# Exploring images with deep learning for classification, retrieval and synthesis

Liu, Y.

# Bibliography


[1] Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks **61** (2015) 85–117

[2] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016) `http://www.deeplearningbook.org`.

[3] Cun, L., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: NIPS. (1990)

[4] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1106–1114

[5] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115** (2015) 211–252

[6] Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: CVPR workshop. (2014)

[7] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015)

[8] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. (2015) 1–9

[9] Srivastava, R.K., Greff, K., Schmidhuber, J.: Training very deep networks. In: NIPS. (2015) 2377–2385

[10] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016) 770–778

[11] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. (2017) 4278–4284

[12] Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC. (2016)

[13] Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.: Deep networks with stochastic depth. In: ECCV. (2016) 646–661

[14] Veit, A., Wilber, M.J., Belongie, S.: Residual networks behave like ensembles of relatively shallow networks. In: NIPS. (2016) 550–558

[15] Agrawal, P., Girshick, R., Malik, J.: Analyzing the performance of multilayer neural networks for object recognition. In: ECCV. (2014) 329–344

[16] Liu, L., Shen, C., van den Hengel, A.: The treasure beneath convolutional layers: cross convolutional layer pooling for image classification. In: CVPR. (2015) 4749–4757

[17] Sermanet, P., Chintala, S., LeCun, Y.: Convolutional neural networks applied to house numbers digit classification. In: ICPR. (2012)

[18] Yang, S., Ramanan, D.: Multi-scale recognition with DAG-CNNs. In: ICCV. (2015) 1215–1223

[19] Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: ICCV. (2003) 1470–1477

[20] Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR. (2010) 3304–3311

[21] Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV. (2010) 143–156

[22] Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: ECCV. (2014) 392–407

[23] Yue-Hei Ng, J., Yang, F., Davis, L.S.: Exploiting local features from deep networks for image retrieval. In: CVPR, Deep Vision workshop. (2015)

[24] Wei, X.S., Gao, B.B., Wu, J.: Deep spatial pyramid ensemble for cultural event recognition. In: ICCV Workshops. (2015)

[25] Yoo, D., Park, S., Lee, J.Y., Kweon, I.S.: Multi-scale pyramid pooling for deep convolutional representation. In: CVPR, DeepVision workshop. (2015)

[26] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015) 3431–3440

[27] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: ICLR. (2015)

[28] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. (2015) 234–241

[29] Bertasius, G., Shi, J., Torresani, L.: Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In: CVPR. (2015) 4380–4389

[30] Shen, W., Wang, X., Wang, Y., Bai, X., Zhang, Z.: Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In: CVPR. (2015) 3982–3991

[31] Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV. (2015) 1395–1403

[32] Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NIPS. (2014) 2366–2374

[33] Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV. (2015) 2650–2658

[34] Wang, X., Fouhey, D.F., Gupta, A.: Designing deep networks for surface normal estimation. In: CVPR. (2015) 539–547

[35] Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: CVPR. (2015) 5455–5463

[36] Wang, L., Lu, H., Ruan, X., Yang, M.H.: Deep networks for saliency detection via local estimation and global search. In: CVPR. (2015) 3183–3192

[37] Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X.: Saliency detection with recurrent fully convolutional networks. In: ECCV. (2016) 825–841

[38] Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. TOMCCAP **2** (2006) 1–19

[39] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR. (2007)

[40] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110

[41] Wengert, C., Douze, M., Jégou, H.: Bag-of-colors for improved image search. In: ACM Multimedia. (2011) 1437–1440

[42] 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: ACM Multimedia. (2014) 157–166

[43] Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: ECCV. (2016) 241–257

[44] Radenović, F., Tolias, G., Chum, O.: Cnn image retrieval learns from bow: Unsupervised

fine-tuning with hard examples. In: ECCV. (2016) 3–20

[45] Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.S.: Neural codes for image retrieval. In: ECCV. (2014) 584–599

[46] Zheng, L., Wang, S., He, F., Tian, Q.: Seeing the big picture: Deep embedding with contextual evidences. CoRR **abs/1406.0132** (2014)

[47] Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: ECCV. (2008) 304–317

[48] Zheng, L., Yang, Y., Tian, Q.: SIFT meets CNN: A decade survey of instance retrieval. TPAMI **40** (2018) 1224–1244

