References

[1] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[2] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proc. of ICCV*, 2011, pp. 2556–2563.

[3] R. Nawaratne, D. Alahakoon, D. De Silva, and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Trans. on Industrial Informatics*, 2019, vol. 16, no. 1, pp. 393–402.

[4] H. Liu, T. Liu, Z. Zhang, A. K. Sangaiah, B. Yang, and Y. Li, "ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human-computer interaction," *IEEE Trans. on Industrial Informatics*, 2022, vol. 18, no. 10, pp. 7107–7117.

[5] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proc. of CVPR*, 2018, pp. 8739–8748.

[6] A. G. Del Molino, C. Tan, J.-H. Lim, and A.-H. Tan, "Summarization of egocentric videos: A comprehensive survey," *IEEE Trans. on Human-Machine Systems*, 2016, vol. 47, no. 1, pp. 65–76.

[7] M. Yu, A. Rhuma, S. M. Naqvi, L. Wang, and J. Chambers, "A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment," *IEEE Trans. on ITB*, 2012, vol. 16, no. 6, pp. 1274–1286.

[8] J. Fan, W. Xu, Y. Wu, and Y. Gong, "Human tracking using convolutional neural networks," *IEEE Trans. on Neural Networks*, 2010, vol. 21, no. 10, pp. 1610–1623.

[9] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proc. of CVPR*, 2013, pp. 3626–3633.

[10] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. of the ICCV*, 2015, pp. 3119–3127.

[11] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *preprint arXiv:1406.2199*, 2014, pp. 1–10.

[12] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. of CVPR*, 2015, pp. 4489–4497.

[13] R. Christoph and F. A. Pinz, "Spatiotemporal residual networks for video action recognition," *Proc. of NIPS*, 2016, pp. 3468–3476.

[14] L. Wang, W. Li, W. Li, and L. Van Gool, "Appearance-and-relation networks for video classification," in *Proc. of CVPR*, 2018, pp. 1430–1439.

[15] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proc. of CVPR*, 2018, pp. 6546–6555.

[16] O. Kopuklu, N. Kose, A. Gunduz, and G. Rigoll, "Resource efficient 3d convolutional neural networks," in *Proc. of the ICCVW*, 2019, pp. 1–10.

[17] J. Platt and S. Nowlan, "A convolutional neural network hand tracker," *Proc. of NIPS*, 1995, pp. 901–908.

[18] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler, "Learning human pose estimation features with convolutional networks," *preprint arXiv:1312.7302*, 2013, pp. 1–10.

[19] A. Jain, J. Tompson, Y. LeCun, and C. Bregler, "Modeep: A deep learning framework using motion features for human pose estimation," in *Proc. of ACCV*, 2015, pp. 302–315.

[20] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *Proc. of CVPR*, 2015, pp. 2568–2577.

[21] J. Shao, K. Kang, C. Change Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proc. of CVPR*, 2015, pp. 4657–4666.

[22] Y. Xiong, K. Zhu, D. Lin, and X. Tang, "Recognize complex events from static images by fusing deep channels," in *Proc. of CVPR*, 2015, pp. 1600–1609.

[23] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, 1980, vol. 36, no. 4, pp. 193–202.

[24] R. Sigala, T. Serre, T. Poggio, and M. Giese, "Learning features of intermediate complexity for the recognition of biological motion," in *Proc. of ICANN*, 2005, pp. 241–246.

[25] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *2007 IEEE 11th international conference on computer vision*, 2007, pp. 1–8.

[26] M. A. Giese and T. Poggio, "Neural mechanisms for the recognition of biological movements," *Nature Reviews Neuroscience*, 2003, vol. 4, no. 3, pp. 179–192.

[27] H.-J. Kim, J. S. Lee, and H.-S. Yang, "Human action recognition using a modified convolutional neural network," in *Proc. of International Symposium on Neural Networks*, 2007, pp. 715–723.

[28] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. on PAMI*, 2012, vol. 35, no. 1, pp. 221–231.

[29] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo, "3d human activity recognition with reconfigurable convolutional neural networks," in *Proc. of ACM*, 2014, pp. 97–106.

[30] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. on PAMI*, 2017, vol. 40, no. 6, pp. 1510–1517.

[31] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proc. of ICCV*, 2015, pp. 4597–4605.

[32] E. P. Ijjina and C. K. Mohan, "Human action recognition based on recognition of linear patterns in action bank features using convolutional neural networks," in *Proc. of ICMA*, 2014, pp. 178–182.

[33] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. of CVPR*, 2012, pp. 1234–1241.

[34] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Transactions on Graphics (ToG)*, 2014, vol. 33, no. 5, pp. 1–0.

[35] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proc. of ICCV*, 2015, pp. 3218–3226.

[36] Y. Wang, J. Song, L. Wang, L. Van Gool, and O. Hilliges, "Two-Stream SR-CNNs for Action Recognition in Videos." in *Bmvc*, 2016.

[37] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. of ECCV*, 2016, pp. 20–36.

[38] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "Untrimmednets for weakly supervised action recognition and detection," in *Proc. of CVPR*, 2017, pp. 4325–4334.

[39] Y. Xiong, L. Wang, Z. Wang, B. Zhang, H. Song, W. Li, D. Lin, Y. Qiao, L. Van Gool, and X. Tang, "Cuhk & ethz & siat submission to activitynet challenge 2016," *arXiv preprint arXiv:1608.00797*, 2016.

