Universiteit Leiden
The Netherlands

# Exploring deep learning for intelligent image retrieval

Chen, W.

**Citation**

Chen, W. (2021, October 13). *Exploring deep learning for intelligent image retrieval*. Retrieved from https://hdl.handle.net/1887/3217054

| | |
|---|---|
| Version: | Publisher's Version |
| License: | Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden |
| Downloaded from: | https://hdl.handle.net/1887/3217054 |

**Note:** To cite this publication please use the final published version (if applicable).

# Bibliography


[1] Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 1349–1380

[2] Zhang, L., Rui, Y.: Image search from thousands to billions in 20 years. ACM Transactions on Multimedia Computing, Communications, and Applications **9** (2013) 36

[3] Zheng, L., Yang, Y., Tian, Q.: SIFT meets CNN: A decade survey of instance retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **40** (2018) 1224–1244

[4] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110

[5] Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: European Conference on Computer Vision. (2006) 404–417

[6] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2005) 886–893

[7] Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2003) 1470–1477

[8] Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. In: Proceedings of the International Conference on Neural Information Processing Systems. (1999) 487–493

[9] Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: European Conference on Computer Vision. (2010) 143–156

[10] Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C.G., Bimbo, A.D.: Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. ACM Computing Surveys (CSUR) **49** (2016) 14

[11] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2009) 248–255

[12] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of the International Conference on Neural Information Processing Systems. (2012) 1097–1105

[13] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2016) 770–778

[14] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 4700–4708

[15] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2014) 580–587

[16] Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE/CVF International Conference on

Computer Vision. (2015) 1440–1448

[17] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the International Conference on Neural Information Processing Systems. (2015) 91–99

[18] Wang, X., Shrivastava, A., Gupta, A.: A-fast-RCNN: Hard positive generation via adversary for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017)

[19] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40** (2018) 834–848

[20] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2015) 3431–3440

[21] Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39** (2017) 2481–2495

[22] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning. (2016) 1060–1069

[23] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 5907–5915

[24] Gregor, K., Danihelka, I., Graves, A., Rezende, D., Wierstra, D.: DRAW: A recurrent neural network for image generation. In: International Conference on Machine Learning, PMLR (2015) 1462–1471

[25] Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2015) 1269–1277

[26] Kalantidis, Y., Mellina, C., Osindero, S.: Cross-dimensional weighting for aggregated deep convolutional features. In: European Conference on Computer Vision. (2016) 685–701

[27] Wan, J., Wang, D., Hoi, S.C.H., Wu, P., Zhu, J., Zhang, Y., Li, J.: Deep learning for content-based image retrieval: A comprehensive study. In: Proceedings of the ACM International Conference on Multimedia. (2014) 157–166

[28] Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2010) 3304–3311

[29] Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: European Conference on Computer Vision. (2014) 392–407

[30] Yue-Hei Ng, J., Yang, F., Davis, L.S.: Exploiting local features from deep networks for image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. (2015) 53–61

[31] Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. In: International Conference on Learning Representations. (2015) 1–12

[32] Radenović, F., Tolias, G., Chum, O.: CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In: European Conference on Computer Vision. (2016) 3–20

[33] Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22** (1951) 79–86

[34] Chen, W., Pu, N., Liu, Y., Bakker, E.M., Lew, M.S.: Domain uncertainty based on information theory for cross-modal hash retrieval. In: IEEE International Conference on Multimedia and Expo. (2019) 43–48

[35] Chen, W., Liu, Y., Bakker, E.M., Lew, M.S.: Integrating information theory and adversarial learning for cross-modal retrieval. Pattern Recognition **117** (2021) 107983

[36] McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation. (1989) 109–165

[37] Chen, W., Liu, Y., Wang, W., Tuytelaars, T., Bakker, E.M., Lew, M.S.: On the exploration of incremental learning for fine-grained image retrieval. In: The British Machine Vision Conference. (2020) 1–10

[38] Chen, W., Liu, Y., Pu, N., Wang, W., Liu, L., Lew, M.S.: Feature estimations based correlation distillation for incremental image retrieval. IEEE Transactions on Multimedia (2021)

[39] Chen, W., Wang, W., Liu, L., Lew, M.: New ideas and trends in deep multimodal content understanding: A review. Neurocomputing (2020) 195–215

[40] Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications **2** (2006) 1–19

[41] Cao, Y., Long, M., Wang, J., Zhu, H., Wen, Q.: Deep quantization network for efficient image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2016) 3457–3463

[42] Alzu'bi, A., Amira, A., Ramzan, N.: Semantic content-based image retrieval: A comprehensive study. Journal of Visual Communication and Image Representation **32** (2015) 20–54

[43] Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. (2014) 806–813

[44] Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Proceedings of the International Conference on Neural Information Processing Systems. (2014) 3320–3328

[45] Jiménez, A., Alvarez, J.M., Giró Nieto, X.: Class-weighted convolutional features for visual instance search. In: The British Machine Vision Conference. (2017) 1–12

[46] Do, T.T., Hoang, T., Tan, D.K.L., Le, H., Nguyen, T.V., Cheung, N.M.: From selective deep convolutional features to compact binary representations for image retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications **15** (2019) 1–22

[47] Xu, J., Wang, C., Qi, C., Shi, C., Xiao, B.: Unsupervised part-based weighting aggregation of deep convolutional features for image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2018) 7436–7443

[48] Liu, Y., Guo, Y., Wu, S., Lew, M.S.: Deepindex for accurate and efficient image retrieval. In: Proceedings of the ACM on International Conference on Multimedia Retrieval. (2015) 43–50

[49] Wu, P., Hoi, S.C., Xia, H., Zhao, P., Wang, D., Miao, C.: Online multimodal deep similarity learning with application to image retrieval. In: Proceedings of the ACM International Conference on Multimedia. (2013) 153–162

[50] Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: European Conference on Computer Vision. (2014) 584–599

[51] Huang, C.Q., Yang, S.M., Pan, Y., Lai, H.J.: Object-location-aware hashing for multi-label image retrieval via automatic mask learning. IEEE Transactions on Image Processing **27** (2018) 4490–4502

[52] Garcia, N., Vogiatzis, G.: Learning non-metric visual similarity for image retrieval. Image and Vision Computing **82** (2019) 18–25

[53] Ong, E.J., Husain, S., Bober, M.: Siamese network of deep fisher-vector descriptors for image retrieval. arXiv:1702.00338 (2017)

