模态信息映射到统一表征空间并进行学习；而协同表征对单模态信息分别处理，通过相似性约束使跨模态表征能够协同学习。本文从以上两个方向对传统跨模态表征学习现状进行介绍。

1.1.1 跨模态统一表征学习

1）基于神经网络的表征学习。Ngiam等人（2011）提出通过双模态自动编码器学习跨模态的共享表征。该方法将视频和音频编码器最后一层隐藏层表示进行拼接，使其作为自动编码器的输入进行跨模态的表征学习。通过共享的表示学习层，自动编码器模型能够对两种模态的特征进行协同学习，在给定其中一种模态输入数据的条件下，能够从中复原另一模态的对应数据。Silberer和Lapata（2014）在此基础上提出通过半监督学习目标训练层叠多模态自动编码器，对标注匹配语义表征学习进行求解。

该方法在文本和视觉模态的共享表示学习层上增加了归一化指数函数，从而能够更好地适应任务中的目标标注形式。除此之外，Silberer 和 Lapata（2014）提出一种灵活的半监督损失函数，能够帮助模型学习到更具区分度的模态表征，进而适应目标分类等任务。上述跨模态表征学习方法对不同模态设计不同的网络结构分别学习，并通过共享的表示层或归一化指数函数进行表征学习，Rastegar 等人（2016）提出 MDL-CW（multimodal deep learning framework with cross）方法，通过对不同模态的编码器结构进行跨模态权重学习，使编码器能够逐渐学习到跨模态的信息交互，通过理论分析得出自底向上的跨模态权重学习能够产生更多模态间的交互信息，并通过实验结果证明了模型的跨模态学习能力。

2）基于概率图模型的表征学习。基于概率图模型的跨模态表征学习研究以潜在随机变量对表征学习任务进行建模，通过给定数据对随机变量在联合空间中的概率分布进行构建。Hinton 等人（2006）以受限玻尔兹曼机（Boltzmann machines）作为基础模块，构建了对比散度的受限玻尔兹曼机模型（restricted Boltzmann machines，RBM）。受限玻尔兹曼机模块与神经网络类似，依次连接的受限玻尔兹曼机也对模态语义进行逐层抽象，进而学习到多级的语义信息。玻尔兹曼机通过概率分布对表征进行建模，因此在训练过程中不需要有标注数据。Srivastava 和 Salakhutdinov（2012）提出基于多模态深度信念网络的深度玻尔兹曼机，通过合并不同模态无向图构建多模态信息的联合概率分布。Kim 等人（2013）在此基础上对不同模态分别设计深度信念网络，进而组合获得统一表征。这类基于概率图模型的主要优势在于其具备生成能力，因此即使在一个或多个模态信息缺失的条件下，模型依然能够有较好表现。同时基于概率图的表征学习模型能够通过一种模态输入生成另一模态的样本。

1.1.2 跨模态协同表征学习

不同于将多种模态表征映射到统一表征空间，跨模态协同表征学习对不同模态分别进行表征学习，通过相似性约束对不同模态表征进行协同学习。跨模态协同表征学习主要适用于不同模态耦合度低的任务，如跨模态检索。该类研究主要可以分为基于特征相似约束的表征学习和基于结构相似约束的表征学习。

1）基于特征相似约束的表征学习。基于特征相似约束的方法通过最小化不同模态特征在联合空间中的距离对多种模态的表征学习进行约束。Weston 等人（2010）提出基于图像特征嵌入的标签分类方法 WSABIE（web scale annotation by image embedding），通过构建视频特征和标注特征的联合空间，使用线性函数对不同模态特征进行映射并最大化关联样本的内积，从而获得图像对应的标签。谷歌提出基于神经网络的深度视觉语义嵌入模型 DeViSE（deep visual sementic embedding），在保留 WSABIE 方法中的联合空间内积相似度以及基于排序的损失函数的条件下，使用更为复杂的基于神经网络的视觉文本编码模块（Frome 等，2013）。Kiros 等人（2014）通过长短期记忆网络（long short-term memory，LSTM）编码模块对文本上下文进行更好的表征学习，同时设计了正样本和负样本的损失函数。

2）基于结构相似约束的表征学习。基于结构的相似约束在上述特征语义约束的基础上，根据不同任务对表征学习其余方面提出更强的约束条件。基于结构相似约束的表征学习主要用于跨模态哈希任务，该任务需要将高维模态特征映射到低维二进制表征，除了需要满足不同模态关联样本在二进制表征空间相似度高的条件，还要求表征满足指定大小的二进制码约束。Jiang 和 Li（2017）提出在图像和文本描述中通过可学习的深度神经网络结构对二进制表征进行编码。Cao 等人（2016）在其基础上提出更复杂的 LSTM 编码模块。

另一个相关的任务是典型相关性分析（CCA）。在跨模态任务中，基于典型相关分析的方法通过映射函数最大化不同模态之间的相关性，从而得到跨模态关联的模态表征。除此之外，映射函数还需要满足映射后的随机变量之间正交的约束。Lai 和 Fyfe（2000）以及 Andrew 等人（2013）提出将传统典型相关分析中的线性映射替换成更为复杂的非线性映射，前者提出基于核方法的典型相关分析，后者则针对基于核方法的典型相关分析中伸缩性差的问题提出另一种非线性映射函数。

1.2 多模态大模型表征学习

预训练大模型现已成为全球人工智能领域瞩目的焦点。与此前常用的有监督学习范式不同，预训练大模型能够充分利用大规模的无标注数据来学习训练，并使用下游任务的少量有标注数据进行模型

微调。与直接训练具体任务的模型相比，预训练模型具有更好的通用性与迁移能力，在其基础上微调得到的模型在各种下游任务上均取得了显著性能提升。随着单模态预训练模型的快速发展，多模态大模型受到学术界和工业界的广泛关注，旨在将文本、语音、图像和视频等多模态内容联合起来进行学习，专注多模态内容之间的关联特性及跨模态转换问题，由此实现多模态数据从浅层语义到通用知识的跨越。按照模型结构类型，当前的多模态预训练模型可分为3类，即基于Transformer编码器的模型结构、基于Transformer解码器的模型结构和联合Transformer编码器与解码器的模型结构。

1.2.1 基于Transformer编码器

最早的多模态预训练模型方法大多是基于Transformer编码器的方法，根据网络结构又分为单流模型与双流模型，二者主要区别在于，在单流模型中不同模态的数据直接同时输入到Transformer编码器中，从底层开始进行多模态的交互；而双流模型中每一个模态的数据需要首先输入到该模态对应的编码器，然后在高层通过跨模态编码器实现模态间的交互。

1）单流模型。经典的单流模型通过预训练一个统一的Transformer来捕获不同模态和跨模态关系的元素。Li等人（2019）提出了VisualBert（visual bidirectional encoder representation from transformers）模型，在结构上使用了堆叠的Transformer编码器，利用自监督学习机制对文本和图像信息进行对齐和融合，其视觉输入是Faster-RCNN（faster region convolutional neural network）（Ren等，2015）提取的图像区域特征和相应的位置编码，而语言输入是原始词嵌入。该方法设置了两个预训练任务，一是与BERT（bidirectional encoder representation from transformers）一样的掩蔽语言建模，二是句子图像预测，即判断输入的句子是否为对应图像的描述。Chen等人（2020b）提出图像—文本表征学习模型UNITER（universal image-text representation learning），在文字和图像区域之间添加一个匹配模块，进一步建立图像和文本之间的细粒度关联，并据此设计了掩蔽语言建模、图像—文本匹配和掩蔽图像区域建模3个预训练任务。Kim等人（2021b）提出ViLT（vision and language transformer）模型，使用预训练的ViT（vision transformer）来初始化Transformer，从而可以直接使用交互层来处理视觉特征，而无需添加额外的视觉编码器。此外，ViLT还提出了全词掩码技术，即掩蔽连续子词标记的技术，避免仅通过词上下文进行预测。Sun等人（2019）提出VideoBert（video BERT），将BERT扩展到大规模视频—语言表征学习，为了对应文本中的标记，继续原BERT中的掩蔽语言建模任务，使用K均值聚类对所有提取的视频特征向量进行聚类，并以类中心作为视觉标记，每个视频特征向量由其所属的类中心表示。

2）双流模型。Lu等人（2019）提出了ViLBERT（vision and language BERT），将BERT扩展为双流模型，该模型由两个并行网络组成，分别用于视觉和语言处理，其后是共同注意力转换器层。预训练任务分为重建任务和匹配任务。重建任务包含文本输入的掩蔽语言建模和图像的掩蔽区域建模；匹配任务是预测图像—文本对是否匹配，即文本是否描述图像。Tan和Bansal（2019）在ViLBERT的基础上增加了对象关系编码器，并提出了额外的预训练任务，即基于兴趣区域的特征回归和图像问答。经典的CLIP（contrastive language image pretraining）模型也采用双流架构（Radford等，2021），联合训练图像编码器和文本编码器来预测一批图像—文本训练样本的正确配对。通过使用从互联网收集的4亿个图像文本对进行训练，CLIP的零样本性能可与许多数据集上的有监督方法相媲美。Jia等人（2021）提出了ALIGN（a large scale image and noisy text embedding），使用对比损失训练了一个简单的双编码器模型，利用包含超过10亿个噪声图像—文本对的数据集来扩展视觉和视觉语言表征学习，其预训练得到的视觉和视觉语言表示在广泛的任务上实现了非常强大的性能。如上所述，CLIP和ALIGN主要关注将图像和文本映射到跨模态的共享嵌入空间。而Florence（Yuan等，2021）则着重于如何使模型适应各种下游任务，并设计了一个由多模态预训练模型和适应模型组成的工作流。对于任务适应，使用动态头部适配器将学习到的视觉特征表示从场景扩展到对象，提出了CoSwin适配器来学习视频表示，并使用METER（multimodal end-to-end transformer）适配器将模型应用到依赖细粒度视觉—语言表示的视觉语言任务。

1.2.2 基于Transformer解码器

GPT-3（generative pretrain transformer）（Brown

等，2020）是一种典型的基于Transformer解码器的预训练模型，在各种文本生成任务中表现出优异的性能。基于Transformer解码器，Ramesh等人（2021）提出一种文本到图像生成模型DALL-E，该模型在4亿个图像—文本对上进行训练。通过结合VQVAE（vector quantisation variational auto encoder）（van den Oord等，2017）和GPT（Brown等，2020）可以生成对应图像，称为图像版GPT。同时，DALL-E有能力对生成的图像中的物体进行操作和重新排列，能创造出一些根本不存在的东西。虽然DALL-E在一定程度上提供了对少量物体属性和位置的可控性，但成功率取决于文字的措辞。当引入更多的对象时，DALL-E容易混淆对象及其颜色之间的关联，成功率会急剧下降。Wang等人（2022）设计并训练了一个生成式图像到文本转换器GIT（generative image-to-text transformer），以统一图像/视频描述和问答等视觉语言任务。GIT模型包含一个图像编码器和一个文本解码器。图像编码器部分是一个类似SWIN（shifted window）的视觉Transformer，它基于对比学习任务在大量图像—文本对进行预训练。而文本解码器部分则在视觉部分编码的基础上，用自回归的方法来生成文本。该模型在图像/视频描述、问答等多项任务上都取得了良好的性能。