[40] C. Liu, W. Xu, Q. Wu, and G. Yang, "Learning motion and content-dependent features with convolutions for action recognition," *Multimedia Tools and Applications*, 2016, vol. 75, pp. 13 023–13 039.

[41] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proc. of ICPR*, vol. 3, 2004, pp. 32–36.

[42] S. Singh, C. Arora, and C. Jawahar, "First person action recognition using deep learned descriptors," in *Proc. of CVPR*, 2016, pp. 2620–2628.

[43] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimedia*, 2012, vol. 19, no. 2, pp. 4–10.

[44] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. Ogunbona, "Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring," in *Proc. of ACM*, 2015, pp. 1119–1122.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, 2017, vol. 60, no. 6, pp. 84–90.

[46] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Trans. on Human-Machine Systems*, 2015, vol. 46, no. 4, p. 498509.

[47] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, "Deep convolutional neural networks for action recognition using depth map sequences," *preprint arXiv:1501.04686*, 2015.

[48] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Proc. of CVPRW*. IEEE, 2010, pp. 9–14.

[49] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Proc. of CVPRW*, 2012, pp. 20–27.

[50] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. of CVPR*, 2012, pp. 1290–1297.

[51] T. Dobhal, V. Shitole, G. Thomas, and G. Navada, "Human activity recognition using binary motion image and deep learning," *Procedia computer science*, 2015, vol. 58, pp. 178–185.

[52] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, 1995, vol. 3361, no. 10, p. 1995.

[53] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. of ICCV*, vol. 2. IEEE, 2005, pp. 1395–1402.

[54] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. of CVPR*, 2014, pp. 1725–1732.

[55] H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, "Video-based human action recognition using deep learning: a review," *preprint arXiv:2208.03775*, 2022.

[56] L. Mo, F. Li, Y. Zhu, and A. Huang, "Human physical activity recognition based on computer vision with deep learning model," in *Proc. of IIMT*, 2016, pp. 1–6.

[57] D. W. Ruck, S. K. Rogers, and M. Kabrisky, "Feature selection using a multilayer perceptron," *Journal of Neural Network Computing*, 1990, vol. 2, no. 2, pp. 40–48.

[58] P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," *Knowledge-Based Systems*, 2018, vol. 158, pp. 43–53.

[59] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Trans. on Circuits and Systems for Video Technology*, 2016, vol. 28, no. 3, pp. 807–811.

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

[61] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. of CVPR*, 2015, pp. 1–9.

[62] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," *preprint arXiv:1507.02159*, 2015.

[63] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. of CVPR*, 2016, pp. 1933–1941.

[64] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[65] L. Wang, Z. Wang, Y. Xiong, and Y. Qiao, "CUHK&SIAT submission for thumos15 action recognition challenge," *THUMOS Action Recognition challenge*, 2015, pp. 1–3.

[66] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. of CVPR*, 2009, pp. 2929–2936.

[67] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of ICCV*, 2013, pp. 3551–3558.

[68] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. of CVPR*, 2011, pp. 3169–3176.

[69] C. Beaudry, R. Péteri, and L. Mascarilla, "An efficient and sparse approach for large scale human action recognition in videos," *Machine vision and Applications*, 2016, vol. 27, pp. 529–543.

[70] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Action recognition with joints-pooled 3d deep convolutional descriptors." in *IJCAI*, vol. 1, 2016, p. 3.

[71] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. of ICCV*, 2013, pp. 3192–3199.

[72] W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *Proc. of ICCV*, 2013, pp. 2248–2255.

[73] I. Lillo, A. Soto, and J. Carlos Niebles, "Discriminative hierarchical modeling of spatio-temporally composable human activities," in *Proc. of CVPR*, 2014, pp. 812–819.

[74] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proc. of CVPR*, 2016, pp. 3034–3042.

[75] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of CVPR*, 2016, pp. 770–778.

[76] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. of CVPR*, 2017, pp. 4768–4777.

[77] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proc. of HBU*, 2011, pp. 29–39.

[78] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. of CVPR*, 2015, pp. 4694–4702.

[79] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. of CVPR*, 2015, pp. 2625–2634.

[80] A. Giel and R. Diaz, "Recurrent neural networks and transfer learning for action recognition," 2015.

[81] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proc. of CVPR*, 2016, pp. 1971–1980.

[82] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proc. of CVPR*, 2016, pp. 1961–1970.

[83] Q. Li, Z. Qiu, T. Yao, T. Mei, Y. Rui, and J. Luo, "Action recognition by learning deep multi-granular spatio-temporal video representation," in *Proc. of ACM*, 2016, pp. 159–166.

[84] J. Wu, G. Wang, W. Yang, and X. Ji, "Action recognition with joint attention on multi-level deep features," *arXiv preprint arXiv:1607.02556*, 2016.

[85] H. Chen, J. Chen, R. Hu, C. Chen, and Z. Wang, "Action recognition with temporal scale-invariant deep learning framework," *China Communications*, 2017, vol. 14, no. 2, pp. 163–172.

[86] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. of CVPR*, 2015, pp. 1110–1118.

[87] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. of AAAI*, vol. 31, no. 1, 2017.

[88] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. of CVPR*, vol. 30, no. 1, 2016.

[89] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in *Proc. of ECCV*, 2016, pp. 203–220.