[49] Yan, F., Mikolajczyk, K.: Deep correlation for matching images and text. In: CVPR. (2015) 3441–3450

[50] Ranjan, V., Rasiwasia, N., Jawahar, C.V.: Multi-label cross-modal retrieval. In: ICCV. (2015) 4094–4102

[51] Klein, B., Lev, G., Sadeh, G., Wolf, L.: Associating neural word embeddings with deep image representations using fisher vectors. In: CVPR. (2015) 4437–4446

[52] Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: ICCV. (2015) 2623–2631

[53] Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: CVPR. (2016) 5005–5013

[54] Wei, Y., Zhao, Y., Lu, C., Wei, S., Liu, L., Zhu, Z., Yan, S.: Cross-modal retrieval with cnn visual features: A new baseline. IEEE Transactions on Cybernetics **47** (2017) 449–460

[55] Karpathy, A., Li, F.F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR. (2015) 3128–3137

[56] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR. (2015) 3156–3164

[57] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: VQA: Visual question answering. In: ICCV. (2015) 2425–2433

[58] Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: ICCV. (2015) 1–9

[59] Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: CVPR. (2015) 2927–2936

[60] Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. In: CVPR. (2016) 69–77

[61] Hotelling, H.: Relations between two sets of variates. Biometrika **28** (1936) 321–377

[62] Andrew, G., Arora, R., Livescu, K., Bilmes, J.: Deep canonical correlation analysis. In: ICML. (2013) 1247–1255

[63] Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. IJCV **106** (2014) 210–233

[64] Liu, Y., Guo, Y., Bakker, E.M., Lew, M.S.: Learning a recurrent residual fusion network for multimodal matching. In: ICCV. (2017) 4107–4116

[65] Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Hierarchical multimodal lstm for dense visual-semantic embedding. In: ICCV. (2017) 1881–1889

[66] Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. In: NIPS workshop. (2014)

[67] Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: CVPR. (2017) 299–307

[68] Huang, Y., Wang, W., Wang, L.: Instance-aware image and sentence matching with selective multimodal lstm. In: CVPR. (2017) 2310–2318

[69] Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Shen, Y.: Dual-path convolutional image-text

embedding. CoRR **abs/1711.05535** (2017)

[70] Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: ACM Multimedia. (2017) 154–162

[71] Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: ACM Multimedia. (2014) 7–16

[72] Habibian, A., Mensink, T., Snoek, C.G.: Videostory: A new multimedia embedding for few-example recognition and translation of events. In: ACM Multimedia. (2014) 17–26

[73] Rastegar, S., Soleymani, M., Rabiee, H.R., Mohsen Shojaee, S.: Mdl-cw: A multimodal deep learning framework with cross weights. In: CVPR. (2016) 2601–2609

[74] Vukotić, V., Raymond, C., Gravier, G.: Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications. In: ICMR. (2016) 343–346

[75] Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: CVPR. (2017) 3174–3183

[76] Eisenschtat, A., Wolf, L.: Linking image and text with 2-way nets. In: CVPR. (2017) 4601–4611

[77] Gu, J., Cai, J., Joty, S., Niu, L., Wang, G.: Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In: CVPR. (2018)

[78] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR. (2016) 2387–2395

[79] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. (2014) 2672–2680

[80] Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. (2016) 694–711

[81] Shuhui Jiang, Y.F.: Fashion style generator. In: IJCAI. (2017) 3721–3727

[82] Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: ECCV. (2016) 702–716

[83] Ulyanov, D., Vedaldi, A., Lempitsky, V.: Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In: CVPR. (2017) 4105–4113

[84] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML. (2016) 1060–1069

[85] Zhu, S., Fidler, S., Urtasun, R., Lin, D., Chen, C.L.: Be your own prada: Fashion synthesis with structural coherence. In: ICCV. (2017) 1689–1697

[86] Mirza, M., Osindero, S.: Conditional generative adversarial nets. CoRR **abs/1411.1784** (2014)

[87] Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. In: ECCV. (2016) 776–791

[88] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. (2017) 5967–5976

[89] Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: NIPS. (2016) 469–477

[90] Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NIPS. (2017) 700–708

[91] Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. In: ICLR. (2017)

[92] Benaim, S., Wolf, L.: One-sided unsupervised domain mapping. In: NIPS. (2017) 752–762

