# References


[1] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," in *ACM transactions on graphics (TOG)*, vol. 25, pp. 835–846, ACM, 2006. viii, 9

[2] "Structure-from-Motion example." `https://openmvg.readthedocs.io/en/latest/_images/structureFromMotion.png`. Accessed: 2019-11-12. viii, 11

[3] "Sparse/dense point cloud example." `http://www.visual-experiments.com/category/ogre3d/page/2/`. Accessed: 2019-11-13. viii, 13

[4] S. Pillai and J. Leonard, "Monocular slam supported object recognition," *arXiv preprint arXiv:1506.01732*, 2015. ix, 14, 15

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017. ix, x, 16, 17, 21, 44

[6] M. Crocco, C. Rubino, and A. Del Bue, "Structure from motion with objects," in *Computer Vision and Pattern Recognition*, pp. 782–788, IEEE, 2016. ix, 18, 19, 22, 28, 30, 31, 72, 88, 97

[7] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc.*

*IEEE Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017. ix, xiv, 8, 20, 49, 72, 75, 87, 88

[8] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Computer Vision and Pattern Recognition*, pp. 1386–1393, 2014. ix, 32, 39, 47

[9] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, 2017. x, 34, 35

[10] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that?," *arXiv preprint arXiv:1611.07450*, 2016. x, 43

[11] "Google ARCcore." `https://developers.google.com/ar`. Accessed: 2019-11-13. xii, 83

[12] "Microsoft Hololens." `https://www.microsoft.com/en-us/hololens`. Accessed: 2019-11-13. xii, 83

[13] "Apple ARKit." `https://developer.apple.com/augmented-reality/`. Accessed: 2019-11-13. xii, 83

[14] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. xiv, 16, 62, 85, 91

[15] C. Rubino, M. Crocco, and A. Del Bue, "3d object localisation from multi-view image detections," *IEEE transactions on pattern analysis and machine*

*intelligence*, vol. 40, no. 6, pp. 1281–1294, 2017. xiv, xv, 18, 19, 22, 62, 64, 88, 91, 92, 93

[16] J. Chen, D. Bautembach, and S. Izadi, "Scalable real-time volumetric surface reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 4, p. 113, 2013. 8

[17] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An evaluation of the rgb-d slam system.," in *Icra*, vol. 3, pp. 1691–1696, 2012. 8

[18] Q.-Y. Zhou and V. Koltun, "Dense scene reconstruction with points of interest," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 4, p. 112, 2013. 8

[19] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3d reconstruction at scale using voxel hashing," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 6, p. 169, 2013. 8

[20] J. Kybic and M. Unser, "Fast parametric elastic image registration," *IEEE transactions on image processing*, vol. 12, no. 11, pp. 1427–1442, 2003. 8

[21] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016. 9

[22] C. Wu, "Towards linear-time incremental structure from motion," in *2013 International Conference on 3D Vision-3DV 2013*, pp. 127–134, IEEE, 2013. 9

[23] D. G. Lowe, D. G. Lowe, and D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on*

*Computer Vision-Volume 2 - Volume 2*, ICCV '99, (Washington, DC, USA), pp. 1150–, IEEE Computer Society, 1999. 10

[24] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, June 1981. 12

[25] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003. 12, 28, 63

[26] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—a modern synthesis," in *International workshop on vision algorithms*, pp. 298–372, Springer, 1999. 13

[27] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *2011 IEEE international conference on robotics and automation*, pp. 1817–1824, IEEE, 2011. 15

[28] D. Lin, S. Fidler, and R. Urtasun, "Holistic scene understanding for 3d object detection with rgbd cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1417–1424, 2013. 15, 16

[29] B.-s. Kim, S. Xu, and S. Savarese, "Accurate localization of 3d objects from rgb-d data using segmentation hypotheses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3182–3189, 2013. 15

[30] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao, "Histogram of oriented normal vectors for object recognition with a depth sensor," in *Asian conference on computer vision*, pp. 525–538, Springer, 2012. 15

[31] S. Song and J. Xiao, "Sliding shapes for 3d object detection in depth images," in *European conference on computer vision*, pp. 634–651, Springer, 2014. 15