[90] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *Proc. of ECCV*, 2016, pp. 816–833.

[91] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3D human-skeleton sequences for action recognition," in *Proc. of CVPR*, 2016, pp. 3054–3062.

[92] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, 1996, vol. 381, no. 6583, pp. 607–609.

[93] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," *Proc. of NIPS*, 2006, vol. 19.

[94] K. Yu, Y. Lin, and J. Lafferty, "Learning image representations from the pixel level via hierarchical sparse coding," in *Proc. of CVPR*. IEEE, 2011, pp. 1713–1720.

[95] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proc.of ICML*, 2007, pp. 759–766.

[96] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," in *Proc. of CVPR*, 2010, pp. 3517–3524.

[97] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. of CVPR*, 2009, pp. 1794–1801.

[98] Y. Zhu, X. Zhao, Y. Fu, and Y. Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," in *Proc. of ACCV*, 2011, pp. 660–671.

[99] Z. Lu and Y. Peng, "Latent semantic learning by efficient sparse coding with hypergraph regularization," in *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[100] ——, "Latent semantic learning with structured sparse representation for human action recognition," *Pattern Recognition*, 2013, vol. 46, no. 7, pp. 1799–1809.

[101] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Trans. on PAMI*, 2011, vol. 34, no. 8, pp. 1576–1588.

[102] A. Alfaro, D. Mery, and A. Soto, "Action recognition in video using sparse coding and relative features," in *Proc. of CVPR*, 2016, pp. 2688–2697.

[103] I. Ullah and A. Petrosino, "A strict pyramidal deep neural network for action recognition," in *Proc. of ICIAP*, 2015, pp. 236–245.

[104] ——, "Spatiotemporal features learning with 3DPyraNet," in *Proc. of ACIVS*, 2016, pp. 638–647.

[105] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," *IEEE Trans. on PAMI*, 2017, vol. 40, no. 3, pp. 667–681.

[106] S. L. Phung and A. Bouzerdoum, "A pyramidal neural network for visual pattern recognition," *IEEE Trans. on neural networks*, 2007, vol. 18, no. 2, pp. 329–343.

[107] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *CVPR 2011*, 2011, pp. 3361–3368.

[108] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural image statistics: A probabilistic approach to early computational vision.* Springer Science & Business Media, 2009, vol. 39.

[109] M. Mudrova and A. Procházka, "Principal component analysis in image processing," in *Proc. of the MATLAB technical computing conference, Prague*, 2005.

[110] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. of CVPR*, 2008, pp. 1–8.

[111] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *Proc. of CVPR*, 2009, pp. 1996–2003.

[112] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *Proc. of ICML*, 2015, pp. 843–852.

[113] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei, "Unsupervised learning of long-term motion dynamics for videos," in *Proc. of CVPR*, 2017, pp. 2203–2212.

[114] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *Proc. of NIPS*, 2017, pp. 34–45.

[115] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. of ECCV*, 2018, pp. 305–321.

[116] X. Long, C. Gan, G. De Melo, J. Wu, X. Liu, and S. Wen, "Attention clusters: Purely attention based local feature integration for video classification," in *Proc. of CVPR*, 2018, pp. 7834–7843.

[117] N. Hussein, E. Gavves, and A. W. Smeulders, "PIC: Permutation Invariant Convolution for Recognizing Long-range Activities," *preprint arXiv:2003.08275*, 2020, pp. 1–10.

[118] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. of ECCV*, 2018, pp. 401–417.

[119] T. Nagarajan, C. Feichtenhofer, and K. Grauman, "Grounded human-object interaction hotspots from video," in *Proc. of ICCV*, 2019, pp. 8688–8697.

[120] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. of CVPR*, 2015, pp. 4694–4702.

[121] N. Hussein, E. Gavves, and A. W. Smeulders, "Timeception for complex action recognition," in *Proc. of CVPR*, 2019, pp. 254–263.

[122] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, "Deformable 3D Convolution for Video Super-Resolution," *preprint arXiv:2004.02803*, 2020, pp. 1–5.

[123] M. Pominova, E. Kondrateva, M. Sharaev, A. Bernstein, S. Pavlov, and E. Burnaev, "3D Deformable Convolutions for MRI classification," in *Proc. of ICMLA*, 2019, pp. 1710–1716.

[124] S. Sun, J. Pang, J. Shi, S. Yi, and W. Ouyang, "Fishnet: A versatile backbone for image, region, and pixel level prediction," in *Proc. of NIPS*, 2018, pp. 754–764.

[125] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. of ICCV*, 2017, pp. 764–773.

[126] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of ICML*, 2015, p. 448–456.

[127] R. A. Rensink, "The dynamic representation of scenes," *Visual Cognition*, 2000, vol. 7, no. 1-3, pp. 17–42.

[128] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proc. of CVPR*, 2018, pp. 8359–8367.

[129] X. Wang, Z. Miao, R. Zhang, and S. Hao, "I3d-lstm: A new model for human action recognition," in *IOP Conference Series: Materials Science and Engineering*, vol. 569, no. 3, 2019, pp. 32–35.

[130] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. of NIPS*, 2015, pp. 802–810.

[131] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. of ICML*, 2015, pp. 2048–2057.

[132] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. on PAMI*, 2017, vol. 40, no. 6, pp. 1452–1464.