[93] Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: CVPR. (2017) 95–104

[94] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. (2017) 2223–2232

[95] Kiapour, M.H., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L.: Where to buy it: Matching street clothing photos in online shops. In: ICCV. (2015) 3343–3351

[96] Al-Halah, Z., Stiefelhagen, R., Grauman, K.: Fashion forward: Forecasting visual style in fashion. In: ICCV. (2017) 388–397

[97] Liu, S., Liang, X., Liu, L., Lu, K., Lin, L., Cao, X., Yan, S.: Fashion parsing with video context. IEEE Transactions on Multimedia **17** (2015) 1347–1358

[98] Guan, P., Reiss, L., Hirshberg, D.A., Weiss, E., Black, M.J.: Drape: Dressing any person. ACM Trans. Graph. **31** (2012) 35:1–35:10

[99] Zhou, Z., Shu, B., Zhuo, S., Deng, X., Tan, P., Lin, S.: Image-based clothes animation for virtual fitting. In: SIGGRAPH Asia. (2012)

[100] Movania, M.M., Farbiz, F.: Depth image based cloth deformation for virtual try-on. In: ACM SIGGRAPH. (2013)

[101] Yang, S., Ambert, T., Pan, Z., Wang, K., Yu, L., Berg, T.L., Lin, M.C.: Detailed garment recovery from a single-view image. CoRR **abs/1608.01250** (2016)

[102] Hauswiesner, S., Straka, M., Reitmayr, G.: Virtual try-on through image-based rendering. IEEE Transactions on Visualization and Computer Graphics **19** (2013) 1552–1565

[103] Pons-Moll, G., Pujades, S., Hu, S., Black, M.J.: Clothcap: Seamless 4d clothing capture and retargeting. ACM Trans. Graph. **36** (2017) 73:1–73:15

[104] Gultepe, U., Gudukbay, U.: Real-time virtual fitting with body measurement and motion smoothing. Computers & Graphics **43** (2014) 31–43

[105] Jetchev, N., Bergmann, U.: The conditional analogy gan: Swapping fashion articles on people images. In: ICCV Workshop. (2017)

[106] Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: VITON: an image-based virtual try-on network. In: CVPR. (2018)

[107] Liu, Y., Guo, Y., S. Lew, M.: On the exploration of convolutional fusion networks for visual recognition. In: MMM. (2017) 277–289

[108] Liu, Y., Guo, Y., Georgiou, T., Lew, M.S.: Fusion that matters: convolutional fusion networks for visual recognition. Multimedia Tools and Applications (2018)

[109] Canny, J.: A computational approach to edge detection. TPAMI **8** (1986) 679–698

[110] Xiaofeng, R., Bo, L.: Discriminatively trained sparse code gradients for contour detection. In: NIPS. (2012) 593–601

[111] Leordeanu, M., Sukthankar, R., Sminchisescu, C.: Generalized boundaries from multiple image interpretations. TPAMI **36** (2014) 1312–1324

[112] Sironi, A., Lepetit, V., Fua, P.: Projection onto the manifold of elongated structures for accurate extraction. In: ICCV. (2015) 316–324

[113] Kivinen, J.J., Williams, C.K.I., Heess, N.: Visual boundary prediction: A deep neural prediction network and quality dissection. In: AISTATS. (2014)

[114] Liu, Y., Lew, M.S.: Learning relaxed deep supervision for better edge detection. In: CVPR. (2016) 231–240

[115] Liu, Y., Guo, Y., Wu, S., Lew, M.S.: Deepindex for accurate and efficient image retrieval. In: ICMR. (2015) 43–50

[116] Liu, Y., Guo, Y., Liu, L., Bakker, E.M., Lew, M.S.: Cyclematch: A cycle-consistent embedding network for image-text matching. (2018)

[117] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., DollÃąr, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. (2014) 740–755

[118] Liu, Y., Liu, L., Guo, Y., Lew, M.S.: Learning visual and textual representations for multimodal matching and classification. Pattern Recognition **84** (2018) 51–67

[119] Liu, Y., Guo, Y., Chen, W., Lew, M.S.: An extensive study of cycle-consistent generative networks for image-to-image translation. In: ICPR. (2018)