[32] Z. Ren and E. B. Sudderth, "Three-dimensional object detection and layout prediction using clouds of oriented gradients," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1525–1533, 2016. 16

[33] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European conference on computer vision*, pp. 345–360, Springer, 2014. 16

[34] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 564–571, 2013. 16

[35] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena, "Semantic labeling of 3d point clouds for indoor scenes," in *Advances in neural information processing systems*, pp. 244–252, 2011. 16

[36] X. Ren, L. Bo, and D. Fox, "Rgb-(d) scene labeling: Features and algorithms," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2759–2766, IEEE, 2012. 16

[37] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*, pp. 746–760, Springer, 2012. 16

[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for

semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015. 16

[39] S. Song and J. Xiao, "Deep sliding shapes for amodal 3d object detection in rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 808–816, 2016. 16

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, 2015. 16, 28

[41] J. Lahoud and B. Ghanem, "2d-driven 3d object detection in rgb-d images," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4622–4630, 2017. 17

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. 21, 46, 50, 92

[43] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999. 23, 70

[44] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015. 23, 68

[45] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski, *Bundle Adjustment in the Large*, pp. 29–42. Springer Berlin Heidelberg, 2010. 28

[46] R. Toldo, R. Gherardi, M. Farenzena, and A. Fusiello, "Hierarchical

structure-and-motion recovery from uncalibrated images," *Computer Vision and Image Understanding*, vol. 140, pp. 127 – 143, 2015. 28

[47] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, *et al.*, "Building rome on a cloudless day," in *ECCV 2010*, pp. 368–381, Springer, 2010. 28

[48] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Computer Vision and Pattern Recognition*, 2016. 28

[49] G. Biegelbauer and M. Vincze, "Efficient 3d object detection by fitting superquadrics to range image data for robot's object manipulation," in *Robotics and Automation, 2007 IEEE International Conference on*, pp. 1086–1091, IEEE, 2007. 28

[50] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Close-range scene segmentation and reconstruction of 3d point cloud maps for mobile manipulation in domestic environments," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pp. 1–6, IEEE, 2009. 28

[51] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel, "Visual question answering: A survey of methods and datasets," *arXiv preprint arXiv:1607.05910*, 2016. 28

[52] P. Gay, J. Stuart, and A. Del Bue, "Visual graphs from motion (vgfm): Scene understanding with object geometry reasoning," *CoRR*, vol. abs/1807.05933, 2018. 28

[53] T. Shao, W. Xu, K. Zhou, J. Wang, D. Li, and B. Guo, "An interactive approach to semantic modeling of indoor scenes with an rgbd camera," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 6, p. 136, 2012. 28

[54] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013. 28

[55] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 567–576, 2015. 28

[56] B.-S. Kim, P. Kohli, and S. Savarese, "3d scene understanding by voxel-crf," in *International Conference on Computer Vision*, pp. 1425–1432, 2013. 28

[57] S. Y. Bao, M. Bagra, Y.-W. Chao, and S. Savarese, "Semantic structure from motion with points, regions, and objects," in *Computer Vision and Pattern Recognition*, pp. 2703–2710, IEEE, 2012. 28

[58] N. D. Reddy, P. Singhal, V. Chari, and K. M. Krishna, "Dynamic body vslam with semantic constraints," in *Intelligent Robots and Systems*, pp. 1897–1904, IEEE, 2015. 28

[59] M. Sung, V. G. Kim, R. Angst, and L. Guibas, "Data-driven structural priors for shape completion," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, p. 175, 2015. 28

[60] N. Fioraio and L. Di Stefano, "Joint detection, tracking and mapping by semantic bundle adjustment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1538–1545, 2013. 28

[61] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid, "Dense reconstruction using 3d object shape priors," in *Computer Vision and Pattern Recognition*, pp. 1288–1295, IEEE, 2013. 28

[62] A. Del Bue, "Adaptive non-rigid registration and structure from motion from image trajectories," *International Journal of Computer Vision*, vol. 103, pp. 226–239, June 2013. 28