1.2.3　联合Transformer编码器与解码器

Transformer编码器通过双向的注意力机制来学习对数据的理解能力，而解码器通过单向的注意力机制学习生成能力。为了使模型同时具备这两种能力，从而可以在更广泛的下游任务上应用，一些工作联合Transformer编码器与解码器进行多模态预训练，取得了不错的效果。Cho等人（2021）提出VL-T5（vision language tasks）模型，将多个多模态任务统一为文本生成。具体地，该模型由Transformer编码器和自回归的解码器组成，主要创新点在于针对训练任务与数据的不同采用不同的输入文本与输出文本的构造方式，这种将模型结构和目标任务统一的方法可以充分利用不同任务的数据来训练模型，提高模型的泛化性。Zhou等人（2020）提出了Unified VLP（unified vision language pretrain），编码器和解码器共享同一个Transformer网络。该方法通过设置注意力掩码来控制网络为编码器或解码器。具体地，当使用编码器时，注意力掩码为双向掩码，任意位置都可建模前后两个方向的依赖关系；当使用解码器功能时，注意力掩码设置为单向，每一位置只能建模前文的依赖关系。这种编解码共享的方式能够减少参数量，使网络更加简洁。

1.3　图像到文本的跨模态转换

1.3.1　图像语义描述

多种图像语义描述算法主要分为3类，即基于端到端的方法、基于语义概念的方法和基于风格化的方法。本节从以上3个角度介绍国际上图像语义描述的研究现状。

1）基于端到端的方法。端到端方法在图像语义描述生成任务上得到了广泛的应用，该方法采用编码器—解码器结构（Cho等，2014），编码器负责提取图像特征，解码器负责描述文本的生成。该技术最早用于自然语言处理领域的翻译任务，而图像语义描述生成可以视做一个不同模态间的“翻译”任务。图像编码器与文本解码器最早分别采用卷积神经网络CNN和循环神经网络（recurrent neural network，RNN）（Wang等，2016）。在近年的工作中，注意力机制的变形被大量使用。如Zhang等人（2021）利用注意力机制隐式地探索图像区域之间的视觉关系，从而在文本描述词和视觉区域之间提供良好的对齐。

2）基于语义概念的方法。基于语义概念的图像语义描述方法旨在训练神经网络中的隐状态，学习图像中的具有重要语义的对象（概念），辅助解码器生成细化且连贯的文本描述。Nguyen等人（2021）利用场景图标签进行竞争性图像语义描述生成，其基本思想是减少从输入图像获得的图形与其描述之间的语义差距。

3）基于风格化的方法。图像语义描述的另一个热门研究方向是通过控制图像语义描述的风格生成更具表现力和吸引力的文本描述。该方向因其在现实场景中的潜在应用价值而被工业界所重视。例如，当人们在社交媒体平台上传照片时，往往需要一个吸引人的、风格化的标题，而这是传统的事实性图像语义描述模型难以做到的。Li和Harrison（2021）使用生成的风格向量融合图像区域的局部语义以及全局上下文元素，生成更有吸引力的描述。Li和Harrison（2022）为评估模型的风格化描述生成能力设计了两种新的自动化指标。一种在没有标注参考答案的情况下评估生成的描述捕获给定样式的程度；另一种在流行的方法基础上加入偏置以强调风格化词语，从而更好地衡量描述风格化的质量。

1.3.2 视频语义描述

在图像语义描述的基础上，视频语义描述任务扩展到了针对多帧时空角度连续的图像的语义描述之上，在融合多模态表征的同时，还要兼顾帧与帧之间的关联性，从序列的角度去建模视觉特征并与语言特征进行交互。

国际上的前沿研究基本上是以图像、视频编码模型的发展以及序列建模方式的发展为基本脉络的。Venugopalan等人(2015)在设计的S2VT(sequence to sequence video to text)模型中，首次提出先使用深度卷积网络对视频的图像帧特征和光流帧特征进行双路提取，再分别送入RNN网络中，完成两路解耦编码，同时在解码阶段将两路编码进行融合，再使用RNN进行解码。这个方法为跨模态端到端学习提供了一个最初的解决思路。

随着计算机视觉和自然语言的技术发展，越来越多的学者不局限于视频、语言表征的全局编码、交互，而是使用一些前沿技术对视频语义特征进行细粒度编码，并相应地划分出对应短时间片段内的事件信息进行特征的精确融合编码。得益于Johnson等人(2016)提出的全卷积网络以及Ren等人(2015)提出的Faster R-CNN架构，学者们有了很多强有力的特征提取架构将视频内的图像帧打散为局部的密集语义区域，并使用注意力机制进行细粒度语言和视觉信息的关联交互，以得到更加鲁棒准确的融合表征。而Transformer的引入(Vaswani等，2017)极大程度上解决了序列模型的遗忘等痛点问题，对视频到语言模型的性能改善起到了质的作用。

总之，更好的视频和语言表示提取、更丰富的模态交互、更高效准确的时序建模是视频语义描述的关键，对这些问题，国际上有很多非常好的基础性探索。

1.3.3 视频字幕语义分析

随着多媒体的迅速发展，每天都有大量的多模态视频(带有音频和/或文本)发布在网络中。纯粹的视频语义描述任务只是对视觉内容进行简单的语义描述，而在现实应用中，视频通常与其他形式相关联，例如电影或电视节目的字幕以及现场观众的收音等，这些不同的模态通常涉及人们之间丰富的社交互动，包括活动和对话。

目前，在多媒体领域已经提出了多种基于电影、卡通和电视节目构建的多模态数据集。例如Hendricks等人(2017)提出的DiDeMo数据集、Krishna等人(2017)提出的ActivityNet Captions数据集以及Gao等人(2017)提出的CharadesSTA数据集，这些数据集使用单一的视频进行定位，并没有涉及字幕等复杂语义信息。由于电视字幕可以提供一些隐含的但是非常有用的语义线索来解释演员的情绪和意图，因此，为了更好地从视频语料库中检索相关时刻，Lei等人(2020)提出了一项新的视频字幕语义描述任务，并提出了TVC(TV show caption)数据集以及多模态Transformer(multi-modality transformer, MMT)。MMT首先通过外观、动作和文本形式分别表示每个视频及其字幕。然后，模型直接将所有模态连接起来作为原始Transformer的输入以生成字幕。TVC数据集与从前的数据集有两点不同。1)从前的数据集将视频统一分块并让注释者选择一个(或多个)编写明确的描述。这种粗略的时间注释不能很好地与自然时刻对齐。在TVC中，为了更准确地捕捉重要时刻，注释者可以自由选择时间窗口；2)从前的数据集将为整个视频编写的段落转换为单独的查询语句。虽然注释者在段落中使用了时间连接词(例如first，then)以及代词，但这些词减弱了上下文之间的语义信息的关联性，使得单个句子并不适合作为检索查询。相比之下，TVC注释过程鼓励注释者单独编写查询语句，而不需要考虑段落的上下文信息。Li等人(2020c)也提出了基于字幕的视频文本匹配任务，并提出一种用于大规模视频和语言相结合的表征学习的新框架HERO(hierarchical encoder for video language omni representation pretraining)。该模型将外观和运动模态连接为视觉模态，然后通过交叉注意机制对视觉和文本模态之间的相互关系进行建模。

1.3.4 变化语义描述

变化语义描述算法用于定位和描述一个场景中的语义变化，主要分为基于像素差异的方法和基于表征差异的方法两类。本节从语义变化建模的角度介绍国际上变化语义描述的现状。

1)基于像素差异的算法。美国卡内基梅隆大学的Jhamtani和Berg-Kirkpatrick(2018)在2018年发布了一个来自监控场景的变化语义描述数据集。该数据集中的图像对从固定角度拍摄，有着良好的对齐关系。基于这个前提，提出了一个DDLA(different description with latent alignment)模型来计算图像对像素级别的差异，并将其送入模型完成变化语义描

述。事实上，除了语义变化，动态环境中的图像对间会出现无关变化的干扰。例如在视角变化下，两幅图像中的物体在外观和位置上会出现偏移，导致二者不能完全对齐。而基于像素差异的方法需要建立在两幅图像完全对齐的前提下，所以仍然不能适应变化语义描述的各种场景。

2）基于表征差异的算法。为了使该研究更符合动态环境的设定，美国加州大学伯克利分校的Park等人（2019）发布了一个包含轻微视角变化的数据集。随后，韩国首尔大学的Kim等人（2021a）发布了一个包含极端视角变化的数据集。在上述两个数据集中，图像对间存在两种设定。一是同时存在语义变化和视角变化；二是仅存在视角变化。在差异建模的时候，相关研究工作主要利用基于图像对的特征表征进行建模。Park等人（2019）提出了一个DUDA（dual dynamic attention model）模型。首先利用预训练的CNN提取两幅图像的特征表征；然后利用作差的方法计算出两个表征间的差异表征；最后利用注意力模型和LSTM网络将差异表征转化成文本描述。然而，由于视角的改变，两幅图像的表征在外观和位置上存在轻微的偏移。因此，直接作差的方法导致建模的差异表征存在一定的噪声。为了在视角变化中区分和描述语义变化，新加坡南洋理工大学的Shi等人（2020）提出一个M-VAM（mirrored viewpoint-adapted matching）模型，通过语义相似度的方法首先预测出两幅图像中相似的特征作为未变化特征，进而求出变化特征。随后，基于相似度的范式被韩国首尔大学的Kim等人（2021a）和日本产业技术综合研究所的Qiu等人（2021）的研究团队所沿用。此外，加拿大曼尼托巴大学和华为公司研究团队利用循环一致性模型来提升图像对和语义描述的语义一致性（Hosseinzadeh和Wang，2021）。

1.3.5　视觉问答

随着注意力机制在自然语言处理领域的流行，国际上对于视觉问答模型的研究主要集中在以注意力机制为基础的多模态融合模型上，主要分为基于共同注意力的方法、基于检测注意力的方法和基于关系注意力的方法。本节从注意力机制的角度介绍国际上视觉问答的研究现状。

1）基于共同注意力的方法。共同注意力模型是对称的，通过视觉特征可以引导产生问题的注意力，文本特征可以引导产生图像的注意力。Lu等人（2016）构建了一个层次结构，分别在单词层面、短语层面和句子层面构建共同注意力，提出了平行共同注意力和可选共同注意力两种构建方式。局限在于只学习了多模态实例的粗糙交互，而所学习的共同注意力不能推断出每个图像区域和每个问题词之间的相关性。