[54] Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: European Conference on Computer Vision. (2016) 241–257

[55] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture

for weakly supervised place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2016) 5297–5307

[56] Xu, J., Wang, C., Qi, C., Shi, C., Xiao, B.: Iterative manifold embedding layer learned by incomplete data for large-scale image retrieval. IEEE Transactions on Multimedia **21** (2018) 1551–1562

[57] Radenović, F., Tolias, G., Chum, O.: Fine-tuning CNN image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence **41** (2017) 1655–1668

[58] Liu, C., Yu, G., Volkovs, M., Chang, C., Rai, H., Ma, J., Gorti, S.K.: Guided similarity separation for image retrieval. In: Proceedings of the International Conference on Neural Information Processing Systems. (2019) 1554–1564

[59] Chang, C., Yu, G., Liu, C., Volkovs, M.: Explore-exploit graph traversal for image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 9423–9431

[60] Shen, Y., Qin, J., Chen, J., Yu, M., Liu, L., Zhu, F., Shen, F., Shao, L.: Auto-encoding twin-bottleneck hashing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 2818–2827

[61] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)

[62] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2015) 1–9

[63] Razavian, A.S., Sullivan, J., Carlsson, S., Maki, A.: Visual instance retrieval with deep convolutional networks. ITE Transactions on Media Technology and Applications (2016) 251–258

[64] Jun, H., Ko, B., Kim, Y., Kim, I., Kim, J.: Combination of multiple global descriptors for image retrieval. arXiv:1903.10663 (2019)

[65] Li, Y., Kong, X., Zheng, L., Tian, Q.: Exploiting hierarchical activations of neural network for image retrieval. In: Proceedings of the ACM International Conference on Multimedia. (2016) 132–136

[66] Qi, C., Shi, C., Xu, J., Wang, C., Xiao, B.: Spatial weighted fisher vector for image retrieval. In: IEEE International Conference on Multimedia and Expo. (2017) 463–468

[67] Mohedano, E., McGuinness, K., Giró-i Nieto, X., O'Connor, N.E.: Saliency weighted convolutional features for instance search. In: International Conference on Content-based Multimedia Indexing. (2018) 1–6

[68] Yang, F., Li, J., Wei, S., Zheng, Q., Liu, T., Zhao, Y.: Two-stream attentive CNNs for image retrieval. In: Proceedings of the ACM International Conference on Multimedia. (2017) 1513–1521

[69] Deng, C., Yang, E., Liu, T., Li, J., Liu, W., Tao, D.: Unsupervised semantic-preserving adversarial hashing for image search. IEEE Transactions on Image Processing **28** (2019) 4032–4044

[70] Hu, H., Wang, K., Lv, C., Wu, J., Yang, Z.: Semi-supervised metric learning-based anchor graph hashing for large-scale image retrieval. IEEE Transactions on Image Processing (2018) 739–754

[71] Deng, D., Wang, R., Wu, H., He, H., Li, Q., Luo, X.: Learning deep similarity models with focus ranking for fabric image retrieval. Image and Vision Computing **70** (2018) 11–20

[72] Zhou, K., Liu, Y., Song, J., Yan, L., Zou, F., Shen, F.: Deep self-taught hashing for image retrieval. In: Proceedings of the ACM International Conference on Multimedia. (2015) 1215–1218

[73] Yan, K., Wang, Y., Liang, D., Huang, T., Tian, Y.: CNN vs. SIFT for image retrieval: Alternative or complementary? In: Proceedings of the ACM International Conference on

Multimedia. (2016) 407–411

[74] Wei, X.S., Luo, J.H., Wu, J., Zhou, Z.H.: Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Transactions on Image Processing **26** (2017) 2868–2881

[75] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. In: The British Machine Vision Conference. (2014)

[76] Piras, L., Giacinto, G.: Information fusion in content based image retrieval: A comprehensive overview. Information Fusion **37** (2017) 50–60

[77] Wang, J., Zhang, T., Sebe, N., Shen, H.T., et al.: A survey on learning to hash. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018) 769–790

[78] Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2014) 1717–1724

[79] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: A survey. International Journal of Computer Vision **128** (2020) 261–318

[80] Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S.: A survey of the recent architectures of deep convolutional neural networks. Artificial Intelligence Review **53** (2020) 5455–5516

[81] Azizpour, H., Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: Factors of transferability for a generic convnet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **38** (2016) 1790–1802

[82] Mohedano, E., McGuinness, K., O'Connor, N.E., Salvador, A., Marqués, F., Giro-i Nieto, X.: Bags of local convolutional features for scalable instance search. In: Proceedings of the ACM on International Conference on Multimedia Retrieval. (2016) 327–331

[83] Sharif Razavian, A., Sullivan, J., Maki, A., Carlsson, S.: A baseline for visual instance retrieval with deep convolutional networks. In: International Conference on Learning Representations. (2015)

[84] Cao, J., Liu, L., Wang, P., Huang, Z., Shen, C., Shen, H.T.: Where to focus: Query adaptive matching for instance retrieval using convolutional feature maps. arXiv:1606.06811 (2016)

[85] Reddy Mopuri, K., Venkatesh Babu, R.: Object level deep feature pooling for compact image representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. (2015) 62–70

[86] Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: European Conference on Computer Vision. (2014) 391–405

[87] Mairal, J., Koniusz, P., Harchaoui, Z., Schmid, C.: Convolutional kernel networks. In: Proceedings of the International Conference on Neural Information Processing Systems. (2014) 2627–2635

[88] Song, J., Yu, Q., Song, Y.Z., Xiang, T., Hospedales, T.M.: Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 5552–5561

[89] Salvador, A., Giró-i Nieto, X., Marqués, F., Satoh, S.: Faster R-CNN features for instance search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. (2016) 9–16

[90] Ng, T., Balntas, V., Tian, Y., Mikolajczyk, K.: SOLAR: Second-order loss and attention for image retrieval. In: European Conference on Computer Vision. (2020) 253–270

[91] Yu, D., Liu, Y., Pang, Y., Li, Z., Li, H.: A multi-layer deep fusion convolutional neural network for sketch based image retrieval. Neurocomputing **296** (2018) 23–32

[92] Yu, W., Yang, K., Yao, H., Sun, X., Xu, P.: Exploiting the complementary strengths of multi-layer CNN features for image retrieval. Neurocomputing **237** (2017) 235–241