[133] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier *et al.*, "The kinetics human action video dataset," *preprint arXiv:1705.06950*, 2017, pp. 1–22.

[134] R. Goyal, S. E. Kahou, V. Michalski *et al.*, "The Something Something Video Database for Learning and Evaluating Visual Common Sense." in *Proc. of ICCV*, 2017, pp. 5843–5851.

[135] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding," in *Proc. of CVPR*, 2015, pp. 961–970.

[136] H. Kuehne, A. Arslan, and T. Serre, "The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities," in *Proc. of CVPR*, 2014, pp. 780–787.

[137] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. of CVPR*, 2018, pp. 7794–7803.

[138] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "STM: SpatioTemporal and motion encoding for action recognition," in *Proc. of ICCV*, 2019, pp. 2000–2009.

[139] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "Convnet architecture search for spatiotemporal feature learning," *preprint arXiv:1708.05038*, 2017, pp. 1–12.

[140] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. of CVPR*, 2017, pp. 6299–6308.

[141] Y. Zhou, X. Sun, Z.-J. Zha, and W. Zeng, "Mict: Mixed 3d/2d convolutional tube for human action recognition," in *Proc. of CVPR*, 2018, pp. 449–458.

[142] D. He, Z. Zhou, C. Gan, F. Li, X. Liu, Y. Li, L. Wang, and S. Wen, "Stnet: Local and global spatial-temporal modeling for action recognition," in *Proc. of AAAI*, 2019, pp. 8401–8408.

[143] M. Zolfaghari, K. Singh, and T. Brox, "Eco: Efficient convolutional network for online video understanding," in *Proc. of ECCV*, 2018, pp. 695–712.

[144] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek, "Videolstm convolves, attends and flows for action recognition," *Computer Vision and Image Understanding*, 2018, vol. 166, pp. 41–50.

[145] A. Diba, M. Fayyaz, V. Sharma, M. Mahdi Arzani, R. Yousefzadeh, J. Gall, and L. Van Gool, "Spatio-temporal channel correlation networks for action classification," in *Proc. of ECCV*, 2018, pp. 284–299.

[146] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: Motion-Augmented RGB Stream for Action Recognition," in *Proc. of CVPR*, 2019, pp. 7874–7883.

[147] L. Zhu, D. Tran, L. Sevilla-Lara, Y. Yang, M. Feiszli, and H. Wang, "FASTER Recurrent Networks for Efficient Video Classification." in *Proc. of AAAI*, 2020, pp. 13 098–13 105.

[148] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. of ECCV*, 2018, pp. 399–417.

[149] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. of CVPR*, 2018, pp. 6450–6459.

[150] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. of ICCV*, 2019, pp. 6202–6211.

[151] D. Tran, H. Wang, L. Torresani, and M. Feiszli, "Video classification with channel-separated convolutional networks," in *Proc. of ICCV*, 2019, pp. 5552–5561.

[152] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *Proc. of ICCV*, 2017, pp. 5533–5541.

[153] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L 1 optical flow," in *Proc. of JPRS*, 2007, pp. 214–223.

[154] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in *Proc. of ECCV*, 2018, pp. 715–731.

[155] L. Yang, J. Han, T. Zhao, T. Lin, D. Zhang, and J. Chen, "Background-Click Supervision for Temporal Action Localization," *IEEE Trans. on PAMI*, 2021, pp. 1–15.

[156] T. Zhao, J. Han, L. Yang, B. Wang, and D. Zhang, "Soda: Weakly supervised temporal action localization based on astute background response and self-distillation learning," *International Journal of Computer Vision*, 2021, vol. 129, no. 8, pp. 2474–2498.

[157] D. Zhang, H. Tian, and J. Han, "Few-cost salient object detection with adversarial-paced learning," *Proc. of NIPS*, 2020, vol. 33, pp. 12 236–12 247.

[158] D. Zhang, J. Han, Y. Zhang, and D. Xu, "Synthesizing supervision for learning deep saliency network without human annotation," *IEEE trans. on PAMI*, 2019, vol. 42, no. 7, pp. 1755–1769.

[159] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proc. of ICCV*, 2019, pp. 7083–7093.

[160] Z. Qiu, T. Yao, C.-W. Ngo, X. Tian, and T. Mei, "Learning spatio-temporal representation with local and global diffusion," in *Proc. of CVPR*, 2019, pp. 12 056–12 065.

[161] Y. Quan, Y. Chen, R. Xu, and H. Ji, "Attention with structure regularization for action recognition," *Computer Vision and Image Understanding*, 2019, vol. 187, p. 102794.

[162] N. Nigam, T. Dutta, and H. P. Gupta, "FactorNet: Holistic Actor, Object and Scene Factorization for Action Recognition in Videos," *IEEE Trans. on CSVT.*, 2021, pp. 1–15.

[163] Z. Zheng, G. An, D. Wu, and Q. Ruan, "Global and local knowledge-aware attention network for action recognition," *IEEE Trans. on NNLS*, 2020, vol. 32, no. 1, pp. 334–347.

[164] M. Esat Kalfaoglu, S. Kalkan, and A. Aydin Alatan, "Late Temporal Modeling in 3D CNN Architectures with BERT for Action Recognition," *arXiv e-prints*, 2020, pp. 1–19.

[165] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *preprint arXiv:2010.11929*, 2020, pp. 1–22.