2）基于检测注意力的方法。此前的图像注意力是基于卷积神经网络特征，相当于将图像均等分割成若干区域然后进行筛选，选择图像中前K个候选区作为视觉特征，通过提取图中多个对象作为输入视觉特征。基于检测注意力的方法将开放式注意力与检测注意力结合形成新的共同注意力，加强模型的表达能力。检测注意力作用受限于其检测类别的广度。

3）基于关系注意力的方法。Wu等人（2018）首次提出了关系注意力的概念。现有的大多数工作都集中在通过融合图像特征和文本特征来计算注意力分布，而不需要在不同图像对象之间进行比较。作为关注的主要属性，选择性取决于不同对象之间的比较。对象间的比较提供了更多信息，能够更好地分配注意力。

1.4　跨模态图像生成

1.4.1　不同模态信息的跨模态联合表示方法

同样语义的信息可能表现为不同模态的形式，例如文本和图像都可以表现一个人的外貌。为了达成跨模态图像生成的目标，首先需要设法对不同模态的信息的语义进行联合表示，以对跨模态生成提供约束和评价的标准。由于高层语义信息抽取这一问题的复杂性，目前的工作均基于深度神经网络搭建。现有方法的共同点在于均设法对不同模态的信息搭建了编码器神经网络，将原始模态的信息映射到隐空间中的向量上，以向量之间的余弦相似度建模信息之间的语义一致性。语义一致性越高的信息，它们的隐向量之间的余弦距离越小，反之亦然。

1）基于小规模特定领域跨模态信息对的联合表示方法。当待对齐的信息的语义集中在某个特定领域内时，可以采用针对单个小数据分布训练专用的跨模态联合表示模型。这些模型通常规模较小，易于训练，在特定的领域中有优秀的表现。

文本—图像跨模态生成开山之作GAN-INT-CLS（GAN-interpolation-conditional-latent-space）（Reed等，2016）中提出，将文本—图像联合表示的模块嵌

入GAN中的判别器中，将原生GAN以文本为条件改造为条件GAN（Mirza和Osindero，2014），以判别器的输出结果为跨模态语义对齐与否的标准。

GAN-INT-CLS中的判别器D可以理解为文本编码器φ和图像编码器的结合。文本编码器将文本抽象为特征后，直接将该特征拼接入图像编码器，随后再将拼接后的特征神经网络最终输出单个概率值，表述为$D(\hat{x}, \phi(t))$。它的训练方式与条件GAN的方式一致，对于那些不匹配的文本—图像对，也通过损失函数迫使判别器D输出接近0的值即可。

但是该结构的缺陷也是很明显的。它将文本、图像的编码器嵌入判别器，导致这两个编码器无法独立使用。所以事实上它的可扩展性非常有限。随着人们对于深度神经网络结构的进一步研究，自注意力机制在高层语义任务中取得了巨大的成功（Vaswani等，2017）。Devlin等人（2019）和Dosovitskiy等人（2021）基于自注意力机制的核心网络Transformer设计出了更加强大的跨模态联合表示模型。

AttnGAN（attention GAN）（Xu等，2018）中提出了深度注意力跨模态相似性模型（deep attentional multimodal similarity model，DAMSM），采用Transformer为对齐部分结构的基础。

AttnGAN中的文本编码器是基于长短时记忆网络（Hochreiter和Schmidhuber，1997）这一适用于处理序列信息的网络而搭建的，图像编码器则是采用传统的卷积神经网络搭建的。两个编码器分别将文本、图像各自编码为隐空间中的向量之后，对两个隐向量采用注意力机制（Vaswani等，2017）进行联合编码，给出它们的匹配分数，并且通过最大化匹配的文本—图像对的上述分数和最小化不匹配的文本—图像对的上述分数这一目标，训练文本及图像编码器。实验证明，DAMSM取得了优秀的结果。这一模型自从在AttnGAN中提出之后，广泛地应用在如Li等人（2020a）、Zhu等人（2019）、Qiao等人（2019）、Zhang等人（2017，2019）和Tao等人（2022）等多个文本—图像跨模态生成模型中，活力一直保持至今。

由于上述小规模模型的拟合能力有限，上述的跨模态联合表示模型主要应用在小规模的数据集上，数据需要分布在某个特定领域中。它们的优点是易于训练和易于部署，而缺点也十分明显。它们不能处理那些未在数据集中出现的数据，因此它们的应用范围是高度受限的。

2）基于对比学习的通用跨模态联合表示方法。基于小规模数据集训练的跨模态联合表示模型具有通用性不足的缺陷。为了解决这样的问题，有学者提出，构建足够庞大的跨模态数据集和足够有拟合能力的模型，之后采用对比学习的方式，从这个足够庞大的数据集中构建出各自模态下的编码器，使编码器有能力处理通用的跨模态数据。CLIP（contrastive language-image pre-training）（Radford等，2021）是基于这一方法的著名工作。它是一个文本—图像跨模态联合表示模型，基于一个爬取自互联网的超大规模文本—图像数据集，包含超过4亿对数据。

CLIP分别构建了一个文本编码器和一个图像编码器，在训练时对于单个批输入的N对文本—图像对，最大化相匹配的文本—图像对的隐向量的余弦距离，并最小化不匹配的文本—图像对的隐向量的余弦距离，其基本逻辑非常简单。然而，得益于大规模数据集中语义的丰富程度以及足够强大的计算力，CLIP最终取得的效果非常优秀，在无先验分类任务上取得了最佳性能。已有大量的工作基于CLIP展开，它的强大能力使得使用它充当跨模态语义对齐模型，构建下游任务成为了可能。

相应地，对于视频—文本跨模态对齐任务也已经有类似于CLIP的大规模工作。CLIP4CLIP将CLIP直接应用在连续的视频帧上，取得了优秀的视频检索结果。它通过将CLIP复用在时域上的方法，使得视频模态的信息也能由几乎同样的方式与图像、文本模态进行对齐。HD-VILA（high-resolution and diversified video-language pre-training）则是参考了CLIP的训练方式，收集了超大规模的视频—文本数据对，训练了相似的模型（Xue等，2022）。

为了节省计算资源，单个视频段采用了部分帧输入高分辨率图像、部分帧输入低分辨率图像的训练方式，有效利用了视频的帧间关联性，减少了冗余信息的输入。它提供的联合描述向量可以有丰富的下游应用。高层的如视频检索、视频编辑；低层的如视频超分等。这些都证明了基于对比学习的大规模模型具有强大的生命力。

目前，大部分跨模态联合表示模型都关注文本—图像或文本—视频这样的可由人类直接解读的模态的语义对齐。事实上，模态是一个非常广义的概念。例如，传统多媒体中的每一种媒体都可以成

为一种承载信息的模态。因此对跨模态的联合表示方法的研究还有很广阔的探索空间。

1.4.2　图像的跨模态生成技术

高质量图像的跨模态生成技术需要构建在前述的跨模态联合表示的基础上。跨模态联合表示为图像的跨模态生成提供了语义方面的约束以及定量的评价指标。现有的工作大致分为两类，一类基于预训练好的生成模型，设法将跨模态语义约束与预训练的生成模型的隐空间进行连接，以达到基于已有生成模型进行跨模态生成的目的；另一类从头训练一个新的生成模型，将跨模态语义约束设法加入训练时的损失函数，以达成直接训练一个跨模态生成模型的目的。

生成模型即是设法建模生成的图像落在真实图像数据集中的概率，并构建适当的神经网络，以最大似然作为目标函数，拟合该概率的模式。形式化为

$$\max p(G(\boldsymbol{z})) \tag{1}$$

式中，$p(\cdot)$是图像属于该数据集的概率，而$G(\cdot)$表示生成函数。常见的生成模型包括GAN、VAE（variational auto encoder）、Flow-model、DDPM（denoising diffusion probabilistic models）等。其中GAN和DDPM的应用最为广泛、取得的成果最为丰富。下面简述这两种生成模型，作为跨模态图像生成的基础。

GAN于2014年提出（Goodfellow等，2014），基于它的进一步研究和改进一直在持续。它的核心思路非常巧妙。既然一幅图像落在某个图像数据集中的概率不易直接建模，那么就直接使用一个深度神经网络充当判别器，用它来判断一幅图像落在该指定数据集中的概率。判别器的目的是对于那些来自于数据集的真实图像，给出尽可能接近1的输出，而对于那些虚假的图像给出尽可能接近0的输出。而生成器的目的则是尽可能生成符合数据集特征的图像，使得判别器无法成功地区分真实图像与虚假图像。在训练过程中，判别器和分类器的参数按照上述描述的目标依次更新。这个过程如同生成器和判别器在互相对抗，这也是其对抗生成模型得名的原因。

具体来讲，判别器和生成器的损失函数各自由交叉熵损失给出，具体为

$$L_D = \log(D(\boldsymbol{x})) - \log(1 - D(G(\boldsymbol{z}))) \tag{2}$$

$$L_G = -\log(D(G(\boldsymbol{z})) \tag{3}$$

式中，$\boldsymbol{x}$表示来自数据集的真实图像样本，$\boldsymbol{z}$表示隐空间中的向量，L表示损失函数。对于GAN的摸索以及对于GAN损失函数的探究从未停止，包括Mirza和Osindero（2014）、Arjovsky等人（2017）和Gulrajani等人（2017）的工作。GAN以及它的改进型已经取得了大量令人印象深刻的成果。

Sohl-Dickstein等人（2015）首先提出DDPM的思想，并在2020年发展完善。它的灵感来自于物理学中的扩散现象。具体来讲，首先试图通过向一幅图像中逐步加入高斯噪声的方式冲淡原有的图像，直至最终整幅图看起来几乎与一幅真正的高斯噪声图像没有区别，仿佛原本的图像扩散在高斯噪声之中，这个过程称为前向过程。随后，采用适当的算法，建模了前向过程的逆过程，借助深度学习的方法构建了一个从完全的高斯噪声图中逐步去噪，直至完全恢复到原始图像，这个过程称为逆向过程。上述两个过程也是其去噪扩散概率模型得名的原因。

前向过程可以形式化为

$$q\left(\boldsymbol{x}_t \middle| \boldsymbol{x}_{t-1}\right) = N\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\,\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right) \tag{4}$$

式中，β为预先指定的超参数，表示当前扩散步骤加入的高斯噪声的强度，N表示高斯分布，$\boldsymbol{I}$表示图像。由式（4）可得

$$q\left(\boldsymbol{x}_t \middle| \boldsymbol{x}_0\right) = N\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_{t-1}, (1-\bar{\alpha}_t)\boldsymbol{I}\right) \tag{5}$$