[93] Shen, C., Zhou, C., Jin, Z., Chu, W., Jiang, R., Chen, Y., Hua, X.S.: Learning feature embedding with strong neural activations for fine-grained retrieval. In: Proceedings of the

ACM International Conference on Multimedia. (2017) 424–432

[94] Ding, Z., Song, L., Zhang, X., Xu, Z.: Selective deep ensemble for instance retrieval. Multimedia Tools and Applications (2018) 1–17

[95] Kim, W., Goyal, B., Chawla, K., Lee, J., Kwon, K.: Attention-based ensemble for deep metric learning. In: European Conference on Computer Vision. (2018) 736–751

[96] Bui, T., Ribeiro, L., Ponti, M., Collomosse, J.: Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. Computers & Graphics **71** (2018) 77–87

[97] Ozaki, K., Yokoo, S.: Large-scale landmark retrieval/recognition under a noisy and diverse dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop. (2019)

[98] Xuan, H., Souvenir, R., Pless, R.: Deep randomized ensembles for metric learning. In: European Conference on Computer Vision. (2018) 723–734

[99] Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K., O'Connor, N.E.: Shallow and deep convolutional networks for saliency prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2016) 598–606

[100] Chen, B.C., Davis, L.S., Lim, S.N.: An analysis of object embeddings for image retrieval. arXiv:1905.11903 (2019)

[101] Boureau, Y.L., Ponce, J., LeCun, Y.: A theoretical analysis of feature pooling in visual recognition. In: International Conference on Machine Learning. (2010) 111–118

[102] Wang, F., Zhao, W.L., Ngo, C.W., Merialdo, B.: A hamming embedding kernel with informative bag-of-visual words for video semantic indexing. ACM Transactions on Multimedia Computing, Communications, and Applications **10** (2014) 1–20

[103] Jegou, H., Perronnin, F., Douze, M., Sánchez, J., Perez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 1704–1716

[104] Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: Theory and practice. International Journal of Computer Vision **105** (2013) 222–245

[105] Li, R., Jia, J.: Visual question answering with question representation update (QRU). In: Proceedings of the International Conference on Neural Information Processing Systems. (2016) 4655–4663

[106] Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Largescale image retrieval with attentive deep local features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 3456–3465

[107] Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting human eye fixations via an LSTM-based saliency attentive model. IEEE Transactions on Image Processing **27** (2018) 5142–5154

[108] Cao, J., Huang, Z., Shen, H.T.: Local deep descriptors in bag-of-words for image retrieval. In: Proceedings of the ACM International Conference on Multimedia. (2017) 52–58

[109] Kim, J., Yoon, S.E.: Regional attention based deep feature for image retrieval. In: The British Machine Vision Conference. (2018) 209–223

[110] Chen, B., Deng, W.: Hybrid-attention based decoupled metric learning for zero-shot image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 2750–2759

[111] Deng, C., Yang, E., Liu, T., Tao, D.: Two-stream deep hashing with class-specific centers for supervised image search. IEEE Transactions on Neural Networks and Learning Systems (2019)

[112] Liu, L., Shen, F., Shen, Y., Liu, X., Shao, L.: Deep sketch hashing: Fast free-hand sketch-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 2862–2871

[113] Kang, R., Cao, Y., Long, M., Wang, J., Yu, P.S.: Maximum-margin hamming hashing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 8252–8261

[114] Liu, H., Wang, R., Shan, S., Chen, X.: Deep supervised hashing for fast image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2016) 2064–2072

[115] Zhao, F., Huang, Y., Wang, L., Tan, T.: Deep semantic ranking based hashing for multi-label image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2015) 1556–1564

[116] Long, F., Yao, T., Dai, Q., Tian, X., Luo, J., Mei, T.: Deep domain adaptation hashing with adversarial learning. In: The International ACM SIGIR Conference on Research & Development in Information Retrieval. (2018) 725–734

[117] Cao, Y., Liu, B., Long, M., Wang, J., KLiss, M.: HashGAN: Deep learning to hash with pair conditional wasserstein GAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 1287–1296

[118] Yang, E., Liu, T., Deng, C., Liu, W., Tao, D.: DistillHash: Unsupervised deep hashing by distilling data pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 2946–2955

[119] Carreira-Perpinán, M.A., Raziperchikolaei, R.: Hashing with binary autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2015) 557–566

[120] Do, T.T., Le Tan, D.K., Pham, T.T., Cheung, N.M.: Simultaneous feature aggregating and hashing for large-scale image search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 6618–6627

[121] Gu, Y., Wang, S., Zhang, H., Yao, Y., Yang, W., Liu, L.: Clustering-driven unsupervised deep hashing for image retrieval. Neurocomputing **368** (2019) 114–123

[122] Song, J.: Binary generative adversarial networks for image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2017)

[123] Dizaji, K.G., Zheng, F., Nourabadi, N.S., Yang, Y., Deng, C., Huang, H.: Unsupervised deep generative adversarial hashing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 3664–3673

[124] Erin Liong, V., Lu, J., Wang, G., Moulin, P., Zhou, J.: Deep hashing for compact binary codes learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2015) 2475–2483

[125] Cakir, F., He, K., Bargal, S.A., Sclaroff, S.: Hashing with mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence **41** (2019) 2424–2437

[126] Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: European Conference on Computer Vision. (2008) 304–317

[127] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2007) 1–8

[128] Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 403–412

[129] Wang, J., Zhou, F., Wen, S., Liu, X., Lin, Y.: Deep metric learning with angular loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 2593–2601

[130] Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: Proceedings of the International Conference on Neural Information Processing Systems. (2016) 1857–1865

[131] Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured

feature embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2016) 4004–4012

[132] Wang, X., Hua, Y., Kodirov, E., Hu, G., Garnier, R., Robertson, N.M.: Ranked list loss for deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 5207–5216

[133] Chen, L., He, Y.: Dress fashionably: Learn fashion collocation with deep mixed-category metric learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2018) 2103–2110

[134] Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 360–368

[135] Kim, S., Kim, D., Cho, M., Kwak, S.: Proxy anchor loss for deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 3238–3247

[136] Zheng, W., Chen, Z., Lu, J., Zhou, J.: Hardness-aware deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 72–81

[137] Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2014) 1386–1393

[138] Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., Moreno-Noguer, F.: Discriminative learning of deep convolutional feature point descriptors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2015) 118–126

[139] Song, J., He, T., Gao, L., Xu, X., Shen, H.T.: Deep region hashing for efficient large-scale instance search from images. arXiv:1701.07901 (2017)