[120] Liu, Y., Chen, W., Liu, L., Lew, M.S.: Swapgan: A multi-stage generative approach for person-to-person fashion style transfer. (2018)

[121] Babenko, A., Lempitsky, V.S.: Aggregating local deep features for image retrieval. In: ICCV. (2015) 1269–1277

[122] Lin, M., Chen, Q., Yan, S.: Network in network. In: ICLR. (2014)

[123] Gregor, K., LeCun, Y.: Emergence of complex-like cells in a temporal product network with local receptive fields. CoRR **abs/1006.0448** (2010)

[124] Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: CVPR. (2015) 2892–2900

[125] Lee, C., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: AISTATS. (2015)

[126] Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. TPAMI **33** (2011) 898–916

[127] Dollár, P., Zitnick, C.L.: Fast edge detection using structured forests. TPAMI **37** (2015) 1558–1570

[128] Krizhevsky, A.: Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto. (2009)

[129] Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. IJCV **111** (2015) 98–136

[130] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding. In: ACM Multimedia. (2014) 675–678

[131] Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A.C., Bengio, Y.: Maxout networks. In: ICML. (2013) 1319–1327

[132] Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR. (2015)

[133] Liang, M., Hu, X.: Recurrent convolutional neural network for object recognition. In: CVPR. (2015) 3367–3375

[134] Jin, X., Xu, C., Feng, J., Wei, Y., Xiong, J., Yan, S.: Deep learning with s-shaped rectified linear activation units. In: AAAI. (2016) 1737–1743

[135] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. (2015) 448–456

[136] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006) 2169–2178

[137] Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR. (2009) 413–420

[138] Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP. (2008) 722–729

[139] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology (2011)

[140] Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: ECCV. (2008) 304–317

[141] Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR. (2006) 2161–2168

[142] Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27

[143] Hariharan, B., Arbelaez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: ICCV. (2011) 991–998

[144] Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr,

P.H.S.: Conditional random fields as recurrent neural networks. In: ICCV. (2015) 1529–1537

[145] Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: NIPS. (2015) 2017–2025

[146] De Brabandere, B., Jia, X., Tuytelaars, T., Van Gool, L.: Dynamic filter networks. In: NIPS. (2016) 667–675

[147] Ferrari, V., Fevrier, L., Jurie, F., Schmid, C.: Groups of adjacent contour segments for object detection. TPAMI **30** (2008) 36–51

[148] Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. IJCV **59** (2004) 167–181

[149] Dollár, P., Tu, Z., Belongie, S.: Supervised learning of edges and object boundaries. In: CVPR. (2006) 1964–1971

[150] Isola, P., Zoran, D., Krishnan, D., Adelson, E.H.: Crisp boundary detection using pointwise mutual information. In: ECCV. (2014) 799–814

[151] Hallman, S., Fowlkes, C.C.: Oriented edge forests for boundary detection. In: CVPR. (2015) 1732–1740

[152] Lim, J., Zitnick, C.L., Dollár, P.: Sketch tokens: A learned mid-level representation for contour and object detection. In: CVPR. (2013) 3158–3165

[153] Hwang, J., Liu, T.: Pixel-wise deep learning for contour detection. In: ICLR. (2015)

[154] Bertasius, G., Shi, J., Torresani, L.: High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision. In: ICCV. (2015) 504–512

[155] Ganin, Y., Lempitsky, V.S.: N^4 -fields: Neural network nearest neighbor fields for image transforms. In: ACCV. (2014) 536–551

[156] Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: CVPR. (2014) 891–898

[157] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1097–1105

[158] Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: CVPR. (2015) 5188–5196

[159] Cun, L., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: NIPS. (1990)

[160] Yan, Z., Zhang, H., Piramuthu, R., Jagadeesh, V., DeCoste, D., Di, W., Yu, Y.: Hd-cnn: Hierarchical deep convolutional neural network for large scale visual recognition. In: ICCV. (2015) 2740–2748

[161] Sironi, A., TÂÍÂźretken, E., Lepetit, V., Fua, P.: Multiscale centerline detection. TPAMI (2015)

[162] Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. (2012) 746–760

[163] Dollár, P., Zitnick, C.L.: Structured forests for fast edge detection. In: ICCV. (2013) 1841–1848