式中，$\alpha_t = 1 - \beta_t$和$\bar{\alpha}_t$均为超参数。

前向过程相对较容易形式化。难点在于如何形式化逆向过程。根据贝叶斯原理和马尔可夫性，可进行具体计算，即

$$q\left(\boldsymbol{x}_{t-1} \middle| \boldsymbol{x}_t, \boldsymbol{x}_0\right) = N\left(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right), \tilde{\beta}_t \boldsymbol{I}\right) \tag{6}$$

$$\tilde{\boldsymbol{\mu}}\left(\boldsymbol{x}_t, \boldsymbol{x}_0\right) = \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\,\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0 \tag{7}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{8}$$

式中，$\tilde{\mu}$表示高斯分布中的均值，$\tilde{\beta}$表示方差。

由此可见，如果能够通过一个深度神经网络，从带噪图像样本$\boldsymbol{x}_t$中恢复出原始图像样本$\boldsymbol{x}_0$，那么就可以按上式实现逆向过程的采样，最终迭代地得到原始图像样本$\boldsymbol{x}_0$。而这两幅图像之间的差异，正是一个噪声z_t。于是，DDPM的核心在于训练一个预测噪声的神经网络，具体为

$$\hat{z}_t = z_\theta(\boldsymbol{x}_t, t) \tag{9}$$

训练方式是使用深度学习最小化交叉熵，具体为

$$L_{\mathrm{CE}} = -\log(p_\theta(\boldsymbol{x}_0)) \tag{10}$$

式中，$p(\theta)$是建模的图像分布。该损失即为希望真实图像$\boldsymbol{x}_0$落在建模的图像分布中。经过数学计算及实验验证，上述损失可以表示为

$$L_{\mathrm{MSE}} = \left\| z_t - z_\theta(\boldsymbol{x}_t, t) \right\|^2 \tag{11}$$

上述损失即为预测的噪声$z_\theta(\boldsymbol{x}_t, t)$与真实噪声$z_t$之间的最小二乘损失。至此，DDPM的训练方式与推断方式已全部阐述完毕。近年来，学术界涌现出了大量基于DDPM的大型工作（Ramesh等，2021，2022；Nichol等，2022；Saharia等，2022a，b；Lugmayr等，2022；Gafni等，2022），这些工作生成的图像的质量之高令人印象深刻，这也从侧面证明了DDPM生成模型的能力。

1.4.3　基于图像—跨模态信息对训练的跨模态生成技术

以跨模态的联合表示为约束，可以训练图像跨模态生成的模型。GAN-INT-CLS是文本—图像跨模态生成的早期作品（Reed等，2016）。它基于一个条件GAN构建，将文本引导设计为条件GAN中的条件输入，以此训练文本—图像跨模态生成模型。该工作作为领域内的早期作品，最终达到的主观质量有限，但是它的开创性价值不容忽视。在Xu等人（2018）提出DAMSM之后，基于DAMSM及其改进型的跨模态生成模型大量出现。其中最为优秀的是DF-GAN（deep fusion GAN）（Tao等，2022）。DF-GAN同样基于一个条件GAN搭建，将文本编码器的输出特征逐步地加入生成的上采样生成模块中，最终取得了优秀的主观质量。

上述工作均基于生成模型GAN。近年来，基于DDPM的模型大量出现，得益于DDPM模型强大的适应能力与生成能力，它们生成的图像具有细腻的纹理，同时不拘泥于某些特定的领域，吸引了学界巨大的关注。

GLIDE（guided language to image diffusion for generation and editing）（Nichol等，2022）是首个基于DDPM的文本—图像跨模态生成模型。它基于Nichol和Dhariwal（2021）以及Dhariwal和Nichol（2021）提出的条件DDPM，在带噪图像上精调CLIP模型充当跨模态语义约束器。DALL-E2（Ramesh等，2022）将跨模态生成过程拆分为从文本到CLIP图像隐向量和从CLIP图像隐向量到图像这两个过程，使用两个DDPM分别训练，取得了比GLIDE更加精细而高质量的生成结果，其内容丰富，体现了强大的生成能力。

1.4.4　基于预训练的特定域图像生成模型的图像跨模态生成方法

前述跨模态图像生成模型均是基于跨模态数据对训练了新的生成模型以达成目标的。然而得益于近年来生成模型的进步，可以借助已有的预训练生成模型，直接设法将它与跨模态语义对齐模块相连接，以达成跨模态图像生成的目标。这类做法的优点在于利用已有的模型，大幅减小训练对于计算资源的需求，而缺点在于难以生成预训练模型可生成的图像域之外的图像。目前，最为常用的预训练大规模图像生成模型之一是StyleGAN（style GAN），代表性工作包括Karras等人（2019，2020，2021）提出的方法。StyleGAN提供的多个隐空间（Wu等，2021）及优秀的解耦能力，为基于它搭建下游任务提供了可能。StyleGAN接收一个隐向量作为输入，通过迭代扩大分辨率的方式，逐步生成一幅高质量的图像。

研究人员想到可以借助将CLIP提供的跨模态隐向量映射到StyleGAN的隐空间中的方法实现跨模态图像编辑。StyleCLIP（Patashnik等，2021）首先做出了这样的尝试。它试图通过一个深度神经网络将描述待编辑的属性映射为一个原始的Style-GAN隐向量的残差，以实现对原图像编辑的目标。该神经网络借助CLIP进行跨模态的语义约束。其优点在于灵活使用了StyleGAN和CLIP这两个大型预训练模型，使得任务事半功倍。但是缺点也十分明显，不能用于图像生成，仅能用于已有图像的编辑。

TediGAN（text-guided diverse image generation and manipulation via GAN）（Xia等，2021）则借助隐向量优化的方法达到上述的目标。首先在StyleGAN的隐空间中随机选取一个起点并生成一幅初始图像，然后用CLIP约束初始图像与待编辑的文本之间的语义一致性，以此为目标对StyleGAN中随机初始化的向量进行优化，最终使得生成的图像与给定的文本之间取得语义一致。由于初始向量的随机性，所以TediGAN的表现很不稳定。StyleGAN-NADA（StyleGAN non-adversarial domain adaptation）（Gal

等，2022）采用精调预训练的StyleGAN中的参数的方式，使StyleGAN有能力生成其原本的生成域之外的图像。目标域的语义则由CLIP进行约束。其缺点同样是无法进行从无到有的图像生成，仅能基于已有的图像进行修改。

2 国内研究现状

2.1 传统跨模态表征学习

近年来，国内相关研究者对于跨模态协同表征学习进行了较为深入的研究。不同于WSABIE和DeViSE等方法对跨模态特征较为粗粒度的相似约束设计，You等人（2018）提出从全局和局部分别对多模态特征进行语义编码，从而进行细粒度的表征学习。通过多层深度神经网络对文本和视觉信息分别编码，得到其在联合空间中的全局表征，在对多模态全局表征进行相似约束学习之外，You等人（2018）还提出对不同模态中间表示层也施以相似度约束，从而保证模型对多模态表征学习能够更为充分。具体而言，You等人（2018）提出对编码器中间表示层进行线性变换，并根据中间表示层和全局变量的相关性对变换后表征进行调整，进而得到用以计算跨模态相似度的局部表征。

Dong等人（2019）也从全局和局部表征学习角度出发，提出了Dual-Encoding方法。Dual-Encoding对视频和文本使用相同的多层级编码网络，对帧级别或单词级别的特征采用平均池化操作得到全局表征，对双向门控循环单元网络（bidirectional gated recurrent unit，BiGRU）所有时刻隐状态取平均操作得到时序模式表征，对BiGRU产生的所有隐状态的2维拼接结果使用不同卷积核大小的1维卷积，得到蕴含多尺度局部信息的表征，最后将这3种编码拼接起来映射到视频—文本共同空间中，并进行基于相似度约束的学习。

Wang等人（2021）将Dual-Encoding作为视频和文本的特征提取器，利用图神经网络（graph neural networks，GNN）进行结构化建模，并对节点之间的关系进行调整。具体来说，对于文本到视频检索，给定一个文本查询，建立以该查询、前K个检索视频和前K个检索文本为节点的全连接图结构，边的特征使用余弦、曼哈顿和欧氏距离的拼接值来初始化，对于每一层图神经网络，节点特征和边特征交替更新，并且设计了打分机制，根据边特征选出新的相似视频集合进入下一层，重复上述操作直至完成这个从粗粒度到细粒度的过程，从而找到最相似的视频。

Chen等人（2020a）提出层次化图推理（hierarchical graph reasoning，HGR）模型，通过层级图推理将视频—文本匹配拆分为从全局到局部的层级，从而实现精细匹配。模型分为3个部分，即层级文本编码、层级视频编码和视频—文本匹配，构建文本的语义关系图结构时，动词作为动作节点与事件节点连接，名词短语作为个体节点与不同的动作节点连接。视频语义关系图则根据帧级、事件级以及全局3个不同层级构建。该模型从文本的语义结构出发，分层级理解文本蕴含的信息，对视频也做出相应的调整，可实现有效的多层级精细匹配。

2.2 多模态大模型表征学习

2.2.1 基于Transformer编码器的方法

基于Transformer编码器的方法根据网络结构不同同样可分为单流模型与双流模型。

1）单流模型。Li等人（2020b）在Unicoder（Huang等，2019）的基础上提出Unicoder-VL（universal encoder for vision and language），以解决不同模态间信息难以融合、没有统一表征方式这一问题。采用一个前置的目标检测模型识别图像中的目标区域，并提取每个区域的特征表示作为图像侧的输入。在预训练任务设置上，Unicoder-VL不仅采用了带掩膜的语言建模（masked language modeling，MLM）方式，还引入了带掩膜的目标分类（masked object classification，MOC）方式。具体而言，MOC方式会对图像中的部分物体进行遮挡，其任务是对被遮挡的物体进行标签预测，该标签直接来源于目标检测识别的结果。Su等人（2020）在VisualBERT（Li等，2019）的基础上主要增加了视觉特征嵌入模块，提出了VL-BERT（visual-linguistic BERT）。具体而言，视觉特征嵌入由外观特征和几何特征两部分组成。外观特征是通过Faster-RCNN（Ren等，2015）对图像中感兴趣区域提取得到的特征信息。对于视觉信息，感兴趣的图像区域是对应内容边界框划定的区域；而对于文本词汇和指示信息，则是完整的图像。几何特征指感兴趣区域边界框相对于图像的位置信息。外观特征和几何特征经过拼接后经全连接层的映射最终得到视觉特征嵌入。