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015. 28

[64] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik, "Fast, accurate detection of 100,000 object classes on a single machine," in *Computer Vision and Pattern Recognition*, pp. 1814–1821, IEEE, 2013. 28

[65] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Computer Vision and Pattern Recognition*, June 2016. 28

[66] D. A. Forsyth, S. Ioffe, and J. Haddon, "Bayesian structure from motion," in *ICCV*, vol. 1, pp. 660–665, IEEE, 1999. 29

[67] J. E. Solem, F. Kahl, and A. Heyden, "Visibility constrained surface evolution," in *Computer Vision and Pattern Recognition*, 2005. 29

[68] A. Del Bue, X. Llado, and L. Agapito, "Non-rigid metric shape and motion recovery from uncalibrated images using priors," in *Computer Vision and Pattern Recognition*, vol. 1, pp. 1191–1198, IEEE, 2006. 29

[69] S. Y. Bao and S. Savarese, "Semantic structure from motion," in *Computer Vision and Pattern Recognition*, pp. 2025–2032, IEEE, 2011. 29, 72

[70] L. Torresani, A. Hertzmann, and C. Bregler, "Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors," *Transactions on Pattern Analysis and Machine Intelligence*, 2008. 29

[71] S. I. Olsen and A. Bertoli, "Implicit non-rigid structure-from-motion with priors," *Journal of Mathematical Imaging and Vision*, 2008. 29

[72] A. Del Bue, "A factorization approach to structure from motion with shape priors," in *Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008. 29

[73] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Trajectory space: A dual representation for nonrigid structure from motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1442–1456, 2011. 30

[74] P. F. Gotardo and A. M. Martinez, "Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 2051–2065, 2011. 30

[75] J. Cho, M. Lee, and S. Oh, "Complex non-rigid 3d shape recovery using a procrustean normal distribution mixture model," *International Journal of Computer Vision*, vol. 117, no. 3, pp. 1–21, 2015. 30

[76] L. Quan and T. Kanade, "A factorization method for affine structure from line correspondences," in *Computer Vision and Pattern Recognition*, pp. 803–808, IEEE, 1996. 30

[77] G. Schindler, P. Krishnamurthy, and F. Dellaert, "Line-based structure from motion for urban environments," in *3D Data Processing, Visualization, and Transmission, Third International Symposium on*, pp. 846–853, IEEE, 2006. 30

[78] R. Berthilsson, K. Åström, and A. Heyden, "Reconstruction of general

curves, using factorization and bundle adjustment," *International Journal of Computer Vision*, vol. 41, no. 3, pp. 171–182, 2001. 30

[79] F. Mai and Y. Hung, "3d curves reconstruction from multiple images," in *Digital Image Computing: Techniques and Applications*, pp. 462–467, IEEE, 2010. 30

[80] F. Kahl and J. August, "Multiview reconstruction of space curves," in *International Conference on Computer Vision*, pp. 1017–1024, IEEE, 2003. 30

[81] I. Nurutdinova and A. Fitzgibbon, "Towards pointless structure from motion: 3d reconstruction and camera parameters from general 3d curves," in *International Conference on Computer Vision*, vol. 1, pp. 2363–2371, IEEE, 2015. 30

[82] G. Cross and A. Zisserman, "Quadric reconstruction from dual-space geometry," in *International Conference on Computer Vision*, pp. 25–31, IEEE, 1998. 30

[83] F. Kahl and A. Heyden, "Affine structure and motion from points, lines and conics," *International Journal of Computer Vision*, vol. 33, no. 3, pp. 163–180, 1999. 30

[84] L. Reyes and E. Bayro Corrochano, "The projective reconstruction of points, lines, quadrics, plane conics and degenerate quadrics using uncalibrated cameras," *Image and Vision Computing*, vol. 23, no. 8, pp. 693–706, 2005. 30

[85] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2559–2566, June 2010. 31

[86] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *Journal of Machine Learning Research*, vol. 11, no. Mar, pp. 1109–1135, 2010. 31, 33

[87] G. W. Taylor, I. Spiro, C. Bregler, and R. Fergus, "Learning invariance through imitation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2729–2736, IEEE, 2011. 31, 33