[166] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proc. of ICCV*, 2021, pp. 6836–6846.

[167] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "TMMF: Temporal Multi-Modal Fusion for Single-Stage Continuous Gesture Recognition," *IEEE Trans. on IP.*, 2021, vol. 30, pp. 7689–7701.

[168] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, and G. Zhao, "iMiGUE: An identity-free video dataset for micro-gesture understanding and emotion analysis," in *Proc. of CVPR*, 2021, pp. 10 631–10 642.

[169] X. Liu and G. Zhao, "3d skeletal gesture recognition via sparse coding of time-warping invariant riemannian trajectories," in *Proc. of ICMM*, 2019, pp. 678–690.

[170] X. Liu, H. Shi, X. Hong, H. Chen, D. Tao, and G. Zhao, "3D skeletal gesture recognition via hidden states exploration," *IEEE Trans. on IP.*, 2020, vol. 29, pp. 4583–4597.

[171] R. Cui, H. Liu, and C. Zhang, "A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training," *IEEE Trans. on Multimedia*, 2019, vol. 21, no. 7, pp. 1880–1891.

[172] J. Pu, W. Zhou, and H. Li, "Dilated Convolutional Network with Iterative Optimization for Continuous Sign Language Recognition," in *Proc. of AAAI*, 2018, p. 885–891.

[173] Z. Yu, B. Zhou, J. Wan, P. Wang, H. Chen, X. Liu, S. Z. Li, and G. Zhao, "Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition," *IEEE Trans. on IP.*, 2021, vol. 30, pp. 5626–5640.

[174] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Video emotion recognition with transferred deep feature encodings," in *Proc. of ACM*, 2016, pp. 15–22.

[175] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, and K. Keutzer, "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *Proc. of AAAI*, vol. 34, no. 01, 2020, pp. 303–311.

[176] B. Martinez, D. Modolo, Y. Xiong, and J. Tighe, "Action recognition with spatial-temporal discriminative filter banks," in *Proc. of ICCV*, 2019, pp. 5482–5491.

[177] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. of ICML*, 2010, pp. 807–814.

[178] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. of ICMLW*, vol. 3, no. 2, 2013.

[179] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one*, 2018, vol. 13, no. 5, pp. 1–35.

[180] J. Wan, C. Lin, L. Wen, Y. Li, Q. Miao, S. Escalera, G. Anbarjafari, I. Guyon, G. Guo, and S. Z. Li, "ChaLearn Looking at People: IsoGD and ConGD Large-Scale RGB-D Gesture Recognition," *IEEE Trans. on Cybernetics*, 2020, pp. 1–12.

[181] S. Deandrea, E. Lucenteforte, F. Bravi, R. Foschi, C. Vecchia, and E. Negri, "Risk Factors for Falls in Community-Dwelling Older People: A Systematic Review and Meta-Analysis," *Epidemiology (Cambridge, Mass.)*, 09 2010, vol. 21, pp. 658–68.

[182] A. Guillochon, C. Crinquette, C. Gaxatte, V. Pardessus, S. Bombois, V. Deramecourt, E. Boulanger, and F. Puisieux, "[Neurological diseases detected in the Lille Multidisciplinary Falls Consultation]." *Revue neurologique*, 08 2009, vol. 166, pp. 235–41.

[183] Y. Liu, J. S. Chan, and J. H. Yan, "Neuropsychological mechanisms of falls in older adults," *Frontiers in aging neuroscience*, 2014, vol. 6, no. 64, pp. 1–8.

[184] M. C. Nevitt, S. R. Cummings, and E. S. Hudes, "Risk factors for injurious falls: a prospective study," *Journal of Gerontology*, 1991, vol. 46, no. 5, pp. M164–M170.

[185] D. Epstein, B. Chen, and C. Vondrick, "Oops! predicting unintentional action in video," in *Proc. of CVPR*, 2020, pp. 919–929.

[186] G. Zhao, Z. Mei, D. Liang, K. Ivanov, Y. Guo, Y. Wang, and L. Wang, "Exploration and Implementation of a Pre-Impact Fall Recognition Method Based on an Inertial Body Sensor Network," *Sensors*, 2012, vol. 12, no. 11, p. 15338–15355.

[187] H. Lee, M. Jung, and J. Tani, "Recognition of Visually Perceived Compositional Human Actions by Multiple Spatio-Temporal Scales Recurrent Neural Networks," *IEEE Trans. on Cognitive and Developmental Systems*, 2018, vol. 10, no. 4, pp. 1058–1069.

[188] S. A. W. Talha, M. Hammouche, E. Ghorbel, A. Fleury, and S. Ambellouis, "Features and Classification Schemes for View-Invariant and Real-Time Human Action Recognition," *IEEE Trans. on Cognitive and Developmental Systems*, 2018, vol. 10, no. 4, pp. 894–902.

[189] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li, "Trear: Transformer-based RGB-D Egocentric Action Recognition," *IEEE Trans. on Cognitive and Developmental Systems*, 2020, pp. 1–1.

[190] A. Shojaei-Hashemi, P. Nasiopoulos, J. J. Little, and M. T. Pourazad, "Video-based Human Fall Detection in Smart Homes Using Deep Learning," in *Proc. of ISCAS*, 2018, pp. 1–5.