2)双流模型。Zhu和Yang(2020)提出了ActBERT(action BERT),设计了一种全局—局部关系的建模方法,输入包括视频的全局信息,并且利用视频帧中的局部信息加强对于视频内容的理解。另外,提出了掩码动作分类任务,即将动作特征屏蔽,要求模型根据文本和其他视觉特征预测被屏蔽的动作标签。传统的双流网络模型仅考虑两个流之间的实例级对齐,Lu等人(2022)提出了COTS(collaborative two-stream vision-language pre-training model)模型,同时考虑了3个级别的双流交互。(1)传统的实例级交互,使用动量对比学习来学习对齐图像文本;(2)标记(token)级别交互,根据每幅图像未被掩蔽的视觉标记和对应文本的特征进行掩蔽视觉标记预测,类似于进行掩蔽语言标记预测;(3)任务级交互,在文本到图像和图像到文本检索任务之间设计了一种新颖的对齐学习目标,即最小化两个检索任务的概率分布之间的KL(Kullback-Leibler)散度。在CLIP(Radford等,2021)工作的基础上,Yao等人(2022)提出了FILIP(fine-grained interactive language-image pre-training)模型以解决图文匹配中的细粒度(Wei等,2022)匹配问题,通过跨模态后期交互机制实现更细级别的对齐,即计算视觉和文本之间的最大相似度来指导对比目标。仅通过改进对比损失,FILIP就可以利用图像补丁和文本单词之间的细粒度表达,同时保持了大规模数据集训练和推理的效率。Gu等人(2022)发布了一个名为“悟空”的大规模中文跨模态数据集,其中包含了从网络上收集的一亿个中文图像—文本对,以解决领域内缺乏大规模中文数据集及基准的问题,同时基于CLIP和FILIP等工作,还提供了用各种网络架构和方法预训练得到的大规模VLP(visual-linguistic pretrain)模型。Xie等人(2022)提出了一种标记嵌入对齐模块,对基于Transformer编码不同模态信息的方法进行改进,首先显式地对齐视觉标记和文本标记,然后生成标记级匹配分数来度量输入图像和文本之间的细粒度相似性。标记嵌入对齐模块的设计具有显示对齐视觉标记和文本标记的能力,因此它还有良好的可解释性。

2.2.2 基于Transformer解码器的方法

Ding等人(2021)提出了CogView模型,其具有与DALL-E类似的结构,主要面向中文环境的文本到图像生成,采用更少的GPT层数,性能却超越了DALL-E。该模型为了稳定大规模的生成模型训练,提出了一系列有效的策略,包括Sandwich LN和PB-Relax。此外,CogView不仅可以进行零样本的文本到图像生成以及其逆任务(即图像描述生成),在经过微调后也可以应用于超分、风格迁移等任务,具有较强的泛化能力。

2.2.3 联合Transformer编码器与解码器的方法

Luo等人(2020)提出了一种典型的基于编码器—解码器的方法UniVL(unified video and language pre-training),用于多模态理解和生成。首先,单模态编码器用于接收文本和视频输入。然后,利用基于Transformer的交叉编码器来关联文本和视频信息。最后,使用Transformer解码器重建输入文本。UniVL模型能够通过设计的预训练任务执行理解和生成任务,即条件掩蔽语言建模、条件掩蔽视频帧建模、视频文本对齐和文本重建。Xu等人(2021)提出了一种可端到端训练的模型E2E-VLP(end to end VLP),包括Transformer编码器和解码器两部分。其方法除了在编码器端加入掩码语言建模和图文匹配任务,还在解码器端引入了两个新任务,即目标检测和文本描述生成。通过这两个解码器端的任务,可以增强模型对视觉信息的学习能力。Lin等人(2021)提出了类似的编解码器共享的模型M6,该模型面向中文场景的不同任务,设置了更加丰富的注意力,在一系列下游的理解和生成任务上都实现了很好的性能。Liu等人(2021a)提出了多层次多任务统一学习的编解码模型——紫东太初,通过设计样本级、模态级以及Token级多层次自监督学习任务,实现了图文音多模态数据的多粒度关联建模,可有效支撑多模态理解与生成等各种下游任务,并取得了很好的性能。

2.3 图像到文本的跨模态转换

2.3.1 图像语义描述

图像语义描述在视觉障碍助手等辅助任务以及信息检索任务上均有重要的应用前景,然而现有的图像语义描述生成研究大都基于英语语种,近年来,许多国内研究团队开始关注面向中文的图像语义描述生成。与其他跨语言的深度学习研究相似,面向中文的图像语义描述生成的一大难点是数据集的构建,一种可行的方法是基于现有的英文图像语义描述数据集(如MS-COCO(Microsft common object in context),Flickr 30K),利用百度翻译等成熟的中英

翻译接口将英文描述转译成中文，但所得数据集的质量受限于翻译工具的效果，并会因中英语言的差异（词量大小、一词多义等）带来不可避免的噪声。为解决这一问题，张楷文（2021）通过语言模型得到初始化翻译句子对应的符合有关语言表达习惯的分值，过滤掉不符合中文表达习惯的举止，完成数据初步清洗，再在生成过程中使用强化学习方法针对性地进行优化，在奖励函数上反映输出符合语言表达习惯的程度，极大地缓解了数据噪声对模型训练带来的影响。

2.3.2　视频语义描述

国内研究工作在视觉语义描述任务上属于百花齐放的态势。复旦大学团队 Shen 等人（2017）考虑到视觉信息分布的空间离散性和语言描述的密集性，提出了一种基于弱监督的密集视频描述生成法，可以精确到某一区域内物体的动态变化。Wang 等人（2018b）和 Zhou 等人（2018）均从事件的角度出发，以事件为单位，进行视频中响应特征的提取。Wang 等人（2018b）和 Xiong 等人（2018）则是将视觉信息和语言信息的匹配融合交由强化学习算法来实现，取得了可观的性能。Wang 等人（2020）、Zhang 等人（2020）和 Liu 等人（2021a）的工作同时考虑到了视觉信息的动、静态特征，使用 2D、3D 卷积网络结合的方式来丰富化视觉表征。Liu 等人（2021b）则是为了更好地利用视频的时序信息定制化了一种特殊的网络结构。

2.3.3　视频+字幕语义描述

视频+字幕语义描述任务是视觉语义描述领域的一个新研究方向，该任务可以通过字幕帮助模型学习更加抽象的自然语言表征，生成含有高级语义信息的视频描述，能够给观众在浏览和检索视频内容时带来更好的体验。然而，由于字幕是零碎的信息，与视觉形态存在语义差距，因此字幕的有效使用也非常具有挑战性。为了将零碎的信息组织在一起，并为所有模态生成语义相关性更高的全局表示，Tu 等人（2022）提出了 I2Transformer（intra- and inter-relation embedding transformer）模型，通过多模态信息融合实现视频和字幕的全局表示。该模型包括 IAE（intra-relation embedding block）和 IEE（inter-relation embedding block）两部分，用来学习视频中的内部关系和副标题，以及它们之间的相互关系。这有利于理解每种模态的语义和跨模态的语义交互。首先，IAE 通过构建可学习图来捕获每种模态中的内部关系。然后，IEE 作为一个可学习的交叉注意力门，通过学习视觉和字幕的相互关系从每个模态中提取有用的信息作为 Transformer 的输入。哈尔滨工业大学 Nie 等人（2022）设计了一个大规模多模态的预训练网络，通过 5 项任务来加强下游视频表征，并进一步提出了一种基于流的多样化字幕模型，以根据用户的搜索需求生成不同的字幕，该模型通过重建损失在先验和后验之间的 KL 分歧进行优化，从针对用户搜索需求的角度，自动生成文本去描述一个短视频，以满足用户搜索视频的多样化需求。

2.3.4　变化语义描述

国内研究团队在变化语义描述任务上也发表了多项研究成果。这些研究成果与国际研究趋势同步，即研究如何在视角变化中区分和描述语义变化。其中，中国科学院计算技术研究所和昆明理工大学的研究团队在自然语言处理顶级学术会议 ACL（Association for Computational Linguistics）和 EMNLP（Conference on Empirical Methods in Natural Language Processing）上发布了两项研究成果。具体而言，Tu 等人（2021）提出一个 SRDRL（semantic relation-aware difference representation learning）模型来衡量差异表征和图像表征的语义相似度，并将其作为一种先验知识来帮助模型判断是否存在语义变化以及潜在位置。同时，提出了一个 R3Net（recurrent residual refinement network），根据语义相似度重构出每幅图像上未变化的特征，进而求出变化特征。此外，广西大学的研究团队除了计算图像对间的表征差异外，引入了深度（Liao 等，2021）以及语义属性（Huang 等，2021）等额外知识来建模差异信息。中国人民大学的 Yao 等人（2022）则提出一种基于预训练的训练方式来完成变化语义描述。

2.3.5　视觉问答

国内研究进展主要集中在提出表达能力更强的多模态融合模型与提高模型的可解释性与泛化能力上。胡钦太等人（2021）利用深度学习算法进行多模态学习分析，采用深度混合判别受限玻尔兹曼机神经网络模型，建立多模态学习分析模型。从可解释性分析的角度，利用深度学习算法进行多模态学习行为分析的算法设计与实现过程。

3 国内外研究进展比较

3.1 传统跨模态表征学习

跨模态表征学习长久以来受到学术界和工业界的关注，机器学习模型的表现非常依赖数据表征的选择，一方面高质量的跨模态表征能够极大节省工业应用的成本；另一方面其也对下游任务的学术研究提供便利。传统跨模态表征学习的研究主要关注基于机器学习的模态表征学习，分为跨模态统一表征学习和跨模态协同表征学习两个主要研究方向。随着深度学习技术的兴起，国内研究者在协同表征学习方面贡献了越来越多高质量高影响力的工作，对于协同表征学习，进行了更为细粒度的协同约束，同时根据模态特点设计了多尺度、多层级的编码模块以及匹配模块。

3.2 多模态大模型表征学习

近年国际上主要科研机构和大型企业均在抓紧布局大模型技术，国际主要机构XGOpenAI、谷歌、微软、脸书和英伟达等，国内华为、阿里巴巴、百度、中国科学院自动化研究所和清华大学等众多科研机构纷纷加入研发赛道，大模型成果不断推陈出新。多模态大模型在2019年前后的早期工作主要由国外相关学者和研究机构展开，其工作集中于面向多模态理解任务的多模态预训练模型，设计并提出了一系列经典的多模态预训练方法。在2020年前后，国内学者逐渐发力，相关研究开始逐渐占据主流并贡献了大量的优越方法。特别是随着多模态大模型表征学习在多模态任务的广泛应用，近年部分国内团队已实现国际领先水平，同时还针对中文和多语言背景下的多模态大模型学习进行了深入探索。

3.3 图像到文本的跨模态转换

3.3.1 图像语义描述

考虑到图像语义描述生成现实应用的需求，国内外在图像语义描述任务上的差异主要体现在对中英语种的关注程度，国际上更多地进行英文描述生成的研究，而国内近年来对中文描述生成的关注程度越来越高。从方法上，国内外研究团队主流的研究方向基本一致，都是以编码器—解码器结构为载体，通过不同形式的注意力机制变种抽取不同类型的多模态特征信息，以减小图文模态间语义鸿沟为桥梁，实现更准确或更有吸引力的描述生成。代表性成果包括田枫等人(2021)、廖雷双(2021)、Kavi等人(2022)以及Das和Singh(2022)的工作。