[88] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE International Conference on Computer vision*, vol. 2, pp. 1150–1157, IEEE, 1999. 31

[89] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, IEEE, 2005. 31

[90] G. Wang, D. Hoiem, and D. Forsyth, "Learning image similarity from flickr groups using stochastic intersection kernel machines," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 428–435, IEEE, 2009. 32

[91] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. International Conference on Computer Vision*, pp. 309–316, IEEE, 2009. 32

[92] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012. 32

[93] D. T. Nguyen, T. D. Pham, N. R. Baek, and K. R. Park, "Combining deep and handcrafted image features for presentation attack detection in face

recognition systems using visible-light camera sensors," *Sensors*, vol. 18, no. 3, p. 699, 2018. 32

[94] V. Jain and M. Varma, "Learning to re-rank: query-dependent image re-ranking using click data," in *Proceedings of the 20th international conference on World wide web*, pp. 277–286, ACM, 2011. 32

[95] Y. Hu, M. Li, and N. Yu, "Multiple-instance ranking: Learning to rank images for image retrieval," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008. 32

[96] A. Frome, Y. Singer, and J. Malik, "Image retrieval and classification using local distance functions," in *Advances in Neural Information Processing Systems*, pp. 417–424, 2007. 33

[97] D. Parikh and K. Grauman, "Relative attributes," in *Proc. International Conference on Computer Vision*, pp. 503–510, IEEE, 2011. 33

[98] T. Deselaers and V. Ferrari, "Visual and semantic similarity in imagenet," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1777–1784, IEEE, 2011. 33

[99] N. Friedman and S. Russell, "Image segmentation in video sequences: A probabilistic approach," in *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pp. 175–181, Morgan Kaufmann Publishers Inc., 1997. 34

[100] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russell, "Towards robust automatic traffic scene analysis in real-time," in *Proceedings of 12th International Conference on Pattern Recognition*, vol. 1, pp. 126–131, IEEE, 1994. 34

[101] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *European conference on computer vision*, pp. 282–295, Springer, 2010. 34

[102] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting.," in *BMVC*, vol. 2, p. 8, 2014. 34

[103] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1777–1784, 2013. 34

[104] M. Keuper, B. Andres, and T. Brox, "Motion trajectory segmentation via minimum cost multicuts," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3271–3279, 2015. 34

[105] J. Lezama, K. Alahari, J. Sivic, and I. Laptev, "Track to the future: Spatio-temporal video segmentation with long-range motion cues," in *CVPR 2011*, pp. 3369–3376, IEEE, 2011. 34

[106] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 6, pp. 1187–1200, 2013. 34

[107] S. D. Jain, B. Xiong, and K. Grauman, "Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos," in *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2117–2126, IEEE, 2017. 34

[108] P. Tokmakov, K. Alahari, and C. Schmid, "Learning video object segmentation with visual memory," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4481–4490, 2017. 34

[109] P. Tokmakov, K. Alahari, and C. Schmid, "Learning motion patterns in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3386–3394, 2017. 34

[110] S.-K. Weng, C.-M. Kuo, and S.-K. Tu, "Video object tracking using adaptive kalman filter," *Journal of Visual Communication and Image Representation*, vol. 17, no. 6, pp. 1190–1208, 2006. 34

[111] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proceedings of the 1981 DARPA Image Understanding Workshop*, pp. 121–130, 1981. 34

[112] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4836–4845, 2017. 35

[113] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE International Conference on Image Processing*, pp. 3645–3649, IEEE, 2017. 35, 36, 51, 53

[114] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015. 37

[115] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan, "Clothing attributes assisted person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, pp. 869–878, May 2015. 37

[116] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features,"

in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2360–2367, IEEE, 2010. 37

[117] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016. 37

[118] D. Conte, P. Foggia, G. Percannella, and M. Vento, "Performance evaluation of a people tracking system on pets2009 database," in *Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 119–126, Aug 2010. 37