[140] Gordo, A., Almazan, J., Revaud, J., Larlus, D.: End-to-end learning of deep visual representations for image retrieval. International Journal of Computer Vision **124** (2017) 237–254

[141] Lin, J., Morere, O., Veillard, A., Duan, L.Y., Goh, H., Chandrasekhar, V.: Deephash for image instance retrieval: Getting regularization, depth and fine-tuning right. In: Proceedings of the ACM on International Conference on Multimedia Retrieval. (2017) 133–141

[142] Cao, J., Huang, Z., Wang, P., Li, C., Sun, X., Shen, H.T.: Quartet-net learning for visual instance retrieval. In: Proceedings of the ACM International Conference on Multimedia. (2016) 456–460

[143] Wang, X., Zhang, H., Huang, W., Scott, M.R.: Cross-batch memory for embedding learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 6388–6397

[144] Harwood, B., Kumar, B., Carneiro, G., Reid, I., Drummond, T., et al.: Smart mining for deep metric learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 2821–2829

[145] He, K., Lu, Y., Sclaroff, S.: Local descriptors optimized for average precision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 596–605

[146] Revaud, J., Almazán, J., Rezende, R.S., Souza, C.R.d.: Learning with average precision: Training image retrieval with a listwise loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 5107–5116

[147] Brown, A., Xie, W., Kalogeiton, V., Zisserman, A.: Smooth-AP: Smoothing the path towards large-scale image retrieval. In: European Conference on Computer Vision. (2020) 677–694

[148] Aziere, N., Todorovic, S.: Ensemble deep manifold similarity learning using hard proxies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 7299–7307

[149] Donoser, M., Bischof, H.: Diffusion processes for retrieval revisited. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2013) 1320–1327

[150] Iscen, A., Tolias, G., Avrithis, Y., Furon, T., Chum, O.: Efficient diffusion on region mani-

folds: Recovering small objects with compact CNN representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 2077–2086

[151] Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Mining on manifolds: Metric learning without labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 7642–7651

[152] Zhao, Y., Wang, L., Zhou, L., Shi, Y., Gao, Y.: Modelling diffusion process by deep neural networks for image retrieval. In: The British Machine Vision Conference. (2018) 161–174

[153] Song, B., Bai, X., Tian, Q., Latecki, L.J.: Regularized diffusion process on bidirectional context for object retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **41** (2018) 1213–1226

[154] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations. (2017)

[155] Maria, T., Anastasios, T.: Deep convolutional image retrieval: A general framework. Signal Processing: Image Communication **63** (2018) 30–43

[156] Tu, R.C., Mao, X.L., Feng, B.S., Yu, S.Y.: Object detection based deep unsupervised hashing. In: International Joint Conference on Artificial Intelligence. (2019) 3606–3612

[157] Zieba, M., Semberecki, P., El-Gaaly, T., Trzcinski, T.: BinGAN: learning compact binary descriptors with a regularized GAN. In: Proceedings of the International Conference on Neural Information Processing Systems. (2018) 3612–3622

[158] Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2006) 2161–2168

[159] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2008) 1–8

[160] Radenovic, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018)

[161] Zheng, L., Wang, S., Wang, J., Tian, Q.: Accurate image search with multi-scale contextual evidences. International Journal of Computer Vision **120** (2016) 1–13

[162] Alzu'bi, A., Amira, A., Ramzan, N.: Content-based image retrieval with compact deep convolutional features. Neurocomputing **249** (2017) 95–105

[163] Valem, L.P., Pedronette, D.C.G.: Graph-based selective rank fusion for unsupervised image retrieval. Pattern Recognition Letters (2020)

[164] Alemu, L.T., Pelillo, M.: Multi-feature fusion for image retrieval using constrained dominant sets. Image and Vision Computing **94** (2020) 103862

[165] Yang, F., Hinami, R., Matsui, Y., Ly, S., Satoh, S.: Efficient image retrieval via decoupling diffusion into online and offline processing. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2019) 9087–9094

[166] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: Integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations. (2014)

[167] Husain, S.S., Bober, M.: REMAP: Multi-layer entropy-guided pooling of dense CNN features for image retrieval. IEEE Transactions on Image Processing **28** (2019) 5201–5213

[168] Iscen, A., Avrithis, Y., Tolias, G., Furon, T., Chum, O.: Fast spectral ranking for similarity search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 7632–7641

[169] Yang, J., Liang, J., Shen, H., Wang, K., Rosin, P.L., Yang, M.H.: Dynamic match kernel with deep convolutional features for image retrieval. IEEE Transactions on Image Processing **27** (2018) 5288–5302

[170] Yang, H.F., Lin, K., Chen, C.S.: Cross-batch reference learning for deep classification and

retrieval. In: Proceedings of the ACM International Conference on Multimedia. (2016) 1237–1246

[171] Lv, Y., Zhou, W., Tian, Q., Sun, S., Li, H.: Retrieval oriented deep feature learning with complementary supervision mining. IEEE Transactions on Image Processing **27** (2018) 4945–4957

[172] Wang, Q., Lai, J., Claesen, L., Yang, Z., Lei, L., Liu, W.: A novel feature representation: Aggregating convolution kernels for image retrieval. Neural Networks **130** (2020) 1–10

[173] Jiang, Q.Y., Li, W.J.: Deep cross-modal hashing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 3232–3240

[174] Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. (2013) 785–796

[175] Chi, J., Peng, Y.: Dual adversarial networks for zero-shot cross-media retrieval. In: International Joint Conference on Artificial Intelligence. (2018) 663–669

[176] Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L.: A comprehensive survey on cross-modal retrieval. arXiv:1607.06215 (2016)

[177] Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: Proceedings of the ACM International Conference on Multimedia. (2017) 154–162

[178] Li, C., Deng, C., Li, N., Liu, W., Gao, X., Tao, D.: Self-supervised adversarial hashing networks for cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 4242–4251

[179] Shannon, C.E.: A mathematical theory of communication. Bell system technical journal **27** (1948) 379–423

[180] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the International Conference on Neural Information Processing Systems. (2014) 2672–2680

[181] Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. arXiv:1409.7495 (2014)

[182] Huiskes, M.J., Lew, M.S.: The MIR Flickr retrieval evaluation. In: Proceedings of the ACM International Conference on Image and Video Retrieval. (2008) 39–43

[183] Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from national university of singapore. In: Proceedings of the ACM International Conference on Image and Video Retrieval. (2009) 48