3.3.2 视频语义描述

在视频语义描述领域，国外研究者为学术界提供了很多基础性的思路以及解决方案，涉及相关基础性网络的提出、数据集的搭建、任务的定义以及评价指标的设定，比较好地将视频到语言生成任务的评测数据集构建起来，供研究者一同参考。相比于国外研究，国内的研究在数据集以及评价指标和任务定义方面有所不足，但是研究者可以很好地发掘任务过程中遇到的一些关键性问题并给予解决方案，从细粒度的层次不断将任务方法完善成熟。

3.3.3 视频字幕语义描述

基于电影/卡通/电视节目实现视频领域的多模态任务早期主要由国外学者和相关机构展开，其工作主要集中于数据集的制作以辅助实现其他多模态任务。自2020年视频字幕语义描述任务提出后，通过字幕辅助模型学习更加高级的语义表征这一方向展开了研究热潮。国内学者从模态间的表征学习本身出发，将研究重点集中在缩小视频和字幕这两种不同模态之间的语义鸿沟中，相关的研究也开始逐渐占据主流，特别在短视频描述领域处于国际先进水平。

3.3.4 变化语义描述

近年来国内外的学术机构和工业界对变化语义描述的关注程度显著上升，一方面是由于图像大数据的应用场景和用户需求正在发生巨变；另一方面是由于人工智能技术的飞速发展引领了新一轮的技术革新。变化语义描述领域早期工作主要由国外相关学者和研究机构展开。约2021年前后，国内学者逐渐发力，相关研究开始逐渐占据主流并贡献了大量的优越方法，部分国内团队已实现国际领先水平。

3.3.5 视觉问答

国外视觉问答的研究主要集中在发展规模更大的、更平衡的数据集以及提出表达能力更强的多模态融合方法。国内的视觉问答研究在提出表达能力更强的融合模型的基础上，还聚焦于提高模型的可解释性与泛化能力等方面。

4 发展趋势与展望

4.1 传统跨模态表征学习

随着大模型和预训练技术的兴起，跨模态表征学习愈加受到了上下游任务研究者的关注。研究者开始尝试通过大规模预训练模型在海量训练数据上进行预训练，从而学习到高质量的跨模态表征信息，从而为下游具体跨模态任务提供便利。在训练成本巨大的条件下，如何对下游任务设计通用的跨模态预训练框架是表征学习中亟待解决的问题。同时，尽管当前基于预训练的表征学习能够获得较高质量的模态表征，但其可解释性仍然较差，如何通过传统跨模态表征学习的研究成果对大模型表征的可解释性进行提升是需要探索的方向。

4.2 多模态大模型表征学习

在多模态数据表征学习方面，多模态大模型全面颠覆了传统方法，开创了多模态数据分析理解的新纪元。多模态大模型强大的自监督学习与通用知识迁移能力，大幅降低了具体任务上对人工标注数据的依赖程度。目前，多模态大模型表征学习在多种任务上突破了传统方法的性能缺陷，取得了飞跃式的发展。未来“大数据+大模型”这样一种研究范式或还将继续，应关注预训练数据、基础模型、自监督学习以及下游任务模型适配等核心问题，力争实现面向大数据的多模态大模型的高效鲁棒计算与应用部署。具体来说，1）构建高质量大规模的多模态关联数据集，实现基于全网信息的多模态关联数据的自动收集与智能清洗；2）基于Transformer模型的优化改进甚至升级替代，实现面向大规模多模态弱关联数据的鲁棒自监督学习与高效计算；3）优化设计自监督学习算法，并充分考虑多模态数据的弱关联、有噪声且存在模态缺失等问题，实现多模态信息的细粒度语义关联；4）面向下游任务的模型微调，并辅以模型轻量化、推理加速等手段，实现大模型的低代价迁移学习与高效应用部署。

4.3 图像到文本的跨模态转换

在图像语义描述生成领域，仍然有许多有前景的研究方向值得关注，如包括之前的风格化描述在内，现有模型在生成更加丰富有趣的描述方面仍然有提高的空间；另外，无监督学习和强化学习在未来可能会受到更多的欢迎；此外，现有评估图像语义描述生成质量的常用指标大多仍来自于传统机器翻译任务，设计客观全面的评价指标有助于促进现有方法的进步。

视频语义描述任务一直受国内外学术界、工业界广泛关注，原因在于其本身的任务难度颇高、同时非常贴近人们的实际生活。随着互联网的不断发展，视频数据爆炸性增加，完全依赖人工标注的模型训练方法已经渐渐失去了竞争力，未来该任务的发展势必趋向于充分利用当下非常成功的跨模态大模型先验知识，在自监督、半监督条件下从“大数据、小模型”向“大模型、小数据”方向靠拢。

当前，跨媒体综合学习、知识有效获取与利用以及泛化推理是前沿研究热点问题，受益于视频和字幕之间信息的交互式学习，视频+字幕语义描述的研究不仅可以增强视觉模态内的理解，也可以提高模型的认知能力，使模型在人工智能的类人规划和自我学习能力方面实现新的突破。

变化语义描述在图文转换领域还是一个新型课题，虽然国内外学术机构已经取得了一定的研究成果，但现在的数据集无论在规模还是设定方面都与真实场景相距甚远。如何准确定位和描述动态环境中的复杂变化仍然需要国内外研究团队进行深入探索。

视觉问答领域存在的主要挑战为存在不同模态的模型偏好以及与模型本身的表达能力有限的问题。因此视觉问答未来的主要发展方向为构建更全面均衡的数据集以及提高模型的可解释性、鲁棒性与泛化能力。

4.4 图像生成

首先，现有的图像生成技术已有能力生成高分辨率的拟真图像，但在图像生成多样性方面仍然存在欠缺，而可生成图像的多样性高低是图像生成技术的重要标准。通过单个模型生成开放世界的图像是实际应用场景对图像模型的现实要求。因此，未来的图像生成技术发展方向之一是生成模型可生成的图像多样化扩展，以满足实际应用场景需求。其次，现有的图像生成技术还无法实现对生成的图像进行解耦的精细化控制。基于现有的技术，当试图改变生成的图像中某一个对象时，图像的其余部分会不可避免地发生改变，无法对图像进行精细地控制与编辑。因此，设计可解耦的生成模型结构以达到上述目标也是图像生成技术未来的发展方向。

致　谢　本文由中国图象图形学学会多媒体专业委员会组织撰写，该专委会链接为 http://www.csig.org.cn/detail/2391。

参考文献(References)

Andrew G, Arora R, Bilmes J and Livescu K. 2013. Deep canonical correlation analysis//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA: JMLR.org: 1247-1255

Arjovsky M, Chintala S and Bottou L. 2017. Wasserstein generative adversarial networks//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR.org: 214-223

Baltrušaitis T, Ahuja C and Morency L P. 2019. Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2): 423-443 [DOI: 10.1109/TPAMI.2018.2798607]

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D. 2020. Language models are few-shot learners//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 1877-1901

Cao Y, Long M S, Wang J M, Yang Q and Yu P S. 2016. Deep visual-semantic hashing for cross-modal retrieval//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA: ACM: 1445-1454 [DOI: 10.1145/2939672.2939812]

Chen S Z, Zhao Y D, Jin Q and Wu Q. 2020a. Fine-grained video-text retrieval with hierarchical graph reasoning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10635-10644 [DOI: 10.1109/CVPR42600.2020.01065]

Chen Y C, Li L J, Yu L C, El Kholy A, Ahmed F, Gan Z, Cheng Y and Liu J J. 2020b. Uniter: universal image-text representation learning//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 104-120 [DOI: 10.1007/978-3-030-58577-8_7]

Cho J, Lei J, Tan H and Bansal M. 2021. Unifying vision-and-language tasks via text generation//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR: 1931-1942

Cho K, Van Merriënboer B, Gulçehre Ç, Bahdanau D, Bougares F, Schwenk H and Bengio Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics: 1724-1734 [DOI: 10.3115/v1/D14-1179]

Das R and Singh T D. 2022. Assamese news image caption generation using attention mechanism. Multimedia Tools and Applications, 81(7): 10051-10069 [DOI: 10.1007/s11042-022-12042-8]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minnesota, USA: Association for Computational Linguistics: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Dhariwal P and Nichol A. 2021. Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems, 34: 8780-8794

Ding M, Yang Z Y, Hong W Y, Zheng W D, Zhou C, Yin D, Lin J Y, Zou X, Shao Z, Yang H X and Tang J. 2021. CogView: mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems 34: 19822-19835

Dong J F, Li X R, Xu C X, Ji S L, He Y, Yang G and Wang X. 2019. Dual encoding for zero-example video retrieval//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9338-9347 [DOI: 10.1109/CVPR.2019.00957]

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]. [2023-01-05]. https://arxiv.org/pdf/2010.11929.pdf

Frome A, Corrado G S, Shlens J, Bengio S, Dean J, Ranzato M and Mikolov T. 2013. DeViSE: a deep visual-semantic embedding model//Proceedings of the 26th International Conference on Neural Information Processing Systems. Nevada, USA: Curran Associates Inc.: 2121-2129

Gafni O, Polyak A, Ashual O, Sheynin S, Parikh D and Taigman Y. 2022. Make-A-scene: scene-based text-to-image generation with human priors//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 89-106 [DOI: 10.1007/978-3-031-19784-0_6]

Gal R, Patashnik O, Maron H, Bermano A H, Chechik G and Cohen-Or D. 2022. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. ACM Transactions on Graphics, 41(4): #141 [DOI: 10.1145/3528223.3530164]

Gao J Y, Sun C, Yang Z H and Nevatia R. 2017. Tall: temporal activity localization via language query//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5277-5285 [DOI: 10.1109/ICCV.2017.563]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680

Gu J X, Meng X J, Lu G S, Hou L, Niu M Z, Liang X D, Yao L W,

Huang R H, Zhang W, Jiang X, Xu C J and Xu H. 2022. Wukong: a 100 million large-scale Chinese cross-modal pre-training benchmark [EB/OL]. [2023-01-05]. https://arxiv.org/pdf/2202.06767.pdf

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A C. 2017. Improved training of wasserstein GANs//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5769-5779

Hendricks L A, Wang O, Shechtman E, Sivic J, Darrell T and Russell B. 2017. Localizing moments in video with natural language//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5804-5813 [DOI: 10.1109/ICCV.2017.618]

Hinton G E, Osindero S and Teh Y W. 2006. A fast learning algorithm for deep belief nets. Neural Computation, 18(7): 1527-1554 [DOI: 10.1162/neco.2006.18.7.1527]