[119] A. Bhuiyan, A. Perina, and V. Murino, "Exploiting multiple detections to learn robust brightness transfer functions in re-identification systems," in *Proc. IEEE International Conference on Image Processing*, pp. 2329–2333, Sep. 2015. 37

[120] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," in *Advances in Neural information processing systems*, pp. 41–48, 2004. 37

[121] J. Wald, A. Avetisyan, N. Navab, F. Tombari, and M. Niessner, "RIO: 3D object instance re-localization in changing indoor environments," in *Proceedings IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2019. 37

[122] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, IEEE, 2009. 46, 50

[123] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017. 49

[124] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE International Conference on Image Processing*, pp. 3464–3468, Sep. 2016. 53

[125] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," *CoRR*, vol. abs/1905.00953, 2019. 54

[126] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2138–2147, 2019. 55

[127] G. Cross and A. Zisserman, "Quadric reconstruction from dual-space geometry," in *Computer Vision, 1998. Sixth International Conference on*, pp. 25–31, IEEE, 1998. 63

[128] J. Goodman and J. Weare, "Ensemble samplers with affine invariance," *Communications in applied mathematics and computational science*, vol. 5, no. 1, pp. 65–80, 2010. 71

[129] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International Journal of Computer Vision*, vol. 81, pp. 155–166, 2009. 76

[130] "Google Project Tango." `https://support.google.com/faqs/faq/6029402?hl=en`. Accessed: 2019-11-13. 83

[131] "Lenovo Tango-enabled device." https://www.lenovo.com/in/en/tablets/android-tablets/tablet-phab-series/Lenovo-Phab-2-Pro/p/WMD00000220. Accessed: 2019-11-13. 90

[132] H. Izadinia, F. Sadeghi, and A. Farhadi, "Incorporating scene context and object layout into appearance modeling," in *Computer Vision and Pattern Recognition*, pp. 232–239, IEEE, 2014. 98

[133] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *European Conference on Computer Vision*, pp. 852–869, Springer, 2016. 98

[134] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2544–2550, IEEE, 2010.

[135] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. European Conference on Computer Vision*, pp. 472–488, Springer, 2016.

[136] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[137] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. European conference on computer vision*, pp. 850–865, Springer, 2016.

[138] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional

features for visual tracking," in *Proc. IEEE International Conference on Computer Vision*, pp. 3074–3082, 2015.

[139] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 fps with deep regression networks," in *Proc. European Conference on Computer Vision*, pp. 749–765, Springer, 2016.

[140] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE International Conference on Computer Vision*, pp. 3119–3127, 2015.

[141] Z. Kalal, K. Mikolajczyk, J. Matas, *et al.*, "Tracking-learning-detection," *Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, p. 1409, 2012.

[142] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *Proc. European Conference on Computer Vision*, pp. 788–801, Springer, 2008.

[143] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect,"

[144] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, pp. 379–387, 2016.

[145] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[146] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, *et al.*, "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems*, pp. 1223–1231, 2012.

[147] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[148] R. Girshick, "Fast r-cnn," in *Proc. IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.

[149] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[150] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proc. European Conference on Computer Vision*, pp. 21–37, Springer, 2016.

[151] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2288–2295, IEEE, 2012.

[152] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv:1603.00831 [cs]*, March 2016. arXiv: 1603.00831.

[153] D. Matinec and T. Pajdla, "Line reconstruction from many perspective images by factorization," in *Computer Vision and Pattern Recognition*, vol. 1, pp. 491–497, IEEE, 2003.

[154] A. Bartoli and P. Sturm, "Structure-from-motion using lines: Representation, triangulation, and bundle adjustment," *Computer Vision and Image Understanding*, vol. 100, no. 3, pp. 416–441, 2005.

[155] M. Kaess, R. Zboinski, and F. Dellaert, "Mcmc-based multiview reconstruction of piecewise smooth subdivision curves with a variable number of control points," in *European Conference on Computer Vision*, pp. 329–341, Springer, 2004.

[156] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Computer Vision and Pattern Recognition*, pp. 1271–1278, IEEE, 2009.

[157] M. Arsalan, A. Dragomir, and F. John, "3d bounding box estimation using deep learning and geometry," in *Computer Vision and Pattern Recognition*, pp. 1271–1278, IEEE, 2017.

