References

[1] Fernando Camarena, Miguel Gonzalez-Mendoza, Leonardo Chang, and Ricardo Cuevas-Ascencio. An overview of the vision-based human action recognition field. *Mathematical and Computational Applications*, 28(2):61, 2023. 1

[2] Alexandros Andre Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, 34(15):1799–1807, 2013. 1

[3] Saikat Chakraborty, Riktim Mondal, Pawan Kumar Singh, Ram Sarkar, and Debotosh Bhattacharjee. Transfer learning with fine tuning for human action recognition from still images. *Multimedia Tools and Applications*, 80:20547–20578, 2021. 1, 2

[4] Masoumeh chapariniya, Seyed Sajad Ashrafi, and Shahriar B Shokouhi. Knowledge distillation framework for action recognition in still images. In *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 274–277. IEEE, 2020. 2

[5] Masoumeh Chapariniya, Sara Vesali Barazande, Seyed Sajad Ashrafi, and Shahriar B Shokouhi. Attention transfer in self-regulated networks for recognizing human actions from still images. In *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 036–041. IEEE, 2022. 2

[6] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, pages 168–172. IEEE, 2015. 2, 1

[7] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. In *Advances in Neural Information Processing Systems*, pages 3965–3977. Curran Associates, Inc., 2021. 2, 6

[8] Hojat Asgarian Dehkordi, Ali Soltani Nezhad, Hossein Kashiani, Shahriar Baradaran Shokouhi, and Ahmad Ayatollahi. Multi-expert human action recognition with hierarchical super-class learning. *Knowledge-Based Systems*, 250: 109091, 2022. 2

[9] Vincent Delaitre, Ivan Laptev, and Josef Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC 2010-21st British Machine Vision Conference*, 2010. 2

[10] Andrea D'Eusanio, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. Manual annotations on depth maps for human pose estimation. In *Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part I 20*, pages 233–244. Springer, 2019. 1, 2

[11] Guodong Guo and Alice Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10): 3343–3361, 2014. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 2

[13] Samitha Herath, Basura Fernando, and Mehrtash Harandi. Using temporal information for recognizing actions from still images. *Pattern Recognition*, 96:106989, 2019. 2

[14] Seyed Rohollah Hosseyni, Hasan Taheri, Sanaz Seyedin, and Ali Ahmad Rahmani. Human action recognition in still images using convit. *arXiv preprint arXiv:2307.08994*, 2023. 2

[15] Nazli Ikizler, R Gokberk Cinbis, Selen Pehlivan, and Pinar Duygulu. Recognizing actions from still images. In *2008 19th International conference on pattern recognition*, pages 1–4. IEEE, 2008. 2

[16] Aouaidjia Kamel, Bin Sheng, Po Yang, Ping Li, Ruimin Shen, and David Dagan Feng. Deep convolutional neural networks for human action recognition using depth maps and postures. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(9):1806–1819, 2018. 1

[17] Yeonho Kim and Daijin Kim. A cnn-based 3d human pose estimation based on projection of depth and ridge data. *Pattern Recognition*, 106:107462, 2020. 1, 2

[18] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *The International Conference on Learning Representations*, 2021. 2, 6

[19] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022. 1

[20] Pranjal Kumar, Siddhartha Chauhan, and Lalit Kumar Awasthi. Human pose estimation using deep learning: review, methodologies, progress and future research directions. *International Journal of Multimedia Information Retrieval*, 11(4):489–521, 2022. 1

[21] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 6

[22] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 9–14. IEEE, 2010. 2, 1

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6, 2

[24] Manuel J Marin-Jimenez, Francisco J Romero-Ramirez, Rafael Munoz-Salinas, and Rafael Medina-Carnicer. 3d human pose estimation from depth maps using a deep combination of poses. *Journal of Visual Communication and Image Representation*, 55:627–639, 2018. 1, 3

[25] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer*

vision and pattern Recognition, pages 5079–5088, 2018. 1, 3

[26] Md Golam Morshed, Tangina Sultana, Aftab Alam, and Young-Koo Lee. Human action recognition: A taxonomy-based survey, updates, and opportunities. *Sensors*, 23(4): 2182, 2023. 1

[27] Ali Soltani Nezhad, Hojat Asgarian Dehkordi, Seyed Sajad Ashrafi, and Shahriar B Shokouhi. To transfer or not to transfer (tnt):: Action recognition in still image using transfer learning. In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, pages 341–345. IEEE, 2021. 1

[28] Tangquan Qi, Yong Xu, Yuhui Quan, Yaodong Wang, and Haibin Ling. Image-based action recognition using hint-enhanced deep neural networks. *Neurocomputing*, 267:475–488, 2017. 2

[29] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 6, 2

[30] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 3, 4

[31] Fadime Sener, Cagdas Bas, and Nazli Ikizler-Cinbis. On recognizing actions in still images via multiple features. In *Computer Vision–ECCV 2012. Workshops and Demonstrations: Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*, pages 263–272. Springer, 2012. 2

[32] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, et al. Efficient human pose estimation from single depth images. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2821–2840, 2012. 1, 2

[33] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 1

[34] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6, 2

[35] Jinghua Wang and Gang Wang. Hierarchical spatial sum–product networks for action recognition in still images. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(1):90–100, 2016. 2

[36] Jianzhai Wu, Dewen Hu, Fengtao Xiang, Xingsheng Yuan, and Jiongming Su. 3d human pose estimation by depth map. *The Visual Computer*, 36:1401–1410, 2020. 1, 3

[37] L. Xia, C.C. Chen, and JK Aggarwal. View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 20–27. IEEE, 2012. 2, 1

[38] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021. 1, 2

[39] Shiyang Yan, Jeremy S Smith, Wenjin Lu, and Bailing Zhang. Multibranch attention networks for action recognition in still images. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):1116–1125, 2017. 2

[40] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal attention for long-range interactions in vision transformers. In *Advances in Neural Information Processing Systems*, pages 30008–30022. Curran Associates, Inc., 2021. 6, 2

[41] Jianfei Yang, Yuecong Xu, Haozhi Cao, Han Zou, and Lihua Xie. Deep learning and transfer learning for device-free human activity recognition: A survey. *Journal of Automation and Intelligence*, 1(1):100007, 2022. 1

[42] Weilong Yang, Yang Wang, and Greg Mori. Recognizing human actions from still images with latent poses. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2030–2037. IEEE, 2010. 2

[43] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 579–588, 2021. 1, 2

[44] Zhichen Zhao, Huimin Ma, and Shaodi You. Single image action recognition using semantic body part actions. In *Proceedings of the IEEE international conference on computer vision*, pages 3391–3399, 2017. 2

[45] Yunpeng Zheng, Xiangtao Zheng, Xiaoqiang Lu, and Siyuan Wu. Spatial attention based visual semantic learning for action recognition in still images. *Neurocomputing*, 413:383–396, 2020. 2

[46] Haisheng Zhu, Jian-Fang Hu, and Wei-Shi Zheng. Learning hierarchical context for action recognition in still images. In *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part III 19*, pages 67–77. Springer, 2018. 2