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/neco.1997.9.8.1735]

Hosseinzadeh M and Wang Y. 2021. Image change captioning by learning from an auxiliary task//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2724-2733 [DOI: 10.1109/CVPR46437.2021.00275]

Hu Q T, Wu W Y, Feng G, Pan T F and Qiu K X. 2021. A study on interpretable analysis of multimodal learning behavior supported by deep learning learning. E-education Research, 42(11): 77-83 (胡钦太, 伍文燕, 冯广, 潘庭锋, 邱凯星. 2021. 深度学习支持下多模态学习行为可解释性分析研究. 电化教育研究, 42(11): 77-83) [DOI: 10.13811/j.cnki.eer.2021.11.011]

Huang H Y, Liang Y B, Duan N, Gong M, Shou L J, Jiang D X and Zhou M. 2019. Unicoder: a universal language encoder by pre-training with multiple cross-lingual tasks//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics: 2485-2494 [DOI: 10.18653/v1/D19-1252]

Huang Q B, Liang Y, Wei J L, Cai Y, Liang H Y, Leung H F and Li Q. 2021. Image difference captioning with instance-level fine-grained feature representation. IEEE Transactions on Multimedia, 24: 2004-2017 [DOI: 10.1109/TMM.2021.3074803]

Jhamtani H and Berg-Kirkpatrick T. 2018. Learning to describe differences between pairs of similar images//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics: 4024-4034 [DOI: 10.18653/v1/D18-1436]

Jia C, Yang Y F, Xia Y, Chen Y T, Parekh Z, Pham H, Le Q, Sung Y H, Li Z and Duerig T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: [s.n.]: 4904-4916

Jiang Q Y and Li W J. 2017. Deep cross-modal hashing//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 3270-3278 [DOI: 10.1109/CVPR.2017.348]

Johnson J, Karpathy A and Li F F. 2016. DenseCap: fully convolutional localization networks for dense captioning//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 4565-4574 [DOI: 10.1109/CVPR.2016.494]

Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J and Aila T. 2021. Alias-free generative adversarial networks [EB/OL]. [2023-01-05]. https://arxiv.org/pdf/2106.12423.pdf

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 4396-4405 [DOI: 10.1109/CVPR.2019.00453]

Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J and Aila T. 2020. Analyzing and improving the image quality of StyleGAN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 8107-8116 [DOI: 10.1109/CVPR42600.2020.00813]

Kavi P S, Pon K K, Kaliappan J, Selvaraj S K, Nagalakshmi R and Molla B. 2022. Caption generation based on emotions using CSP-DenseNet and BiLSTM with self-attention. Applied Computational Intelligence and Soft Computing, 2022: #2756396 [DOI: 10.1155/2022/2756396]

Kim H, Kim J, Lee H, Park H and Kim G. 2021a. Viewpoint-agnostic change captioning with cycle consistency//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 2075-2084 [DOI: 10.1109/ICCV48922.2021.00210]

Kim W, Son B and Kim I. 2021b. ViLT: vision-and-language transformer without convolution or region supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR: 5583-5594

Kim Y, Lee H and Provost E M. 2013. Deep learning for robust feature generation in audiovisual emotion recognition//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE: 3687-3691 [DOI: 10.1109/ICASSP.2013.6638346]

Kiros R, Salakhutdinov R and Zemel R S. 2014. Unifying visual-semantic embeddings with multimodal neural language models [EB/OL]. [2023-01-05]. https://arxiv.org/pdf/1411.2539.pdf

Krishna R, Hata K, Ren F, Li F F and Niebles J C. 2017. Dense-captioning events in videos//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 706-715 [DOI: 10.1109/ICCV.2017.83]

Lai P L and Fyfe C. 2000. Kernel and nonlinear canonical correlation

analysis. International Journal of Neural Systems, 10(5): 365-377 [DOI: 10.1142/S012906570000034X]

Lei J, Yu L C, Berg T L and Bansal M. 2020. TVR: a large-scale dataset for video-subtitle moment retrieval//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 447-463 [DOI: 10.1007/978-3-030-58589-1_27]

Li B W, Qi X J, Lukasiewicz T and Torr P H S. 2020a. ManiGAN: text-guided image manipulation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7877-7886 [DOI: 10.1109/CVPR42600.2020.00790]

Li C X and Harrison B. 2021. 3M: multi-style image caption generation using multi-modality features under multi-UPDOWN model [EB/OL]. [2023-01-05]. https://arxiv.org/pdf/2103.11186.pdf

Li C X and Harrison B. 2022. StyleM: stylized metrics for image captioning built with contrastive *N*-grams [EB/OL]. [2023-01-05]. https://arxiv.org/pdf/2201.00975.pdf

Li G, Duan N, Fang Y J, Gong M and Jiang D X. 2020b. Unicoder-VL: a universal encoder for vision and language by cross-modal pre-training. Proceedings of 2020 AAAI Conference on Artificial Intelligence, 34(7): 11336-11344 [DOI: 10.1609/aaai.v34i07.6795]

Li L H, Yatskar M, Yin D, Hsieh C J and Chang K W. 2019. VisualBERT: a simple and performant baseline for vision and language [EB/OL]. [2023-01-05]. https://arxiv.org/pdf/1908.03557.pdf

Li L J, Chen Y C, Cheng Y, Gan Z, Yu L C and Liu J J. 2020c. Hero: hierarchical encoder for video+language omni-representation pre-training//Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. [s.l.]: Association for Computational Linguistics: 2046-2065 [DOI: 10.18653/v1/2020.emnlp-main.161]

Liao L S. 2021. A Research on Image Description Based on Attention and Multi-level Vision Features. Shanghai: Shanghai University of Finance and Economics (廖雷双. 2021. 基于注意力机制与多层级视觉特征的图像描述方法研究. 上海: 上海财经大学) [DOI: 10.27296/d.cnki.gshcu.2021.001921]

Liao Z M, Huang Q B, Liang Y, Fu M Y, Cai Y and Li Q. 2021. Scene graph with 3D information for change captioning//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: ACM: 5074-5082 [DOI: 10.1145/3474085.3475712]

Lin J Y, Men R, Yang A, Zhou C, Zhang Y C, Wang P, Zhou J R, Tang J and Yang H X. 2021. M6: multi-modality-to-multi-modality multitask mega-transformer for unified pretraining//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Virtual Event, Singapore: ACM: 3251-3261 [DOI: 10.1145/3447548.3467206]

Liu J, Zhu X X, Liu F, Guo L T, Zhao Z J, Sun M Z, Lu H Q, Wang W N, Lu H Q, Zhou S Y, Zhang J J and Wang J Q. 2021a. OPT: omni-perception pre-trainer for cross-modal understanding and generation [EB/OL]. [2023-01-05]. https://arxiv.org/pdf/2107.00249.pdf

Liu S, Ren Z and Yuan J S. 2021b. SibNet: sibling convolutional encoder for video captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(9): 3259-3272 [DOI: 10.1109/TPAMI.2019.2940007]

Lu H Y, Fei N Y, Huo Y Q, Gao Y Z, Lu Z W and Wen J R. 2022. COTS: collaborative two-stream vision-language pre-training model for cross-modal retrieval//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15671-15680 [DOI: 10.1109/CVPR52688.2022.01524]

Lu J S, Batra D, Parikh D and Lee S. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 13-23

Lu J S, Yang J W, Batra D and Parikh D. 2016. Hierarchical question-image co-attention for visual question answering//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 289-297

Lugmayr A, Danelljan M, Romero A, Yu F, Timofte R and Van Gool L. 2022. RePaint: inpainting using denoising diffusion probabilistic models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11451-11461 [DOI: 10.1109/CVPR52688.2022.01117]

Luo H S, Ji L, Shi B T, Huang H Y, Duan N, Li T R, Li J, Bharti T and Zhou M. 2020. UniVL: a unified video and language pre-training model for multimodal understanding and generation [EB/OL]. [2023-01-05]. https://arxiv.org/pdf/2002.06353.pdf

Mirza M and Osindero S. 2014. Conditional generative adversarial nets [EB/OL]. [2023-01-05]. https://arxiv.org/pdf/1411.1784.pdf

Ngiam J, Khosla A, Kim M, Nam J, Lee H and Ng A Y. 2011. Multimodal deep learning//Proceedings of the 28th International Conference on Machine Learning. Bellevue, USA: Omnipress: 689-696

Nguyen K, Tripathi S, Du B, Guha T and Nguyen T Q. 2021. In defense of scene graphs for image captioning//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1387-1396 [DOI: 10.1109/ICCV48922.2021.00144]

Nichol A Q and Dhariwal P. 2021. Improved denoising diffusion probabilistic models//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR: 8162-8171

Nichol A Q, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I and Chen M. 2022. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 16784-16804

Nie L Q, Qu L G, Meng D, Zhang M, Tian Q and del Bimbo A. 2022. Search-oriented micro-video captioning//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 3234-3243 [DOI: 10.1145/3503161.3548180]

Park D H, Darrell T and Rohrbach A. 2019. Robust change captioning// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4623-4632 [DOI: 10.1109/ICCV.2019.00472]

Patashnik O, Wu Z Z, Shechtman E, Cohen-Or D and Lischinski D. 2021. StyleCLIP: text-driven manipulation of StyleGAN imagery// Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 2065-2074 [DOI: 10.1109/ICCV48922.2021.00209]

Qiao T T, Zhang J, Xu D Q and Tao D C. 2019. MirrorGAN: learning text-to-image generation by redescription//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1505-1514 [DOI: 10.1109/CVPR.2019.00160]

Qiu Y, Yamamoto S, Nakashima K, Suzuki R, Iwata K, Kataoka H and Satoh Y. 2021. Describing and localizing multiple changes with transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1951-1960 [DOI: 10.1109/ICCV48922.2021.00198]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Edinburgh, Scotland: PMLR: 8748-8763

Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M. 2022. Hierarchical text-conditional image generation with CLIP latents [EB/OL]. [2023-01-05]. https://arxiv.org/pdf/2204.06125.pdf

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M and Sutskever I. 2021. Zero-shot text-to-image generation//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR: 8821-8831

Rastegar S, Soleymani M S, Rabiee H R and Shojaee S M. 2016. MDL-CW: a multimodal deep learning framework with cross weights//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2601-2609 [DOI: 10.1109/CVPR.2016.285]

Reed S, Akata Z, Yan X C, Logeswaran L, Schiele B and Lee H. 2016. Generative adversarial text to image synthesis//Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR.org: 1060-1069

Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 91-99

Saharia C, Chan W, Chang H W, Lee C, Ho J, Salimans T, Fleet D and Norouzi M. 2022a. Palette: image-to-image diffusion models// Proceedings of ACM SIGGRAPH 2022 Conference Proceedings. Vancouver, Canada: ACM: #15 [DOI: 10.1145/3528233.3530757]

