(IJCV), 2014, 106(2): 210–233. [DOI: 10.1007/s11263-013-0658-4]

[9] Li D, Dimitrova N, Li M, et al. Multimedia content processing through cross-modal association [C]//Proceeding of ACM International Conference on Multimedia (ACM-MM), 2003: 604–611. [DOI: 10.1145/957013.957143]

[10] Peng Y, Zhai X, Zhao Y, et al. Semi-supervised cross-media feature learning with unified patch graph regularization [J]. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2016, 26(3): 583–596. [DOI: 10.1109/TCSVT.2015.2400779]

[11] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning [C]//Proceeding of International Conference on Machine Learning (ICML), 2011: 689–696.

[12] Andrew G, Arora R, Bilmes J A, et al. Deep canonical correlation analysis [C]//Proceeding of International Conference on Machine Learning (ICML), 2013: 1247–1255.

[13] Peng Y, Huang X, and Qi J. Cross-media shared representation by hierarchical learning with multiple deep networks [C]//Proceeding of International Joint Conference on Artificial Intelligence (IJCAI), 2016: 3846–3853.

[14] Wei Y, Zhao Y, Lu C, et al. Cross-modal retrieval with CNN visual features: A new baseline [J]. IEEE Transactions on Cybernetics (TCYB), 2017, 47(2): 449–460. [DOI: 10.1109/TCYB.2016.2519449]

[15] Wang B, Yang Y, Xu X, et al. Adversarial cross-modal retrieval [C]//Proceeding of ACM Conference on Multimedia (ACM-MM), 2017, pp. 154–162. [DOI: 10.1145/3123266.3123326]

[16] Huang X, Peng Y, and Yuan M. Cross-modal Common Representation Learning by Hybrid Transfer Network [C]//Proceeding of International Joint Conference on Artificial Intelligence (IJCAI), 2017: 1893-1900. [DOI: 10.24963/ijcai.2017/263]

[17] Simonyan K and Zisserman A. Very deep convolutional networks for large-scale image recognition [C]//Proceeding of International Conference on Learning Representations (ICLR), 2014.

[18] Lin Y, Pang Z, Wang D, et al. Task-driven visual saliency and attention-based visual question answering [J]. arXiv preprint arXiv:1702.06700, 2017.

[19] Kim Y. Convolutional neural networks for sentence classification [C]//Proceeding of Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1746–1751. [DOI: 10.3115/v1/D14-1181]

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

[21] Lu J, Yang J, Batra D, et al. Hierarchical question-image co-attention for visual question answering [C]//Proceeding of Conference on Neural Information Processing Systems (NIPS), 2016: 289–297.

[22] Rashtchian C, Young P, Hodosh M, et al. Collecting image annotations using amazon's mechanical turk [C]//Proceeding of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010: 139–147.

[23] Kang C, Xiang S, Liao S, et al. Learning consistent feature representation for cross-modal multimedia retrieval [J]. IEEE Transactions on Multimedia (TMM), 2015, 17(3): 370–381. [DOI: 10.1109/TMM.2015.2390499]

第一作者简介：

綦金玮，1994 年生，男，北京大学计算机科学技术研究所硕士研究生，主要研究方向为跨媒体分析与检索。E-mail: qijinwei@pku.edu.cn

通信作者：

彭宇新，男，北京大学计算机科学技术研究所教授、博士生导师，主要研究方向：跨媒体分析与推理；图像、视频理解与检索；机器学习与人工智能。E-mail: pengyuxin@pku.edu.cn

其他作者简介：

袁玉鑫，男，北京大学计算机科学技术研究所硕士研究生，主要研究方向为跨媒体分析与检索。E-mail: yuanyuxin@pku.edu.cn