[164] Zheng, L., Wang, S., Liu, Z., Tian, Q.: Packing and padding: Coupled multi-index for accurate image retrieval. In: CVPR. (2014) 1947–1954

[165] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. (2014)

[166] Sun, S., Zhou, W., Li, H., Tian, Q.: Search by detection: Object-level feature for image retrieval. In: ICIMCS. (2014) 46–49

[167] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006) 2169–2178

[168] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: ICML. (2014) 647–655

[169] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015)

[170] Jégou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. IJCV **87** (2010) 316–336

[171] Babenko, A., Lempitsky, V.S.: The inverted multi-index. In: CVPR. (2012) 3069–3076

[172] Agrawal, P., Girshick, R., Malik, J.: Analyzing the performance of multilayer neural networks for object recognition. In: ECCV. (2014) 329–344

[173] Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: CVPR. (2012) 2911–2918

[174] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: CVPR. (2008)

[175] Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR. (2006) 2161–2168

[176] Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: VISAPP. (2009) 331–340

[177] Yang, M., Wang, X., Lin, Y., Tian, Q.: Semantic-aware co-indexing for near-duplicate image retrieval. In: ICCV. (2014)

[178] Tolias, G., Avrithis, Y., Jégou, H.: To aggregate or not to aggregate: selective match kernels for image search. In: ICCV. (2013) 1401–1408

[179] Jégou, H., Douze, M., Schmid, C.: On the burstiness of visual elements. In: CVPR. (2009)

[180] Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., Lazebnik, S.: Improving image-sentence embeddings using large weakly annotated photo collections. In: ECCV. (2014) 529–545

[181] Karpathy, A., Joulin, A., Li, F.: Deep fragment embeddings for bidirectional image sentence mapping. In: NIPS. (2014) 1889–1897

[182] Mineiro, P., Karampatziakis, N.: A randomized algorithm for cca. In: NIPS workshop. (2014)

[183] Michaeli, T., Wang, W., Livescu, K.: Nonparametric canonical correlation analysis. In: ICML. (2016) 1967–1976

[184] Hardoon, D.R., Szedmak, S.R., Shawe-taylor, J.R.: Canonical correlation analysis: An overview with application to learning methods. Neural Computation **16** (2004) 2639–2664

[185] Lev, G., Sadeh, G., Klein, B., Wolf, L.: RNN fisher vectors for action recognition and image annotation. In: ECCV. (2016) 833–850

[186] Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS. (2014) 3104–3112

[187] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9** (1997) 1735–1780

[188] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS. (2013) 3111–3119

[189] Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL **2** (2014) 67–78

[190] Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). ICLR (2015)

[191] Lin, X., Parikh, D.: Leveraging visual question answering for image-caption ranking. In: ECCV. (2016) 261–277

[192] Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., Torralba, A.: Learning

cross-modal embeddings for cooking recipes and food images. In: CVPR. (2017) 3020–3028

[193] Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: CVPR. (2017) 1970–1979

[194] Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: ECCV. (2016) 451–466

[195] Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: CVPR. (2016) 49–58

[196] Bucher, M., Herbin, S., Jurie, F.: Improving semantic embedding consistency by metric learning for zero-shot classiffication. In: ECCV. (2016) 730–746

[197] Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: ECCV. (2016) 817–834

[198] Zhang, Y., Yuan, L., Guo, Y., He, Z., Huang, I.A., Lee, H.: Discriminative bimodal networks for visual localization and detection with natural language queries. In: CVPR. (2017) 557–566

[199] He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., Ma, W.Y.: Dual learning for machine translation. In: NIPS. (2016) 820–828

[200] Huang, Y., Wu, Q., Wang, L.: Learning semantic concepts and order for image and sentence matching. In: CVPR. (2018)

[201] Wang, F., Huang, Q., Guibas, L.: Image co-segmentation via consistent functional maps. In: ICCV. (2013) 849–856

[202] Zhou, T., Krähenbühl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3d-guided cycle consistency. In: CVPR. (2016) 117–126

[203] Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR. (2017) 270–279

[204] Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: ICCV. (2017) 2849–2857

[205] Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: ICML. (2017) 1857–1865

[206] Chen, X., Zitnick, C.L.: Mind's eye: A recurrent visual representation for image caption generation. In: CVPR. (2015) 2422–2431