[184] Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: Improving visual-semantic embeddings with hard negatives. In: The British Machine Vision Conference. (2018) 1–10

[185] Wang, D., Wang, Q., He, L., Gao, X., Tian, Y.: Joint and individual matrix factorization hashing for large-scale cross-modal retrieval. Pattern Recognition (2020) 107479

[186] Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: European Conference on Computer Vision. (2018) 686–701

[187] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861 (2017)

[188] Graves, A., Fernández, S., Schmidhuber, J.: Bidirectional LSTM networks for improved phoneme classification and recognition. In: International Conference on Artificial Neural Networks. (2005) 799–804

[189] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. (2017) 1321–1330

[190] Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research **47** (2013) 853–899

[191] Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2** (2014) 67–78

[192] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision. (2014) 740–755

[193] Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 1970–1979

[194] Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2016) 5005–5013

[195] Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of machine learning research **9** (2008)

[196] Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-RNN). In: International Conference on Learning Representations. (2015)

[197] Lev, G., Sadeh, G., Klein, B., Wolf, L.: RNN fisher vectors for action recognition and image annotation. In: European Conference on Computer Vision. (2016) 833–850

[198] Dong, J., Li, X., Snoek, C.G.: Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia **20** (2018) 3377–3388

[199] Huang, Y., Wang, W., Wang, L.: Instance-aware image and sentence matching with selective multimodal LSTM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 2310–2318

[200] Liu, Y., Guo, Y., Bakker, E.M., Lew, M.S.: Learning a recurrent residual fusion network for multimodal matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 4127–4136

[201] Sarafianos, N., Xu, X., Kakadiaris, I.A.: Adversarial representation learning for text-to-image matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 5814–5824

[202] Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 299–307

[203] Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., Shen, Y.D.: Dual-path convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications **16** (2020) 1–23

[204] Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-aware textual-visual matching with latent co-attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 1890–1899

[205] Chen, D., Li, H., Liu, X., Shen, Y., Shao, J., Yuan, Z., Wang, X.: Improving deep visual representation for person re-identification by global and local image-language association. In: European Conference on Computer Vision. (2018) 54–70

[206] Niu, K., Huang, Y., Ouyang, W., Wang, L.: Improving description-based person re-identification by multi-granularity image-text alignments. IEEE Transactions on Image Processing **29** (2020) 5542–5556

[207] Klein, B., Lev, G., Sadeh, G., Wolf, L.: Associating neural word embeddings with deep image representations using fisher vectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2015) 4437–4446

[208] Bousquet, O., Elisseeff, A.: Stability and generalization. The Journal of Machine Learning Research **2** (2002) 499–526

[209] Yu, B.: Stability. Bernoulli **19** (2013) 1484–1500

[210] Sun, W.: Stability of machine learning algorithms. PhD thesis, Purdue University (2015)

[211] Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning. (2015) 1180–1189

[212] Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40** (2017) 2935–2947

[213] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114** (2017) 3521–3526

[214] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015)

[215] Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., Sriperumbudur, B.K.: Optimal kernel choice for large-scale two-sample tests. In: Proceedings of the International Conference on Neural Information Processing Systems. (2012) 1205–1213

[216] Zhai, M., Chen, L., Tung, F., He, J., Nawhal, M., Mori, G.: Lifelong GAN: Continual learning for conditional image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 2759–2768

[217] Shmelkov, K., Schmid, C., Alahari, K.: Incremental learning of object detectors without catastrophic forgetting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 3400–3409

[218] Wu, D., Dai, Q., Liu, J., Li, B., Wang, W.: Deep incremental hashing network for efficient image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 9069–9077

[219] Michieli, U., Zanuttigh, P.: Incremental learning techniques for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. (2019)

[220] Hou, S., Pan, X., Change Loy, C., Wang, Z., Lin, D.: Lifelong learning via progressive distillation and retrospection. In: European Conference on Computer Vision. (2018) 437–452

[221] Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. In: Proceedings of the International Conference on Neural Information Processing Systems. (2017) 2990–2999

[222] Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. In: Proceedings of the International Conference on Neural Information Processing Systems. (2016) 136–144

[223] Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford Dogs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop. (2011)

[224] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 dataset. (2011)

[225] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)

[226] Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 5022–5030

[227] Park, D., Hong, S., Han, B., Lee, K.M.: Continual learning by asymmetric loss approximation with single-side overestimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 3335–3344

[228] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. (2009)

[229] Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Fu, Y.: Large scale incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition. (2019) 374–382

[230] Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. In: Proceedings of the International Conference on Neural Information Processing Systems. (2017) 6467–6476

[231] van de Ven, G.M., Tolias, A.S.: Generative replay with feedback connections as a general strategy for continual learning. arXiv:1809.10635 (2018)

[232] Yao, X., Huang, T., Wu, C., Zhang, R.X., Sun, L.: Adversarial feature alignment: Avoid catastrophic forgetting in incremental task lifelong learning. Neural Computation **31** (2019) 2266–2291

[233] Parshotam, K., Kilickaya, M.: Continual learning of object instances. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. (2020) 224–225

[234] Xie, L., Wang, J., Zhang, B., Tian, Q.: Fine-grained image search. IEEE Transactions on Multimedia **17** (2015) 636–647

[235] Zhou, P., Mai, L., Zhang, J., Xu, N., Wu, Z., Davis, L.S.: M2KD: Multi-model and multi-level knowledge distillation for incremental learning. In: The British Machine Vision Conference. (2020) 1–10

[236] Tian, X., Ng, W., Wang, H., Kwong, S.: Complementary incremental hashing with query-adaptive re-ranking for image retrieval. IEEE Transactions on Multimedia (2020) 1–15

[237] Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision (2021)

[238] Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 4133–4141

[239] Huang, X., Peng, Y.: TPCKT: two-level progressive cross-media knowledge transfer. IEEE Transactions on Multimedia **21** (2019) 2850–2862

[240] Ma, X., Zhang, T., Xu, C.: Multi-level correlation adversarial hashing for cross-modal retrieval. IEEE Transactions on Multimedia **22** (2020) 3101–3114

[241] Peng, Y., Qi, J.: Show and tell in the loop: Cross-modal circular correlation learning. IEEE Transactions on Multimedia **21** (2018) 1538–1550

[242] Peng, Y., Qi, J., Huang, X., Yuan, Y.: CCL: Cross-modal correlation learning with multi-grained fusion by hierarchical network. IEEE Transactions on Multimedia **20** (2017) 405–420