[191] Z. Huang, Y. Liu, Y. Fang, and B. K. P. Horn, "Video-based Fall Detection for Seniors with Human Pose Estimation," in *Proc. of UV*, 2018, pp. 1–4.

[192] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. on KDE*, 2009, vol. 21, no. 9, pp. 1263–1284.

[193] K. E. Bennin, J. Keung, P. Phannachitta, A. Monden, and S. Mensah, "Mahakil: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction," *IEEE Trans. on Software Engineering*, 2017, vol. 44, no. 6, pp. 534–550.

[194] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. of CVPR*, 2019, pp. 9268–9277.

[195] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. of CVPR*, 2016, pp. 4004–4012.

[196] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *preprint arXiv:1808.01340*, 2018, pp. 1–6.

[197] D. Epstein and C. Vondrick, "Learning Goals from Failure," *preprint arXiv:2006.15657*, 2020, pp. 1–10.

[198] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "Convnet architecture search for spatiotemporal feature learning," *preprint arXiv:1708.05038*, 2017.

[199] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. of ICCV*, 2017, pp. 764–773.

[200] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *preprint arXiv:1511.07122*, 2015, pp. 1–13.

[201] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam, "CASENet: Deep Category-Aware Semantic Edge Detection," in *Proc. of CVPR*, 2017, pp. 1761–1770.

[202] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. of ICCV*, 2017, pp. 2980–2988.

[203] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. of ECCV*, 2018, pp. 305–321.

[204] X. Wang, X. Xiong, M. Neumann, A. Piergiovanni, M. S. Ryoo, A. Angelova, K. M. Kitani, and W. Hua, "Attentionnas: Spatiotemporal attention cell search for video classification," in *Proc. of ECCV*. Springer, 2020, pp. 449–465.

[205] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?" *preprint arXiv:2102.05095*, 2021, pp. 1–13.

[206] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. of CVPR*, 2016, pp. 2921–2929.

[207] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "Stm: Spatiotemporal and motion encoding for action recognition," in *Proc. of ICCV*, 2019, pp. 1–10.

[208] W. Wu, D. He, T. Lin, F. Li, C. Gan, and E. Ding, "MVFNet: Multi-View Fusion Network for Efficient Video Recognition," *preprint arXiv:2012.06977*, 2020, pp. 1–10.

[209] N. Nigam, T. Dutta, and D. Verma, "Fall-perceived Action Recognition of Persons with Neurological Disorders using Semantic Supervision," *IEEE Trans. on Cognitive and Developmental Systems*, 2022, pp. 1–10.

[210] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, 2017, pp. 1273–1282.

[211] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," *preprint arXiv:2002.06440*, 2020, pp. 1–10.

[212] W. Sun, S. Lei, L. Wang, Z. Liu, and Y. Zhang, "Adaptive federated learning and digital twin for industrial internet of things," *IEEE Trans. on Industrial Informatics*, 2020, vol. 17, no. 8, pp. 5605–5614.

[213] B. Jiang, J. Li, H. Wang, and H. Song, "Privacy-Preserving Federated Learning for Industrial Edge Computing via Hybrid Differential Privacy and Adaptive Compression," *IEEE Trans. on Industrial Informatics*, 2021, pp. 1–10.

[214] L. Ahmed, K. Ahmad, N. Said, B. Qolomany, J. Qadir, and A. Al-Fuqaha, "Active learning based federated learning for waste and natural disaster image classification," *IEEE Access*, 2020, vol. 8, pp. 208 518–208 531.

[215] P. Yu and Y. Liu, "Federated object detection: Optimizing object detection model with federated learning," in *Proc. of VISP*, 2019, pp. 1–6.

[216] Y. Liu, A. Huang, Y. Luo *et al.*, "Fedvision: An online visual object detection platform powered by federated learning," in *Proc. of AAAI*, vol. 34, no. 08, 2020, pp. 13 172–13 179.

[217] K. Doshi and Y. Yilmaz, "Federated learning-based driver activity recognition for edge devices," in *Proc. of CVPR*, 2022, pp. 3338–3346.

[218] B. A. Erol, A. Majumdar, P. Benavidez, P. Rad, K.-K. R. Choo, and M. Jamshidi, "Toward Artificial Emotional Intelligence for Cooperative Social Human–Machine Interaction," *IEEE Trans. on Computational Social Systems*, 2020, vol. 7, no. 1, pp. 234–246.

[219] I. Dave, Z. Scheffer, A. Kumar, S. Shiraz, Y. S. Rawat, and M. Shah, "GabriellaV2: Towards better generalization in surveillance videos for Action Detection," in *Proc. of WACVW*, 2022, pp. 122–132.

[220] W. Li, G. Zeng, J. Zhang *et al.*, "CogEmoNet: A Cognitive-Feature-Augmented Driver Emotion Recognition Model for Smart Cockpit," *IEEE Trans. on CSS.*, 2021, pp. 1–12.

[221] M. Buzzelli, A. Albé, and G. Ciocca, "A Vision-Based System for Monitoring Elderly People at Home," *Applied Sciences*, 2020, vol. 10.

[222] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. on PAMI*, 2017, vol. 40, no. 6, pp. 1510–1517.

[223] H. Li, J. Huang, M. Zhou, Q. Shi, and Q. Fei, "Self-attention Pooling-based Long-term Temporal Network for Action Recognition," *IEEE Trans. on Cognitive and Developmental Systems*, 2022, pp. 1–1.

