Tu Y, Li L, Yan C, Gao S and Yu Z. 2021. R3Net: Relation-embedded Representation Reconstruction Network for Change Captioning//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 9319-9329 [DOI:10.18653/v1/2021.emnlp-main.735]

Tu Y, Yao T, Li L, Lou J, Gao S, Yu Z and Yan C. 2021. Semantic Relation-aware Difference Representation Learning for Change Captioning. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 63-73 [DOI:10.18653/v1/2021.findings-acl.6]

Van Den Oord A and Vinyals O. 2017. Neural Discrete Representation Learning//Advances in Neural Information Processing Systems (NIPS), 6306-6315

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A and Polosukhin I. 2017. Attention is All You Need//Advances in Neural Information Processing Systems (NIPS), 5998-6008

Venugopalan S, Rohrbach M, Donahue J, Mooney R, Darrell T and Saenko K. 2015. Sequence to Sequence-Video to Text//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 4534-4542 [DOI:10.1109/ICCV.2015.515]

Wang C, Yang H, Bartz C and Meinel C. 2016. Image Captioning with Deep Bidirectional LSTMs//Proceedings of the ACM International Conference on Multimedia (ACM MM), 988-997 [DOI:10.1145/2964284.2964299]

Wang J, Jiang W, Ma L, Liu W and Xu Y. 2018. Bidirectional Attentive Fusion with Context Gating for Dense Video Captioning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7190-7198 [DOI:10.1109/CVPR.2018.00751]

Wang J, Yang Z, Hu X, Li L, Lin K, Gan Z and Wang L. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language[EB/OL].[2022-12-15]. https://arxiv.org/pdf/2205.14100.pdf

Wang L, Shang C, Qiu H, Zhao T, Qiu B and Li H. 2020. Multi-Stage Tag Guidance Network in Video Caption//Proceedings of the ACM International Conference on Multimedia (ACM MM), 4610-4614 [DOI:10.1145/3394171.3416288]

Wang W, Gao J, Yang X and Xu C. 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, Wu J, Wang Y F and Wang W. 2018. Video Captioning via Hierarchical Reinforcement Learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4213-4222 [DOI:10.1109/CVPR.2018.00443]

Wei X S, Song Y Z, Mac Aodh O, Wu J, Peng Y, Tang 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, Liu J, Wang X and Dong X. 2018. Object-difference Attention: A Simple Relational Attention for Visual Question Answering//Proceedings of the ACM International Conference on Multimedia (ACM MM), 519-527 [DOI:10.1145/3240508.3240513]

Wu Z, Lischinski D and Shechtman E. 2021. Stylespace Analysis: Disentangled Controls for Stylegan Image Generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12863-12872 [DOI:10.1109/CVPR46437.2021.01267]

Xia W, Yang Y, Xue J H and Wu B. 2021. TediGAN: Text-Guided Diverse Face Image Generation and Manipulation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2256-2265 [DOI:10.1109/CVPR46437.2021.00229]

Xie C W, Wu J, Zheng Y, Pan P and Hua X S. 2022. Token Embeddings Alignment for Cross-Modal Retrieval//Proceedings of the ACM International Conference on Multimedia (ACM MM), 4555-4563 [DOI:10.1145/3503161.3548107]

Xiong Y, Dai B and Lin D. 2018. Move Forward and Tell: A Progressive Generator of Video Descriptions//Proceedings of the European Conference on Computer Vision (ECCV), 468-483 [DOI:10.1007/978-3-030-01252-6_29]

Xu H, Yan M, Li C, Bi B, Huang S, Xiao W and Huang F. 2021. E2E-VLP: End-to-End Vision-Language Pre-Training Enhanced by Visual Learning//Proceedings of the Association for Computational Linguistics, 503-513 [DOI:10.18653/v1/2021.acl-long.42]

Xu T, Zhang P, Huang Q, Han Z, Gan Z, Huang X and He X. 2018. AttnGAN: Fine-grained Text to Image Generation with Attentional Generative Adversarial Networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1316-1324 [DOI:10.1109/CVPR.2018.00143]

Xue H, Hang T, Zeng Y, Sun Y, Liu B, Yang H and Guo B. 2022. Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5036-5045 [DOI:10.1109/CVPR52688.2022.00498]

Yao L, Huang R, Hou L, Lu G, Niu M, Xu H and Xu C. 2022. Filip: Fine-grained Interactive Language-image Pre-training//International Conference on Learning Representation (ICLR), 1-21

Yao L, Wang W and Jin Q. 2022. Image Difference Captioning with Pre-training and Contrastive Learning//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 3108-3116 [DOI:10.1609/aaai.v36i3.20218]

Yin Q Y, Huang Y, Zhang J G, Wu S and Wang L. 2021. Survey on deep learning based cross-modal retrieval. Journal of Im

age and Graphics, 26(6): 1368-1388 (尹奇跃, 黄岩, 张俊格, 吴书, 王亮. 2021. 基于深度学习的跨模态检索综述. 中国图象图形学报, 26(6):1368-1388) [DOI: 10.11834/jig.200862]

You Q, Zhang Z and Luo J. 2018. End-to-End Convolutional Semantic Embeddings//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5735-5744 [DOI:10.1109/CVPR.2018.00601]

Yuan L, Chen D, Chen Y L, Codella N, Dai X, Gao J and Zhang P. 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, Zhang S, Wang X, Huang X and Metaxas D. 2017. Stackgan: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 5907-5915 [DOI:10.1109/ICCV.2017.629]

Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X and Metaxas D. 2018. 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 Z, Shi Y, Yuan C, Li B, Wang P, Hu W and Zha Z J. 2020. Object Relational Graph with Teacher-recommended Learning for Video Captioning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13278-13288 [DOI:10.1109/cvpr42600.2020.01329]

Zhang Z, 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, Xie Y and Yang L. 2018. Photographic Text-to-image Synthesis with a Hierarchically-nested Adversarial Network//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6199-6208 [DOI:10.1109/CVPR.2018.00649]

Zhou L, Palangi H, Zhang L, Hu H, Corso J and Gao J. 2020. Unified Vision-Language Pre-Training for Image Captioning and VQA//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 34:13041-13049 [DOI:10.1609/AAAI.V34I07.7005]

Zhou L, Zhou Y, Corso J, Socher R and Xiong C. 2018. End-to-End Dense Video Captioning with Masked Transformer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8739-8748 [DOI:10.1109/CVPR.2018.00911]

Zhu L and Yi Y. 2020. ActBERT: Learning Global-Local Video-Text Representations//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8743-8752 [DOI:10.1109/cvpr42600.2020.00877]

Zhu M, Pan P, Chen W and Yang Y. 2019. DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-image Synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5802-5810 [DOI:10.1109/CVPR.2019.00595]

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):7 (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):7) [DOI:10.13811/j.cnki.eer.2021.11.011]

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]

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 System 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]

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]

作者简介：

刘华峰，1988年生，男，博士后，研究方向为多媒体、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，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、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，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