[243] Li, Z., Tang, J., Mei, T.: Deep collaborative embedding for social image understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **41** (2018) 2070–2083

[244] Peng, Z., Li, Z., Zhang, J., Li, Y., Qi, G.J., Tang, J.: Few-shot image recognition with knowledge transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 441–449

[245] Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 1365–1374

[246] Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: European Conference on Computer Vision. (2018) 532–547

[247] Upchurch, P., Gardner, J., Pleiss, G., Pless, R., Snavely, N., Bala, K., Weinberger, K.: Deep feature interpolation for image content changes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 7064–7073

[248] Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2010) 117–128

[249] Aljundi, R., Chakravarty, P., Tuytelaars, T.: Expert gate: Lifelong learning with a network of experts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 3366–3375

[250] Perez-Rua, J.M., Zhu, X., Hospedales, T.M., Xiang, T.: Incremental few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 13846–13855

[251] Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks **113** (2019) 54–71

[252] French, R.M.: Catastrophic forgetting in connectionist networks. Trends in cognitive sciences **3** (1999) 128–135

[253] Yu, L., Yazici, V.O., Liu, X., Weijer, J.v.d., Cheng, Y., Ramisa, A.: Learning metrics from teachers: Compact networks for image embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 2907–2916

[254] Lu, J., Hu, J., Zhou, J.: Deep metric learning for visual understanding: An overview of recent advances. IEEE Signal Processing Magazine **34** (2017) 76–84

[255] Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J.: Dreaming to distill: Data-free knowledge transfer via deepinversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 8715–8724

[256] Haroush, M., Hubara, I., Hoffer, E., Soudry, D.: The knowledge within: Methods for data-free model compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 8494–8502

[257] Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 5007–5016

[258] Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 3967–3976

[259] Rannen, A., Aljundi, R., Blaschko, M.B., Tuytelaars, T.: Encoder based lifelong learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 1320–1328

[260] Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D object representations for fine-grained categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. (2013) 554–561

[261] Wei, K., Deng, C., Yang, X.: Lifelong zero-shot learning. In: International Joint Conference on Artificial Intelligence. (2020) 551–557

[262] Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Proceedings of the International Conference on Neural Information Processing Systems. (2016)

[263] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the International Conference on Neural Information Processing Systems. (2017)

[264] Park, C.C., Kim, Y., Kim, G.: Retrieval of sentence sequences for an image stream via coherence recurrent convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **40** (2018) 945–957

[265] Liang, J., Jiang, L., Cao, L., Kalantidis, Y., Li, L.J., Hauptmann, A.G.: Focal visual-text attention for memex question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence **41** (2019) 1893–1908

[266] Chen, H., Ding, G., Lin, Z., Zhao, S., Han, J.: Show, observe and tell: Attribute-driven attention model for image captioning. In: International Joint Conference on Artificial Intelligence. (2018) 606–612

[267] Cha, M., Gwon, Y.L., Kung, H.: Adversarial learning of semantic relevance in text to image synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 3272–3279

[268] Gorti, S.K., Ma, J.: Text-to-image-to-text translation using cycle consistent adversarial networks. arXiv:1808.04538 (2018)

[269] Wu, L., Wang, Y., Shao, L.: Cycle-consistent deep generative hashing for cross-modal retrieval. IEEE Transactions on Image Processing **28** (2018) 1602–1612

[270] Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: European Conference on Computer Vision. (2018) 684–699

[271] Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 10685–10694

[272] Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 7008–7024

[273] Liu, D., Zha, Z.J., Zhang, H., Zhang, Y., Wu, F.: Context-aware visual policy network for sequence-level image captioning. arXiv:1808.05864 (2018)

[274] Hossain, M., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CSUR) **51** (2019) 118

[275] Wang, H., Wang, H., Xu, K.: Evolutionary recurrent neural network for image captioning. Neurocomputing (2020)

[276] Dai, B., Fidler, S., Urtasun, R., Lin, D.: Towards diverse and natural image descriptions via a conditional GAN. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 2970–2979

[277] Sukhbaatar, S., Weston, J., Fergus, R., et al.: End-to-end memory networks. In: Proceedings of the International Conference on Neural Information Processing Systems. (2015) 2440–2448

[278] Park, C.C., Kim, B., Kim, G.: Towards personalized image captioning via multimodal memory networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)

[279] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 6077–6086

[280] Song, L., Liu, J., Qian, B., Chen, Y.: Connecting language to images: A progressive attention-guided network for simultaneous image captioning and language grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 8885–8892

[281] Jin, J., Nakayama, H.: Annotation order matters: Recurrent image annotator for arbitrary length image tagging. In: International Conference on Pattern Recognition. (2016) 2452–2457

[282] Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 1261–1270

[283] Anderson, P., Gould, S., Johnson, M.: Partially-supervised image captioning. In: Proceedings of the International Conference on Neural Information Processing Systems. (2018) 1879–1890

[284] El, O.B., Licht, O., Yosephian, N.: GILT: Generating images from long text. arXiv:1901.02404 (2019)

[285] Mansimov, E., Parisotto, E., Ba, J.L., Salakhutdinov, R.: Generating images from captions with attention. International Conference on Learning Representations (2016)

[286] Reed, S., van den Oord, A., Kalchbrenner, N., Colmenarejo, S.G., Wang, Z., Chen, Y., Belov, D., de Freitas, N.: Parallel multiscale autoregressive density estimation. In: International Conference on Machine Learning. (2017) 2912–2921

[287] Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 1219–1228

[288] Hong, S., Yang, D., Choi, J., Lee, H.: Inferring semantic layout for hierarchical text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 7986–7994

[289] Zhang, Z., Xie, Y., Yang, L.: Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 6199–6208

[290] Gao, L., Chen, D., Song, J., Xu, X., Zhang, D., Shen, H.T.: Perceptual pyramid adversarial networks for text-to-image synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2019) 8312–8319

[291] Han, Z., Tao, X., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018) 1–1

[292] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 1316–1324

[293] Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: Proceedings of the International Conference on Neural Information Processing Systems. (2016) 217–225

[294] Yuan, M., Peng, Y.: Text-to-image synthesis via symmetrical distillation networks. In: Proceedings of the ACM International Conference on Multimedia. (2018) 1407–1415