[158] D. Jingming and S. Xiaohan, Fei Stefano, "Visual-inertial-semantic scene representation for 3d object detection," in *Computer Vision and Pattern Recognition*, pp. 1371–1378, IEEE, 2017.

[159] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *International Journal of Computer Vision*, vol. 80, no. 1, pp. 3–15, 2008.

[160] Y. J. Xiao and Y. Li, "Optimized stereo reconstruction of free-form space curves based on a nonuniform rational b-spline model," *JOSA A*, vol. 22, no. 9, pp. 1746–1762, 2005.

[161] T. Okatani, T. Yoshida, and K. Deguchi, "Efficient algorithm for low-rank matrix factorization with missing components and performance comparison of latest algorithms," in *International Conference on Computer Vision*, vol. 3, pp. 842–849, IEEE, 2011.

[162] A. Gruber and Y. Weiss, "Factorization with uncertainty and missing data: Exploiting temporal coherence," in *Advances in Neural Information Processing Systems*, MIT, 2003.

[163] R. Basri, D. Jacobs, and I. Kemelmacher, "Photometric stereo with general, unknown lighting," *International Journal of Computer Vision*, vol. 72, no. 3, pp. 239–257, 2007.

[164] T. E. Boult and L. G. Brown, "Factorization-based segmentation of motions," in *Visual Motion, 1991., Proceedings of the IEEE Workshop on*, pp. 179–186, IEEE, 1991.

[165] D. Martinec and T. Pajdla, "3d reconstruction by fitting low-rank matrices with missing data," in *Computer Vision and Pattern Recognition*, vol. 1, pp. 198–205, 2005.

[166] J. Costeira and T. Kanade, "A multi-body factorization method for motion analysis," in *ICCV 1995*, pp. 1071–1076, IEEE, 1995.

[167] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *IJCV 1992*, vol. 9, no. 2, pp. 137–154, 1992.

[168] D. Jacobs, "Linear fitting with missing data: Applications to structure-from-motion and to characterizing intensity images," in *Computer Vision and Pattern Recognition*, pp. 206–212, IEEE, 1997.

[169] H.-Y. Shum, K. Ikeuchi, and R. Reddy, "Principal component analysis with missing data and its application to polyhedral object modeling," *PAMI 1995*, vol. 17, no. 9, pp. 854–867, 1995.

[170] J. Oliensis and R. Hartley, "Iterative extensions of the sturm/triggs algorithm: Convergence and nonconvergence," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2217–2233, 2007.

[171] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3d shape from image streams," in *Computer Vision and Pattern Recognition*, vol. 2, pp. 690–696, IEEE, 2000.

[172] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *ECCV 2006*, pp. 94–106, Springer, 2006.

[173] P. Tresadern and I. Reid, "Articulated structure from motion by factorization," in *CVPR 2005*, vol. 2, pp. 1110–1115, IEEE, 2005.

[174] K. Kanatani and Y. Sugaya, "Factorization without factorization: complete recipe," *Memoirs of the Faculty of Engineering*, vol. 38, no. 1&2, pp. 61–72, 2004.

[175] B. Triggs, "Factorization methods for projective structure and motion," in *CVPR 1996*, pp. 845–851, IEEE, 1996.

[176] M. Irani, "Multi-frame optical flow estimation using subspace constraints," in *ICCV 1999*, vol. 1, pp. 626–633, IEEE, 1999.

[177] M. Z. Zia, M. Stark, and K. Schindler, "Are cars just 3d boxes? jointly estimating the 3d shape of multiple objects," in *CVPR 2014*, pp. 3678–3685, 2014.

[178] S. Mahamud, M. Hebert, Y. Omori, and J. Ponce, "Provably-convergent iterative methods for projective structure from motion," in *Computer Vision and Pattern Recognition*, vol. 1, pp. 1018–1025, IEEE, 2001.

[179] A. Aldoma, T. Faulhammer, and M. Vincze, "Automation of ground truth annotation for multi-view rgb-d object instance recognition datasets," in *IROS 2014*, pp. 5016–5023, IEEE, 2014.