[224] A. Miech, J.-B. Alayrac, I. Laptev, J. Sivic, and A. Zisserman, "RareAct: A video dataset of unusual interactions," *arxiv:2008.01018*, 2020, pp. 1–5.

[225] C. Gao, Y. Zou, and J.-B. Huang, "ican: Instance-centric attention network for human-object interaction detection," *preprint arXiv:1808.10437*, 2018, pp. 1–10.

[226] T. Gupta, A. Schwing, and D. Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," in *Proc. of CVPR*, 2019, pp. 9677–9685.

[227] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning Human-Object Interactions by Graph Parsing Neural Networks," in *Proc. of ECCV*, 2018, pp. 407–423.

[228] T. Nagarajan, C. Feichtenhofer, and K. Grauman, "Grounded Human-Object Interaction Hotspots From Video," in *Proc. of ICCV*, 2019, pp. 8687–8696.

[229] H. Wu, X. Ma, and Y. Li, "Convolutional networks with channel and STIPs attention model for action recognition in videos," *IEEE Trans. on Multimedia*, 2019, vol. 22, no. 9, pp. 2293–2306.

[230] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Weakly-supervised learning of visual relations," in *Proc. of ICCV*, 2017, pp. 5179–5188.

[231] S. Huang and D. Ramanan, "Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters," in *Proc. of CVPR*, 2017, pp. 2243–2252.

[232] D. Epstein, B. Chen, and C. Vondrick, "Oops! Predicting Unintentional Action in Video," in *Proc. of CVPR*, 2020, pp. 1–2.

[233] P. Jain, S. Goenka, S. Bagchi, B. Banerjee, and S. Chaterji, "Federated Action Recognition on Heterogeneous Embedded Devices," *IEEE Trans. on Airtificial Intelligence*, 2021, pp. 1–10.

[234] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proc.of CVPR*, 2018, pp. 6546–6555.

[235] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *preprint arXiv:1704.04861*, 2017, pp. 1–9.

[236] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. of CVPR*, 2018, pp. 6848–6856.

[237] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and less than 0.5 MB model size," *preprint arXiv:1602.07360*, 2016, pp. 1–13.

[238] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. of ICML*, 2019, pp. 6105–6114.

[239] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video Action Transformer Network," *Proc. of CVPR*, 2019, pp. 244–253.

[240] Y. Cai, H. Li, G. Yuan, W. Niu, Y. Li, X. Tang, B. Ren, and Y. Wang, "Yolobile: Real-time object detection on mobile devices via compression-compilation co-design," in *Proc. of the AAAI*, vol. 35, no. 2, 2021, pp. 955–963.

[241] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. of ECCV*, 2014, pp. 740–755.

[242] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. of Int. Conf. on Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[243] D. Damen, H. Doughty, G. M. Farinella *et al.*, "The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines," *IEEE Trans. on PAMI*, 2021, vol. 43, no. 11, pp. 4125–4141.

[244] W. Price and D. Damen, "An evaluation of action recognition models on epic-kitchens," *preprint arXiv:1908.00867*, 2019, pp. 1–6.

[245] S. Li, P. Zheng, J. Fan, and L. Wang, "Toward Proactive Human–Robot Collaborative Assembly: A Multimodal Transfer-Learning-Enabled Action Prediction Approach," *IEEE Trans. on Industrial Electronics*, 2022, vol. 69, no. 8, pp. 8579–8588.

[246] L. Yang, X. Shan, C. Lv, J. Brighton, and Y. Zhao, "Learning Spatio-Temporal Representations With a Dual-Stream 3-D Residual Network for Nondriving Activity Recognition," *IEEE Trans. on Industrial Electronics*, 2022, vol. 69, no. 7, pp. 7405–7414.

[247] A. Hietanen, R. Pieters, M. Lanz, J. Latokartano, and J.-K. Kämäräinen, "AR-based interaction for human-robot collaborative manufacturing," *Robotics and Computer-Integrated Manufacturing*, 2020, vol. 63, p. 101891.

[248] N. Hussein, E. Gavves, and A. Smeulders, "Videograph: Recognizing minutes-long human activities in videos," *arXiv:1905.05143*, 2019, pp. 1–10.

[249] Y. Y. Joefrie and M. Aono, "Action Recognition by Composite Deep Learning Architecture I3D-DenseLSTM," in *Proc. of International Conference of Advanced Informatics: Concepts, Theory and Applications*, 2019, pp. 1–6.

[250] D. Cao, L. Xu, and H. Chen, "Action Recognition in Untrimmed Videos with Composite Self-attention Two-Stream Framework," *Pattern Recognition*, 2020, p. 27–40.

[251] J. Zhou, K.-Y. Lin, H. Li, and W.-S. Zheng, "Graph-based high-order relation modeling for long-term action recognition," in *Proc. of CVPR*, 2021, pp. 8984–8993.

[252] F. Liu, X. Xu, T. Zhang, K. Guo, and L. Wang, "Exploring privileged information from simple actions for complex action recognition," *Neurocomputing*, 2020, vol. 380, pp. 236–245.

[253] F. Sener and A. Yao, "Unsupervised learning and segmentation of complex activities from video," in *Proc. of CVPR*, 2018, pp. 8368–8376.