[295] Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2016) 49–58

[296] Zhang, S., Dong, H., Hu, W., Guo, Y., Wu, C., Xie, D., Wu, F.: Text-to-image synthesis via visual-memory creative adversarial network. In: Pacific Rim Multimedia. (2018) 417–427

[297] Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: Proceedings of the International Conference on Neural Information Processing Systems. (2017) 3856–3866

[298] Song, G., Wang, D., Tan, X.: Deep memory network for cross-modal retrieval. IEEE Transactions on Multimedia **21** (2018) 1261–1275

[299] Zhang, M., Zhang, H., Li, J., Wang, L., Fang, Y., Sun, J.: Supervised graph regularization based cross media retrieval with intra and inter-class correlation. Journal of Visual Communication and Image Representation **58** (2019) 1–11

[300] Wu, Y., Wang, S., Song, G., Huang, Q.: Augmented adversarial training for cross-modal retrieval. IEEE Transactions on Multimedia (2020)

[301] Wang, Y., Luo, X., Nie, L., Song, J., Zhang, W., Xu, X.S.: BATCH: A scalable asymmetric discrete cross-modal hashing. IEEE Transactions on Knowledge and Data Engineering (2020)

[302] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., van den Hengel, A.: Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding **163** (2017) 21–40

[303] Teney, D., Anderson, P., He, X., van den Hengel, A.: Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 4223–4232

[304] Li, Q., Tao, Q., Joty, S., Cai, J., Luo, J.: VQA-E: Explaining, elaborating, and enhancing your answers for visual questions. In: European Conference on Computer Vision. (2018) 552–567

[305] Zhang, Y., Hare, J.S., Prügel-Bennett, A.: Learning to count objects in natural images for visual question answering. International Conference on Learning Representations (2018)

[306] Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., Parikh, D.: Yin and yang: Balancing and answering binary visual questions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2016) 5014–5022

[307] Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. IEEE Computational intelligence magazine **13** (2018) 55–75

[308] Liu, X., Wang, M., Zha, Z.J., Hong, R.: Cross-modality feature learning via convolutional

autoencoder. ACM Transactions on Multimedia Computing, Communications, and Applications **15** (2019) 7

[309] Xu, X., Song, J., Lu, H., Yang, Y., Shen, F., Huang, Z.: Modal-adversarial semantic learning network for extendable cross-modal retrieval. In: Proceedings of the ACM on International Conference on Multimedia Retrieval. (2018) 46–54

[310] Zhu, X., Li, L., Liu, J., Li, Z., Peng, H., Niu, X.: Image captioning with triple-attention and stack parallel LSTM. Neurocomputing **319** (2018) 55–65

[311] Jiang, W., Ma, L., Jiang, Y.G., Liu, W., Zhang, T.: Recurrent fusion network for image captioning. In: European Conference on Computer Vision. (2018) 499–515

[312] Chen, C., Mu, S., Xiao, W., Ye, Z., Wu, L., Ju, Q.: Improving image captioning with conditional generative adversarial nets. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 8142–8150

[313] Dash, A., Gamboa, J.C.B., Ahmed, S., Liwicki, M., Afzal, M.Z.: TAC-GAN-text conditioned auxiliary classifier generative adversarial network. arXiv:1703.06412 (2017)

[314] Gu, J., Joty, S., Cai, J., Zhao, H., Yang, X., Wang, G.: Unpaired image captioning via scene graph alignments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 10323–10332

[315] Xu, W., Keshmiri, S., Wang, G.R.: Adversarially approximated autoencoder for image generation and manipulation. IEEE Transactions on Multimedia (2019)

[316] Feng, Y., Ma, L., Liu, W., Luo, J.: Unsupervised image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 4125–4134

[317] Yu, J., Yang, C., Qin, Z., Yang, Z., Hu, Y., Liu, Y.: Textual relationship modeling for cross-modal information retrieval. arXiv (2018)

[318] Chen, F., Ji, R., Su, J., Wu, Y., Wu, Y.: Structcap: Structured semantic embedding for image captioning. In: Proceedings of the ACM International Conference on Multimedia. (2017) 46–54

[319] Teney, D., Liu, L., van den Hengel, A.: Graph-structured representations for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 1–9

[320] Zhang, D., Cao, R., Wu, S.: Information fusion in visual question answering: A survey. Information Fusion (2019)

[321] Su, Z., Zhu, C., Dong, Y., Cai, D., Chen, Y., Li, J.: Learning visual knowledge memory networks for visual question answering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 7736–7745

[322] Hudson, D.A., Manning, C.D.: Compositional attention networks for machine reasoning. arXiv:1803.03067 (2018)

[323] Fan, H., Zhou, J.: Stacked latent attention for multimodal reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 1072–1080

[324] Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. In: International Conference on Machine Learning. (2016) 2397–2406

[325] Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: European Conference on Computer Vision. (2016) 451–466

[326] Ma, C., Shen, C., Dick, A.R., Wu, Q., Wang, P., van den Hengel, A., Reid, I.D.: Visual question answering with memory-augmented networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 6975–6984

[327] Wu, C., Liu, J., Wang, X., Dong, X.: Object-difference attention: A simple relational attention for visual question answering. In: Proceedings of the ACM International Conference on Multimedia. (2018) 519–527

[328] Singh, J., Ying, V., Nutkiewicz, A.: Attention on attention: Architectures for visual question answering (VQA). arXiv:1803.07724 (2018)

[329] Nguyen, D.K., Okatani, T.: Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 6087–6096

[330] Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 4971–4980

[331] Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 1821–1830

[332] Wang, W., Liu, P., Yang, S., Zhang, W.: Dynamic interaction networks for image-text multimodal learning. Neurocomputing **379** (2020) 262–272

[333] Peng, G., Li, H., You, H., Jiang, Z., Lu, P., Hoi, S., Wang, X.: Dynamic fusion with intra-and inter-modality attention flow for visual question answering. arXiv:1812.05252 (2018)

[334] Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41** (2019) 423–443

[335] Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. (2016) 457–468

[336] Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. arXiv:1610.04325 (2016)

[337] Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: MUTAN: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 2612–2620

[338] Gao, P., Li, H., Li, S., Lu, P., Li, Y., Hoi, S.C., Wang, X.: Question-guided hybrid convolution for visual question answering. In: European Conference on Computer Vision. (2018) 469–485

[339] Liu, X., Li, H., Shao, J., Chen, D., Wang, X.: Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In: European Conference on Computer Vision. (2018) 338–354