[207] van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne. JMLR **9** (2008) 2579–2605

[208] Nandakumar, K., Chen, Y., Dass, S.C., Jain, A.: Likelihood ratio-based biometric score fusion. IEEE TPAMI **30** (2008) 342–347

[209] Zheng, L., Wang, S., Tian, L., He, F., Liu, Z., Tian, Q.: Query-adaptive late fusion for image search and person re-identification. In: CVPR. (2015) 1741–1750

[210] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: EMNLP. (2017) 670–680

[211] Faghri, F., Fleet, D.J., Kiros, R., Fidler, S.: VSE++: improved visual-semantic embeddings. CoRR **abs/1707.05612** (2017)

[212] Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: NIPS. (2015) 3294–3302

[213] Vendrov, I., Kiros, R., Fidler, S., Urtasun, R.: Order-embeddings of images and language. In: ICLR. (2016)

[214] Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M.A., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: NIPS. (2013) 2121–2129

[215] Tenenbaum, J.B., Freeman, W.T.: Separating style and content with bilinear models. Neural Computation **12** (2000) 1247–1283

[216] Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: CVPR. (2016) 317–326

[217] 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: EMNLP. (2016) 457–468

[218] Pham, N., Pagh, R.: Fast and scalable polynomial kernels via explicit feature maps. In: SIGKDD. (2013) 239–247

[219] Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using amazon's mechanical turk. In: Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. (2010) 139–147

[220] Socher, R., Karpathy, A., Le, Q., Manning, C., Ng, A.: Grounded compositional semantics for finding and describing images with sentences. TACL **2** (2014) 207–218

[221] Simon, M., Rodner, E.: Neural activation constellations: Unsupervised part model discovery with convolutional networks. In: ICCV. (2015) 1143–1151

[222] Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: ICCV. (2015) 1449–1457

[223] Zhang, X., Xiong, H., Zhou, W., Lin, W., Tian, Q.: Picking deep filter responses for fine-grained image recognition. In: CVPR. (2016) 1134–1142

[224] Azizpour, H., Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: Factors of transferability for a generic convnet representation. TPAMI **38** (2016) 1790–1802

[225] Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based rcnn for fine-grained detection. In: ECCV. (2014) 834–849

[226] Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: CVPR. (2015) 842–850

[227] Lin, D., Shen, X., Lu, C., Jia, J.: Deep lac: Deep localization, alignment and classification for fine-grained recognition. In: CVPR. (2015) 1666–1674

[228] Qian, Q., Jin, R., Zhu, S., Lin, Y.: Fine-grained visual categorization via multi-stage metric learning. In: CVPR. (2015) 3716–3724

[229] Guo, Y., Liu, Y., Lao, S., Bakker, E.M., Bai, L., Lew, M.S.: Bag of surrogate parts feature for visual recognition. IEEE Trans. on Multimedia (2017)

[230] Xie, L., Wang, J., Lin, W., Zhang, B., Tian, Q.: Towards reversal-invariant image representation. IJCV **123** (2017) 226–250

[231] Cai, S., Zhang, L., Zuo, W., Feng, X.: A probabilistic collaborative representation based approach for pattern classification. In: CVPR. (2016) 2950–2959

[232] Wang, D., Shen, Z., Shao, J., Zhang, W., Xue, X., Zhang, Z.: Multiple granularity descriptors for fine-grained categorization. In: ICCV. (2015) 2399–2406

[233] Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV. (2017) 2813–2821

[234] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR. (2015)

[235] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015)

[236] Tyleček, R., Šára, R.: Spatial pattern templates for recognition of objects with regular structure. In: German Conference on Pattern Recognition. (2013) 364–374

[237] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U.,

Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. (2016) 3213–3223

[238] Yu, Q., Liu, F., SonG, Y.Z., Xiang, T., Hospedales, T., Loy, C.C.: Sketch me that shoe. In: CVPR. (2016) 799–807

[239] Liang, X., Lin, L., Yang, W., Luo, P., Huang, J., Yan, S.: Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval. IEEE Transactions on Multimedia **18** (2016) 1175–1186

[240] Zhang, X., Jia, J., Gao, K., Zhang, Y., Zhang, D., Li, J., Tian, Q.: Trip outfits advisor: Location-oriented clothing recommendation. IEEE Transactions on Multimedia **19** (2017) 2533–2544