[254] Y. Wang and M. Hoai, "Pulling Actions out of Context: Explicit Separation for Effective Combination," in *Proc. of CVPR*, 2018, pp. 7044–7053.

[255] Y. Wang, V. Tran, G. Bertasius, L. Torresani, and M. Hoai, "Attentive Action and Context Factorization," *arXiv:1904.05410*, 2020, pp. 1–10.

[256] O. Ulutan, S. Rallapalli, M. Srivatsa, C. Torres, and B. Manjunath, "Actor conditioned attention maps for video action detection," in *Proc. of ACV*, 2020, pp. 527–536.

[257] Y. Zhang, J. Li, G. Wu, H. Zhang, Z. Shi, Z. Liu, and Z. Wu, "Temporal transformer networks with self-supervision for action recognition," *preprint arXiv:2112.07338*, 2021, pp. 1–13.

[258] A. Graham, *Kronecker products and matrix calculus with applications.* Courier Dover Publications, 2018.

[259] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[260] L. Rayleigh, "XVI. On James Bernouilli's theorem in probabilities," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1899, vol. 47, no. 285, pp. 246–251.

[261] G. A. Sigurdsson, G. Varol, X. Wang *et al.*, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. of ECCV*, 2016, pp. 510–526.

[262] S. Yeung, O. Russakovsky, N. Jin *et al.*, "Every moment counts: Dense detailed labeling of actions in complex videos," *International Journal of Computer Vision*, 2018, vol. 126, no. 2-4, pp. 375–389.

[263] A. Paszke, S. Gross, S. Chintala *et al.*, "Automatic differentiation in pytorch," *Proc. of NIPS*, 2017, pp. 1–4.

[264] Y.-D. Kim, E. Park, S. Yoo, T. Choi, L. Yang, and D. Shin, "Compression of deep convolutional neural networks for fast and low power mobile applications," *preprint arXiv:1511.06530*, 2015, pp. 1–10.

[265] T. Garipov, D. Podoprikhin, A. Novikov, and D. Vetrov, "Ultimate tensorization: compressing convolutional and fc layers alike," *preprint arXiv:1611.03214*, 2016, pp. 1–10.

[266] W. Wang, V. Aggarwal, and S. Aeron, "Efficient low rank tensor ring completion," in *Proc. of ICCV*, 2017, pp. 5697–5705.

[267] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. of ECCV*, 2018, pp. 803–818.

List of Publications

Refereed Journal Papers (Published/Accepted)

1. N. Nigam, T. Dutta and H.P. Gupta, "FactorNet: Holistic Actor, Object, and Scene Factorization for Action Recognition in Videos," in *IEEE Trans. on Circuits and Systems for Video Technology*, 2022, vol. 32, no. 3, pp. 976-991.
2. N. Nigam and T. Dutta, "Emotion and Gesture Guided Action Recognition in Videos Using Supervised Deep Networks," in *IEEE Trans. on Computational Social Systems*, 2022 (Early Access), doi: 10.1109/TCSS.2022.3187198.
3. N.Nigam, T. Dutta and D. Verma, "Fall-perceived Action Recognition of Persons with Neurological Disorders using Semantic Supervision," in *IEEE Trans. on Cognitive and Developmental Systems* (Early Access), doi: 10.1109/TCDS.2022.3157813.

Refereed Journal Papers (Submitted)

1. N. Nigam and T. Dutta, "24x7 Secure Monitoring of Unsafe Behaviour of MCI Patients via Video-based Action Recognition", in *IEEE Trans. on Cognitive and Developmental Systems*, 2023.
2. N. Nigam and T. Dutta, "Leveraging Smartphones for Order and Span Invariant Long-range Complex Action Recognition", in *IEEE Trans. on Industrial Informatics*, 2023.

Refereed Conference Posters

1. N. Nigam, and T. Dutta, "Poster Abstract: A Fast, Multi-Camera, and Intelligent System for Exact Stampede Detection in Large Crowds", in *Proc. of ACM SenSys*, pp. 1-2, 2022.
2. N. Nigam, and T. Dutta, "Poster Abstract: Crowd Crush Detection in Large Mass Gatherings via Federated Learning Across Multicamera Environment", in *Proc. of ACM BuildSys*, pp. 1-2, 2022.

Other Journal Papers

1. R. Bagi, T. Dutta, N. Nigam, D. Verma, and, H. P. Gupta, "Met-MLTS: Leveraging Smartphones for End-to-end Spotting of Multilingual Oriented Scene Texts and Traffic Signs in Adverse Meteorological Conditions," in *IEEE Trans. on Intelligent Transportation System*, 2021, vol. 23, no. 8, pp. 12801-12810.
2. A. Soni, T. Dutta, N. Nigam, D. Verma, and, H. P. Gupta, "Supervised Attention Network for Arbitrary-shaped Text Detection in Edge-fainted Noisy Scene Images," in *IEEE Trans. on Computational Social Systems*, 2022.

Other Conference Papers

1. N. Nigam, T.Dutta, and H. P. Gupta, "Impact of Noisy Labels in Learning Techniques: A Survey" in *Proc. of Springer Conference on Advances in Data and Information Sciences*, 2020, pp. 403-411.

Patent (Applied for)

1. N. Nigam and T.Dutta, "A SECURED VISION-BASED SMART SURVEILLANCE SYSTEM AND METHOD THEREOF" in *Indian Patent*, Application no. 202311030847.