Saharia C, Ho J, Chan W, Salimans T, Fleet D J and Norouzi M. 2022b. Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence: #320446 [DOI: 10.1109/TPAMI.2022.3204461]

Shen Z Q, Li J G, Su Z, Li M J, Chen Y R, Jiang Y G and Xue X Y. 2017. Weakly supervised dense video captioning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5159-5167 [DOI: 10.1109/CVPR.2017.548]

Shi X X, Yang X, Gu J X, Joty S and Cai J F. 2020. Finding it at another side: a viewpoint-adapted matching encoder for change captioning//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 574-590 [DOI: 10.1007/978-3-030-58568-6_34]

Silberer C and Lapata M. 2014. Learning grounded meaning representations with autoencoder//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA: Association for Computational Linguistics: 721-732 [DOI: 10.3115/v1/P14-1068]

Sohl-Dickstein J, Weiss E A, Maheswaranathan N and Ganguli S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics// Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org: 2256-2265

Srivastava N and Salakhutdinov R. 2012. Learning representations for multimodal data with deep belief nets//Proceedings of 2012 International Conference on Machine Learning Workshop, Edinburgh, Scotland: PMLR: 1-8

Su W J, Zhu X Z, Cao Y, Li B, Lu L W, Wei F R and Dai J F. 2020. VL-BERT: pre-training of generic visual-linguistic representations// Proceedings of 2020 International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR: 1-6

Sun C, Myers A, Vondrick C, Murphy K and Schmid C. 2019. VideoBERT: a joint model for video and language representation learning// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 7463-7472 [DOI: 10.1109/ICCV.2019.00756]

Tan H and Bansal M. 2019. LXMert: learning cross-modality encoder representations from transformers//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics: 5100-5111 [DOI: 10.18653/v1/D19-1514]

Tao M, Tang H, Wu F, Jing X Y, Bao B K and Xu C S. 2022. DF-GAN: a simple and effective baseline for text-to-image synthesis// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16494-16504 [DOI: 10.1109/CVPR52688.2022.01602]

Tian F, Sun X Q, Liu F, Li T Y, Zhang L and Liu Z G. 2021. Chinese image caption with dual attention and multi-label image. Computer Systems and Applications, 30(7): 32-40 (田枫, 孙小强, 刘芳, 李婷玉, 张蕾, 刘志刚. 2021. 融合双注意力与多标签的图像中

文描述生成方法. 计算机系统应用, 30(7): 32-40)[DOI: 10.15888/j.cnki.csa.008010]

Tu Y B, Li L, Su L, Gao S X, Yan C G, Zha Z J, Yu Z T and Huang Q M. 2022. I²Transformer: intra- and inter-relation embedding transformer for TV show captioning. IEEE Transactions on Image Processing, 31: 3565-3577[DOI: 10.1109/TIP.2022.3159472]

Tu Y B, Yao T T, Li L, Lou J D, Gao S X, Yu Z T and Yan C G. 2021. Semantic relation-aware difference representation learning for change captioning//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Virtual Event: Association for Computational Linguistics: 63-73[DOI: 10.18653/v1/2021.findings-acl.6]

van den Oord A, Vinyals O and Kavukcuoglu K. 2017. Neural discrete representation learning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6309-6318

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

Venugopalan S, Rohrbach M, Donahue J, Mooney R, Darrell T and Saenko K. 2015. Sequence to sequence-video to text//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 4534-4542[DOI: 10.1109/ICCV.2015.515]

Wang C, Yang H J, Bartz C and Meinel C. 2016. Image captioning with deep bidirectional LSTMs//Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, the Netherlands: ACM: 988-997[DOI: 10.1145/2964284.2964299]

Wang J F, Yang Z Y, Hu X W, Li L J, Lin K, Gan Z, Liu Z C, Liu C and Wang L J. 2022. GIT: a generative image-to-text transformer for vision and language[EB/OL].[2023-01-05]. https://arxiv.org/pdf/2205.14100.pdf

Wang L X, Shang C, Qiu H Q, Zhao T J, Qiu B L and Li H L. 2020. Multi-stage tag guidance network in video caption//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 4610-4614[DOI: 10.1145/3394171.3416288]

Wang W, Gao J Y, Yang X S and Xu C S. 2021. Learning coarse-to-fine graph neural networks for video-text retrieval. IEEE Transactions on Multimedia, 23: 2386-2397[DOI: 10.1109/tmm.2020.3011288]

Wang X, Chen W H, Wu J W, Wang Y F and Wang W Y. 2018b. Video captioning via hierarchical reinforcement learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4213-4222[DOI: 10.1109/CVPR.2018.00443]

Wei X S, Song Y Z, Aodha O M, Wu J X, Peng Y X, Tang J H, Yang J and Belongie S. 2022. Fine-grained image analysis with deep learning: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12): 8927-8948[DOI: 10.1109/TPAMI.2021.3126648]

Weston J, Bengio S and Usunier N. 2010. Large scale image annotation: learning to rank with joint word-image embeddings. Machine Learning, 81(1): 21-35[DOI: 10.1007/s10994-010-5198-3]

Wu C F, Liu J L, Wang X J and Dong X. 2018. Object-difference attention: a simple relational attention for visual question answering//Proceedings of the 26th ACM International Conference on Multimedia. Seoul Korea (South): ACM: 519-527[DOI: 10.1145/3240508.3240513]

Wu Z Z, Lischinski D and Shechtman E. 2021. StyleSpace analysis: disentangled controls for StyleGAN image generation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12858-12867[DOI: 10.1109/CVPR46437.2021.01267]

Xia W H, Yang Y J, Xue J H and Wu B Y. 2021. TediGAN: text-guided diverse face image generation and manipulation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2256-2265[DOI: 10.1109/CVPR46437.2021.00229]

Xie C W, Wu J M, Zheng Y, Pan P and Hua X S. 2022. Token embeddings alignment for cross-modal retrieval//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 4555-4563[DOI: 10.1145/3503161.3548107]

Xiong Y L, Dai B and Lin D H. 2018. Move forward and tell: a progressive generator of video descriptions//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 489-505[DOI: 10.1007/978-3-030-01252-6_29]

Xu H Y, Yan M, Li C L, Bi B, Huang S F, Xiao W M and Huang F. 2021. E2E-VLP: end-to-end vision-language pre-training enhanced by visual learning//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: 503-513[DOI: 10.18653/v1/2021.acl-long.42]

Xu T, Zhang P C, Huang Q Y, Zhang H, Gan Z, Huang X L and He X D. 2018. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1316-1324[DOI: 10.1109/CVPR.2018.00143]

Xue H W, Hang T K, Zeng Y H, Sun Y C, Liu B, Yang H, Fu J L and Guo B N. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5026-5035[DOI: 10.1109/CVPR52688.2022.00498]

Yao L L, Wang W Y and Jin Q. 2022. Image difference captioning with pre-training and contrastive learning. Proceedings of 2022 AAAI Conference on Artificial Intelligence, 36(3): 3108-3116[DOI: 10.1609/aaai.v36i3.20218]

You Q Z, Luo J B and Zhang Z Y. 2018. End-to-end convolutional semantic embeddings//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5735-5744 [DOI: 10.1109/CVPR.2018.00601]

Yuan L, Chen D D, Chen Y L, Codella N, Dai X Y, Gao J F, Hu H D, Huang X D, Li B X, Li C Y, Liu C, Liu M C, Liu Z C, Lu Y M, Shi Y, Wang L J, Wang J F, Xiao B, Xiao Z, Yang J W, Zeng M, Zhou L W and Zhang P C. 2021. Florence: a new foundation model for computer vision [EB/OL]. [2021-11-22]. https://arxiv.org/pdf/2111.11432.pdf

Zhang H, Xu T, Li H S, Zhang S T, Wang X G, Huang X L and Metaxas D. 2017. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5908-5916 [DOI: 10.1109/ICCV.2017.629]

Zhang H, Xu T, Li H S, Zhang S T, Wang X G, Huang X L and Metaxas D N. 2019. StackGAN++: realistic image synthesis with stacked generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8): 1947-1962 [DOI: 10.1109/TPAMI.2018.2856256]

Zhang K W. 2021. Research on Chinese-Oriented Image Caption Generation Method. Harbin: Harbin Institute of Technology (张楷文. 2021. 面向中文的图像描述生成方法研究. 哈尔滨: 哈尔滨工业大学) [DOI: 10.27061/d.cnki.ghgdu.2021.003103]

Zhang Z J, Wu Q, Wang Y and Chen F. 2021. Exploring region relationships implicitly: image captioning with visual relationship attention. Image and Vision Computing, 109: #104146 [DOI: 10.1016/J.IMAVIS.2021.104146]

Zhang Z Q, Shi Y Y, Yuan C F, Li B, Wang P J, Hu W M and Zha Z J. 2020. Object relational graph with teacher-recommended learning for video captioning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13275-13285 [DOI: 10.1109/cvpr42600.2020.01329]

Zhou L W, Palangi H, Zhang L, Hu H D, Corso J and Gao J F. 2020. Unified vision-language pre-training for image captioning and VQA//Proceedings of 2020 AAAI Conference on Artificial Intelligence, 34(7): 13041-13049 [DOI: 10.1609/AAAI.V34I07.7005]

Zhou L W, Zhou Y B, Corso J J, Socher R and Xiong C M. 2018. End-to-end dense video captioning with masked transformer//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8739-8748 [DOI: 10.1109/CVPR.2018.00911]

Zhu L C and Yang Y. 2020. ActBERT: learning global-local video-text representations//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8743-8752 [DOI: 10.1109/cvpr42600.2020.00877]

Zhu M F, Pan P B, Chen W and Yang Y. 2019. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5795-5803 [DOI: 10.1109/CVPR.2019.00595]

作者简介

刘华峰，男，博士后，主要研究方向为多媒体和计算机视觉。E-mail: liu.hua.feng@njust.edu.cn

聂礼强，通信作者，男，教授，主要研究方向为多媒体内容分析与搜索。E-mail: nieliqiang@gmail.com

陈静静，女，副教授，主要研究方向为多媒体内容分析。E-mail: chenjingjing@fudan.edu.cn

李亮，男，副研究员，主要研究方向为视觉与语言建模和机器学习。E-mail: liang.li@ict.ac.cn

鲍秉坤，女，教授，主要研究方向为多媒体计算和计算机视觉。E-mail: bingkunbao@njupt.edu.cn

李泽超，男，教授，主要研究方向为多媒体分析与检索、人工智能和计算机视觉。E-mail: zechao.li@njust.edu.cn

刘家瑛，女，副教授，主要研究方向为智能媒体计算与视觉理解。E-mail: liujiaying@pku.edu.cn