[241] Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR. (2016) 1096–1104

[242] Zhang, L., Liu, M., Chen, L., Hu, Y., Zhang, L., Zimmermann, R.: Online modeling of aesthetic communities using deep perception graph analytics. IEEE Transactions on Multimedia (2017)

[243] Garg, V., Banerjee, R.H., Rajagopal, A.K., Thiruvambalam, S., Warrier, D.: Sales potential: Modeling sellability of fashion product. KDD (2017)

[244] Zheng, Z.H., Zhang, H.T., Zhang, F.L., Mu, T.J.: Image-based clothes changing system. Computational Visual Media **3** (2017) 337–347

[245] Yoo, D., Kim, N., Park, S., Paek, A.S., Kweon, I.S.: Pixel-level domain transfer. In: ECCV. (2016) 517–532

[246] Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: NIPS. (2017) 405–415

[247] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. (2017) 1302–1310

[248] Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: CVPR. (2017) 6757–6765

[249] Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill (2016)

[250] Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NIPS. (2016) 2226–2234

[251] Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Statist. **22** (1951) 79–86

[252] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Trans. on Image Processing **13** (2004) 600–612

[253] Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: CVPR. (2017) 3319–3327

[254] Zhang, Q., Wu, Y.N., Zhu, S.C.: Interpretable convolutional neural networks. In: CVPR. (2018)

[255] Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. TPAMI **38** (2016) 1425–1438

[256] Lei Ba, J., Swersky, K., Fidler, S., salakhutdinov, R.: Predicting deep zero-shot convolutional neural networks using textual descriptions. In: ICCV. (2015) 4247–4255

[257] Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV. (2017) 3774–3782

[258] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR. (2018)

# List of Abbreviations

| Abb. | Full Name |
|---|---|
| **AP** | Average Precision |
| **BN** | Batch Normalization |
| **BoW** | Bag of Words |
| **CBIR** | Content-based Image Retrieval |
| **CBP** | Compact Bilinear Pooling |
| **CCA** | Canonical Correlation Analysis |
| **CFN** | Convolutional Fusion Networks |
| **cGAN** | Conditional Generative Adversarial Networks |
| **CNN** | Convolutional Neural Networks |
| **DSN** | Deeply Supervised Networks |
| **FC** | Fully-connected Layer |
| **FCFN** | Fully Convolutional Fusion Networks |
| **FCN** | Fully Convolutional Networks |
| **FFT** | Fast Fourier Transformation |
| **FV** | Fisher Vector |
| **GAN** | Generative Adversarial Networks |
| **GAP** | Global Average Pooling |
| **HGLMM** | Hybrid Gaussian-Laplacian Mixture Model |
| **I2I** | Image-to-Image Translation |
| **I2T** | Image-to-Text Translation |
| **IoU** | Intersection over Union |
| **IS** | Inception Score |

| Abb. | Full Name |
|---|---|
| **LC** | Locally-connected Layer |
| **LSGAN** | Least Square Generative Adversarial Networks |
| **LSTM** | Long Short-Term Memory |
| **MA** | Multiple Assignment |
| **mAP** | Mean Average Precision |
| **MC-Net** | Multi-modal Classification Network |
| **MDS** | Multi-Dimensional Scaling |
| **MM-Net** | Multi-modal Matching Network |
| **MMC-Net** | Multi-modal Matching and Classification Network |
| **NMS** | Non-maximal Suppression |
| **ODS** | Fixed Contour Threshold |
| **OIS** | Per-image Best Threshold |
| **PCA** | Principal Component Analysis |
| **RDS** | Relaxed Deep Supervision |
| **ReLU** | Rectified Linear Units |
| **RNN** | Recurrent Neural Networks |
| **RRF** | Recurrent Residual Fusion |
| **SGD** | Stochastic Gradient Descent |
| **SIFT** | Scale Invariant Feature Transform |
| **SVM** | Support Vector Machine |
| **T2I** | Text-to-Image Translation |
| **TF** | Term Frequency |
| **TPS** | Thin Plate Spline |
| **t-SNE** | t-Distributed Stochastic Neighbor Embedding |
| **VLAD** | Vector of Aggregate Locally Descriptor |