## REFERENCES


[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision*. 2425–2433.

[2] Yusuf Aytar, O Bilal Orhan, and Mubarak Shah. 2007. Improving semantic concept detection and retrieval using contextual estimates. In *IEEE International Conference on Multimedia and Expo*. 536–539.

[3] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research* 55 (2016), 409–442.

[4] Zhen Cui, Chunyan Xu, Wenming Zheng, and Jian Yang. 2018. Context-dependent diffusion network for visual relationship detection. In *ACM International Conference on Multimedia*. 1475–1482.

[5] Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3076–3086.

[6] Jia Deng, Alexander C Berg, Kai Li, and Li Fei-Fei. 2010. What does classifying more than 10,000 image categories tell us?. In *European Conference on Computer Vision*. 71–84.

[7] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. 2014. Large-scale object classification using label relation graphs. In *European conference on computer vision*. 48–64.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[9] Jia Deng, Jonathan Krause, Alexander C Berg, and Li Fei-Fei. 2012. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3450–3457.

[10] Jia Deng, Sanjeev Satheesh, Alexander C Berg, and Fei Li. 2011. Fast and balanced: Efficient label tree learning for large scale object recognition. In *Advances in Neural Information Processing Systems*. 567–575.

[11] Nan Ding, Jia Deng, Kevin P Murphy, and Hartmut Neven. 2015. Probabilistic label relation graphs with ising models. In *IEEE International Conference on Computer Vision*. 1161–1169.

[12] Ross Girshick. 2015. Fast r-cnn. In *IEEE International Conference on Computer Vision*. 1440–1448.

[13] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. 2016. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*. 241–257.

[14] Seong Jae Hwang, Sathya N Ravi, Zirui Tao, Hyunwoo J Kim, Maxwell D Collins, and Vikas Singh. 2018. Tensorize, factorize and regularize: Robust visual relationship learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1014–1023.

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123 (2017), 32–73.

[16] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017. Scene graph generation from objects, phrases and region captions. In *IEEE International Conference on Computer Vision*. 1261–1270.

[17] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. 2018. Visual relationship detection with deep structural ranking. In *AAAI Conference on Artificial Intelligence*. 7098–7105.

[18] Xiaodan Liang, Lisa Lee, and Eric P Xing. 2017. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 848–857.

[19] Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028* (2002).

[20] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*. 852–869.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. 3111–3119.

[23] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38 (1995), 39–41.

[24] Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. From large scale image categorization to entry-level categories. In *IEEE International Conference on Computer Vision*. 2768–2775.

[25] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. 2017. Video captioning with transferred semantic attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6504–6512.

[26] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7263–7271.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. 91–99.

[28] Mohammad Amin Sadeghi and Ali Farhadi. 2011. Recognition using visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1745–1752.

[29] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *ACM International Conference on Multimedia*. 1300–1308.

[30] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[31] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361* (2015).

[32] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*. 3551–3558.

[33] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5410–5419.

[34] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph r-cnn for scene graph generation. In *European Conference on Computer Vision*. 670–685.

[35] Xu Yang, Hanwang Zhang, and Jianfei Cai. 2018. Shuffle-then-assemble: Learning object-agnostic visual relationship features. In *European Conference on Computer Vision*. 36–52.

[36] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *European Conference on Computer Vision*. 684–699.

[37] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. 2018. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *European Conference on Computer Vision*. 322–338.

[38] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. 2017. Visual relationship detection with internal and external linguistic knowledge distillation. In *IEEE International Conference on Computer Vision*. 1974–1982.

[39] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5831–5840.

[40] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual translation embedding network for visual relation detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5532–5540.

[41] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. 2017. PPR-FCN: Weakly supervised visual relation detection via parallel pairwise R-FCN. In *IEEE International Conference on Computer Vision*. 4233–4241.

[42] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. 2018. Large-scale visual relationship understanding. *arXiv preprint arXiv:1804.10660* (2018).

[43] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. 2017. Open vocabulary scene parsing. In *IEEE International Conference on Computer Vision*. 2002–2010.

[44] Hao Zhou, Chuanping Hu, Chongyang Zhang, and Shengyang Shen. 2019. Visual Relationship Recognition via Language and Position Guided Attention. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2097–2101.

[45] Yaohui Zhu and Shuqiang Jiang. 2018. Deep structured learning for visual relationship detection. In *AAAI Conference on Artificial Intelligence*. 7623–7630.

[46] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. 2017. Towards context-aware interaction recognition for visual relationship detection. In *IEEE International Conference on Computer Vision*. 589–598.