[180] J. Andrews and C. H. Séquin, "Type-Constrained Direct Fitting of Quadric Surfaces," vol. 10, pp. 1–15, 2013.

[181] C. Ax and D. Hua, "On the Symmetric Solutions of Linear Matrix Equations," vol. 7, no. 2, pp. 1–7.

[182] Y. Bian, F. Dong, W. Zhang, H. Wang, C. Tan, and Z. Zhang, "3D reconstruction of single rising bubble in water using digital image processing and characteristic matrix," *Particuology*, vol. 11, pp. 170–183, April 2013.

[183] A. Eriksson and A. van den Hengel, "Efficient computation of robust low-rank matrix approximations in the presence of missing data using the l1 norm," in *Computer Vision and Pattern Recognition*, pp. 771–778, IEEE, 2010.

[184] K.-w. E. Chu, "Symmetric solutions of linear matrix equations by matrix decompositions," *Linear Algebra and its Applications*, vol. 119, pp. 35–50, July 1989.

[185] Y. Dai, H. Li, and M. He, "Element-wise factorization for n-view projective reconstruction," in *European Conference on Computer Vision*, vol. 6314, pp. 396–409, Springer, 2010.

[186] S. K. Divvala, A. A. Efros, and M. Hebert, "How important are "deformable parts" in the deformable parts model?," in *ECCV 2012*, pp. 31–40, Springer, 2012.

[187] A. M. Buchanan and A. W. Fitzgibbon, "Damped newton algorithms for matrix factorization with missing data," in *Computer Vision and Pattern Recognition*, vol. 2, pp. 316–322, IEEE, 2005.

[188] D. Eberly, "Reconstructing an Ellipsoid from its Perspective Projection onto a Plane Determining the Elliptical Cone Bounding the Ellipsoid," pp. 5–9, 2013.

[189] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI 2010*, vol. 32, no. 9, pp. 1627–1645, 2010.

[190] N. Fioraio and L. Di Stefano, "Joint detection, tracking and mapping by semantic bundle adjustment," in *Computer Vision and Pattern Recognition*, pp. 1538–1545, IEEE, 2013.

[191] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," *arXiv preprint arXiv:1609.05130*, 2016.

[192] N. Savinov, C. Häne, M. Pollefeys, *et al.*, "Discrete optimization of ray potentials for semantic 3d reconstruction," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5511–5518, IEEE, 2015.

[193] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[194] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3d traffic scene understanding from movable platforms," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 1012–1025, 2014.

[195] S. J. Prince, *Computer vision: models, learning, and inference*. Cambridge University Press, 2012.

[196] A. Gupta, A. A. Efros, and M. Hebert, "Blocks world revisited: Image understanding using qualitative geometry and mechanics," in *ECCV 2010*, pp. 482–496, Springer, 2010.

[197] P. Gurdjos and V. Charvillat, "Multiple View Reconstruction of a Quadric of Revolution from its Occluding Contours,"

[198] P. Gurdjos, V. Charvillat, and G. Morin, "Multiple View Reconstruction of a Quadric of Revolution from its Occluding Contours," no. september, p. 2009, 2009.

[199] A. B. Morgan, "Investigation into matrix factorization when elements are unknown," *Technical report*, 2004.

[200] M. Hejrati and D. Ramanan, "Analysis by synthesis: 3d object recognition by object reconstruction," in *CVPR 2014*, pp. 2449–2456, IEEE, 2014.

[201] H. V. Henderson and S. Searle, "Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics," *Canadian Journal of Statistics*, vol. 7, no. 1, pp. 65–81, 1979.

[202] S. Nakajima and M. Sugiyama, "Theoretical analysis of bayesian matrix factorization," *Journal of Machine Learning Research*, vol. 12, pp. 2583–2648, 2011.

[203] B. Nasihatkon, R. Hartley, and J. Trumpf, "A generalized projective reconstruction theorem and depth constraints for projective factorization," *International Journal of Computer Vision*, vol. 115, no. 2, pp. 1–28, 2015.

[204] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *ACCV 2012*, pp. 548–562, Springer, 2012.