[340] Wu, Q., Wang, P., Shen, C., Reid, I., van den Hengel, A.: Are you talking to me? reasoned visual dialog generation through adversarial learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 6106–6115

[341] Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning **8** (1992) 229–256

[342] Gao, J., Wang, S., Wang, S., Ma, S., Gao, W.: Self-critical n-step training for image captioning. arXiv:1904.06861 (2019)

[343] Li, Y., Duan, N., Zhou, B., Chu, X., Ouyang, W., Wang, X., Zhou, M.: Visual question generation as dual task of visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 6116–6124

[344] Li, C., Deng, C., Wang, L., Xie, D., Liu, X.: Coupled CycleGAN: Unsupervised hashing network for cross-modal retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2019) 176–183

[345] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2015) 2641–2649

[346] You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2016) 4651–4659

[347] Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via

a visual sentinel for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 375–383

[348] Yan, S., Wu, F., Smith, J.S., Lu, W., Zhang, B.: Image captioning using adversarial networks and reinforcement learning. In: International Conference on Pattern Recognition. (2018) 248–253

[349] Liu, F., Xiang, T., Hospedales, T.M., Yang, W., Sun, C.: Inverse visual question answering: A new benchmark and vqa diagnosis tool. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)

[350] Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J.: Object-driven text-to-image synthesis via adversarial training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 12174–12182

[351] Lao, Q., Havaei, M., Pesaranghader, A., Dutil, F., Jorio, L.D., Fevens, T.: Dual adversarial inference for text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 7567–7576

[352] Joseph, K., Pal, A., Rajanala, S., Balasubramanian, V.N.: C4Synth: Cross-caption cycle-consistent text-to-image synthesis. In: IEEE Winter Conference on Applications of Computer Vision. (2019) 358–366

[353] Qiao, T., Zhang, J., Xu, D., Tao, D.: MirrorGAN: Learning text-to-image generation by re-description. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 1505–1514

[354] Li, B., Qi, X., Lukasiewicz, T., Torr, P.H.: Controllable text-to-image generation. arXiv:1909.07083 (2019)

[355] Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J.: Semantics disentangling for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 2327–2336

[356] Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D.: VQA: Visual question answering. International Journal of Computer Vision **123** (2017) 4–31

[357] Schwartz, I., Schwing, A., Hazan, T.: High-order attention models for visual question answering. In: Proceedings of the International Conference on Neural Information Processing Systems. (2017) 3664–3674

[358] Yu, D., Fu, J., Mei, T., Rui, Y.: Multi-level attention networks for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 4709–4717

[359] Zhu, C., Zhao, Y., Huang, S., Tu, K., Ma, Y.: Structured attentions for visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 1291–1300

[360] Lu, P., Li, H., Zhang, W., Wang, J., Wang, X.: Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2018)

[361] Song, J., Zeng, P., Gao, L., Shen, H.T.: From pixels to objects: Cubic visual attention for visual question answering. In: International Joint Conference on Artificial Intelligence. (2018) 906–912

[362] Osman, A., Samek, W.: Dual recurrent attention units for visual question answering. arXiv:1802.00209 (2018)

[363] Liu, Y., Zhang, X., Huang, F., Li, Z.: Adversarial learning of answer-related representation for visual question answering. In: Proceedings of the ACM International Conference on Information and Knowledge Management. (2018) 1013–1022

[364] Wu, C., Liu, J., Wang, X., Li, R.: Differential networks for visual question answering. Proceedings of the AAAI Conference on Artificial Intelligence (2019)

[365] Liu, F., Liu, J., Fang, Z., Lu, H.: Language and visual relations encoding for visual question answering. In: IEEE International Conference on Image Processing. (2019) 3307–3311

[366] Liu, F., Liu, J., Fang, Z., Hong, R., Lu, H.: Densely connected attention flow for visual question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2019) 869–875

[367] Aytar, Y., Vondrick, C., Torralba, A.: See, hear, and read: Deep aligned representations. arXiv:1706.00932 (2017)

[368] He, X., Peng, Y., Xie, L.: A new benchmark and approach for fine-grained cross-media retrieval. In: Proceedings of the ACM International Conference on Multimedia. (2019) 1740–1748

[369] Fu, C., Pei, W., Cao, Q., Zhang, C., Zhao, Y., Shen, X., Tai, Y.W.: Non-local recurrent neural memory for supervised sequence modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 6311–6320

# List of Abbreviations

| Abbreviation | Full Name / Short Definition |
|---|---|
| **DCNNs** | Deep Convolutional Neural Networks / A regularized version of multilayer perceptrons based on convolution kernels |
| **CBIR** | Content-based Image Retrieval / An image search task according to the content contained in images |
| **MAC** | Maximum Activations of Convolutions / Maximum value over a convolutional feature map |
| **R-MAC** | Regional Maximum Activations of Convolutions / Maximum value over a region on a convolutional feature map |
| **CroW** | Cross-dimensional Weighting / Weighting the activations over different feature maps |
| **SPoC** | Sum-Pooled Convolutional / Sum pooling over different feature maps |
| **ReLU** | Rectified Linear Unit / An activation function returns 0 if it receives any negative input |
| **SPM** | Spatial Pyramid Modeling / An method to model feature in a pyramid way |
| **t-SNE** | t-Distributed Stochastic Neighbor Embedding / A method to visualize high-dimensional data |
| **RPNs** | Region Proposal Networks / A network to obtain proposal for a region or an object |
| **FC** | Fully-Connected (layer) |
| **KNN** | K-Nearest Neighbors |
| **BoW** | Bag-of-Words / Method to embed features according to the number of feature occurrences |
| **VLAD** | Vector of Locally Aggregated Descriptors / Method to embed features based on their residuals w.r.t. each visual word |
| **FV** | Fisher Vector / Method to embed features by using Gaussian mixture model |
| **GeM** | Generalized Mean / A pooling method to apply over each feature map |
| **CAM** | Class Activation Maps / A feature weighting method based on an activated class output |
| **PCA** | Principal Component Analysis |
| **MMD** | Maximum Mean Discrepancy |
| **FGIR** | Fine-Grained Image Retrieval |
| **RKHS** | Reproducing Kernel Hilbert Space |
| **DKD** | Dual Knowledge Distillation // Knowledge distillation based on two teacher models |
| **GCNs** | Graph Convolutional Networks |
| **VQA** | Visual Question Answering / A computer vision task |
| **KL-divergence** | Kullback–Leibler divergence / A metric to measure the distance between two distributions |