[205] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *IJCV 2008*, vol. 80, pp. 3–15, October 2008.

[206] J.-j. Jacq, V. Burdin, C. Roux, G.-e. Bretagne, and C. Couture, "Type-Constrained Robust Fitting of Quadrics with Application to the 3D Morphological Characterization of Saddle-Shaped Articular Surfaces Stéphane Allaire," 2007.

[207] N. Jojic, A. Perina, M. Cristani, V. Murino, and B. Frey, "Stel component analysis: Modeling spatial correlations in image class structure," in *Computer Vision and Pattern Recognition*, pp. 2044–2051, IEEE, 2009.

[208] K. Kang, K. Kangbrownedu, J.-p. Tarel, L. C. Cedex, and D. B. Cooper, "A Unified Linear Fitting Approach for Singular and Non-Singular 3D Quadrics from Occluding Contours,"

[209] K. Kang, J.-P. Tarel, R. Fishman, and D. Cooper, "A linear dual-space approach to 3D surface reconstruction from occluding contours using algebraic surfaces," *ICCV 2001*, vol. 1, pp. 198–204, 2001.

[210] W. C. Karl, G. C. Verghese, and A. S. Willsky, "Reconstructing Ellipsoids from Projections *," no. 617, 1993.

[211] R. G. Lorenz and S. P. Boyd, "Robust minimum variance beamforming," *Signal Processing, IEEE Transactions on*, vol. 53, no. 5, pp. 1684–1696, 2005.

[212] S. Ma, "Conics-based stereo, motion estimation, and pose determination," *IJCV 1993*, vol. 10, pp. 7–25, February 1993.

[213] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[214] N. Pattern, "Ellipsoid Reconstruction from Three Perspective Views*," pp. 344–348, 1996.

[215] M. Pollefeys, R. Koch, and L. Van Gool, "Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters," *IJCV 1999*, vol. 32, no. 1, pp. 7–25, 1999.

[216] A. Heyden, R. Berthilsson, and G. Sparr, "An iterative factorization method for projective structure and motion from image sequences," *Image and Vision Computing*, vol. 17, no. 13, pp. 981–991, 1999.

[217] J. Hyeong Hong and A. Fitzgibbon, "Secrets of matrix factorization: Approximations, numerics, manifold optimization and random restarts," in *International Conference on Computer Vision*, pp. 4130–4138, IEEE, 2015.

[218] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *CVPR 2013*, pp. 1352–1359, IEEE, 2013.

[219] P. Sturm and B. Triggs, "A factorization based algorithm for multi-image projective structure and motion," in *ECCV 1996*, pp. 709–720, Springer, 1996.

[220] Q. Surface and F. Occluding, "Quadric Surface From Occluding Contour," pp. 27–31, 1994.

[221] A. N. Tikhonov and V. Y. Arsenin, "Solutions of ill-posed problems," 1977.

[222] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[223] P. Aguiar, J. Xavier, and M. Stosic, "Spectrally optimal factorization of incomplete matrices," in *Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2008.

[224] S. Vicente, J. Carreira, L. Agapito, and J. Batista, "Reconstructing pascal voc," in *CVPR 2014*, pp. 41–48, IEEE, 2014.

[225] D. S. Wokes and P. L. Palmer, "Perspective Reconstruction of a Spheroid from an Image Plane Ellipse," *IJCV 2010*, vol. 90, pp. 369–379, July 2010.

[226] Y. Xiang and S. Savarese, "Estimating the aspect layout of object categories," in *CVPR 2012*, pp. 3410–3417, IEEE, 2012.

[227] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler, "Detailed 3d representations for object recognition and modeling," *PAMI 2013*, vol. 35, no. 11, pp. 2608–2623, 2013.

[228] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV 2012*, pp. 746–760, Springer, 2012.

[229] C. M. Bishop, "Bayesian pca," in *Advances in Neural Information Processing Systems*, pp. 382–388, MIT, 1999.

[230] S. Christy and R. Horaud, "Euclidean shape and motion from multiple perspective views by affine iterations," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 11, pp. 1098–1104, 1996.