# The Journey of Action Recognition

Xi Ding
Australian National University
Canberra, ACT, Australia
Xi.Ding1@anu.edu.au

Lei Wang*
Australian National University
Canberra, ACT, Australia
Lei.W@anu.edu.au

## Abstract

Action recognition has transitioned from a niche area to a cornerstone of video understanding, propelled by advancements in data, model architectures, and learning paradigms. Early efforts relied on limited datasets and handcrafted features, but the advent of deep learning sparked a revolution. Models like 2D- and 3D-CNNs, spatiotemporal graph networks, and transformers have significantly improved the ability to capture complex, multidimensional actions across diverse, multimodal datasets. Innovative paradigms, including self-supervised, few-shot, and zero-shot learning, have redefined data usage, enabling models to generalize with minimal labeled data. Transformer architectures now excel at capturing long-range temporal dependencies, while video masked autoencoders have advanced the balance of spatial and temporal information, driving breakthroughs in motion dynamics. This paper examines action recognition through three lenses: evolving architectures, diverse data, and novel learning techniques. We trace how these elements have expanded the field's scope to address challenges like anomaly detection, captioning, and video question answering. Notably, large language models have infused semantic context, enhancing both performance and versatility. Reflecting on these developments, we provide a roadmap for future progress. By integrating multimodal, temporal, and semantic insights, action recognition is poised to become central to AI-driven video analysis, paving the way for transformative advancements in understanding and interaction.

## CCS Concepts

• **Computing methodologies** → *Computer vision representations*; **Learning paradigms**; **Machine learning algorithms**; **Activity recognition and understanding**; • **Networks** → **Network architectures**.

## Keywords

Action recognition, Data, Model architectures, Learning paradigm

## 1 Introduction

Action recognition, the task of identifying and understanding human actions in video, has become a pivotal area of research in computer vision and machine learning [85]. It plays a crucial role in a wide range of applications, from surveillance systems and autonomous driving to video indexing and human-computer interaction. Early research works in action recognition focus on small, labeled video datasets and rely heavily on handcrafted features to capture motion and spatial information [9, 48, 117, 199, 202, 250]. However, with the rapid growth of video data and advancements in machine learning techniques, the field has undergone significant transformations, leading to more robust, scalable, and accurate methods [13, 21, 234, 250, 264].

The evolution of action recognition can be understood through three key interconnected dimensions: the data, the learning paradigms, and the model architectures. As datasets grow in scale and complexity, researchers begin to shift from simple, labeled datasets to more diverse and larger video repositories. This shift enables the development of learned representations through deep learning techniques, which outperform traditional methods that rely on handcrafted features [32, 49, 101, 223, 311]. Concurrently, new learning paradigms, such as unsupervised, self-supervised, and few-shot learning, are introduced to better use the expanding volume of unlabeled video data [41, 50, 76, 81]. These paradigms enable models to generalize more effectively, making it possible to learn action recognition tasks without the need for large amounts of manually labeled data.

In parallel, model architectures evolve from simple 2D convolutional networks (CNNs) to more complex 3D and two-stream networks designed to capture spatial and temporal features [33, 183, 215, 231]. Recent advancements in transformer-based models and video masked autoencoders have further pushed the boundaries of action recognition, allowing for better handling of long-range temporal dependencies and improving the capture of both spatial and temporal motion features [70, 230, 251, 277]. The integration of language models and vision-language models into action recognition has further enhanced the field, enabling richer contextual understanding of actions and their relationships to textual descriptions [106, 161, 272].

This paper aims to provide a comprehensive exploration of the journey of action recognition from its early stages to its current state. We discuss the evolution of the field from both data and model perspectives and examine how learning paradigms have shaped the progress of action recognition research. Through this analysis, we uncover valuable insights into the challenges, breakthroughs, and future directions of action recognition, as it continues to advance and become an integral part of broader video processing tasks. The main **contributions** of this paper are as follows:

i. A detailed review of the evolution of action recognition from data, learning, and model perspectives, highlighting key milestones and breakthroughs.
ii. An in-depth exploration of the co-evolution of paradigms, data, and architectures, offering a unified view of the interdependencies that have shaped the field.
iii. A discussion on the future directions and emerging trends in action recognition, emphasizing the integration of multimodal data, transformer-based architectures, and vision-language models, and their potential to address the challenges in video understanding and processing.

*Corresponding author.

**Table 1: The journey of action recognition (Part 1): Methods based on RGB videos, including handcrafted features, 2D CNNs, (2+1)D CNNs, 3D CNNs, two-stream networks, and transformers. Columns detail learning paradigms, data modalities, and publication venues (year).**

| | Method | Venue | Learning | Dataset | Modality |
|---|---|---|---|---|---|
| Handcrafted | HL-STIP[117] | IJCV 2005 | Supervised | Outdoor scenes [117] | RGB |
| | Spatio-temporal Cuboids[48] | VS-PETS 2005 | Supervised | Human Action Dataset[201] | RGB |
| | 3D-SURF[9] | ECCV 2006 | Supervised | Mikolajczyk[163] | RGB |
| | 3D-SIFT[202] | ACM MM 2007 | Supervised | Weizmann[74] | RGB |
| | NNMF Detector [283] | ICCV 2007 | Supervised | KTH[201] | RGB |
| | HOG3D[107] | BMVC 2008 | Supervised | KTH[201], Weizmann[74], Hollywood[118] | RGB |
| | Laptev et al.[118] | CVPR 2008 | Supervised | KTH[201] | RGB+Optical flow |
| | Action MACH[192] | CVPR 2008 | Supervised | KTH[201], Weizmann[74] | RGB+Optical flow |
| | extended SURF[281] | ECCV 2008 | Supervised | KTH[201], TRECVID 2006[179] | RGB |
| | LTP[308] | ICCV 2009 | Supervised | KTH[201], Hollywood[118], Kissing and slapping dataset[192], UCF Sports[192] | RGB |
| | Messing et al.[159] | ICCV 2009 | Supervised | KTH[201] | RGB |
| | Bregonzio et al.[16] | CVPR 2009 | Supervised | KTH[201], Weizmann[74] | RGB |
| | Tracklet Descriptors [190] | ECCV 2010 | Supervised | KTH[201], ADL[159], Hollywood[118] | RGB+Optical flow |
| | Dense Long-Duration Trajectories[224] | ICME 2010 | Supervised | KTH[201] | RGB+Optical flow |
| | Dense Trajectories[240] | IJCV 2013 | Supervised | KTH[201], YouTube[141], Hollywood2[155], UCF Sports[192], IXMAS[279], Olympic Sports[171], UCF50[191], UIUC[232], HMDB51[114] | RGB+Optical flow |
| | iDT[242] | ICCV 2013 | Supervised | Hollywood2[155], HMDB51[114], Olympic Sports[171], UCF50[191] | RGB+Optical flow |
| | Taylor videos [265] | ICML 2024 | Supervised | HMDB51[114],CATER[71],MPII Cooking[193], Kinetics-400[103], -600[19], Something-Something V2[77],NTU RGB+D[142, 205], Kinetics-skeleton[294] | RGB+Skeleton |
| 2D-based | Slow fusion[101] | CVPR 2014 | Supervised | Sports-1M[101], UCF101[221] | RGB |
| | CNN-LSTM[311] | CVPR 2015 | Supervised | Sports-1M[101], UCF101[221] | RGB+Optical flow |
| | LRCN[49] | CVPR 2015 | Supervised | UCF101[221] | RGB+Optical flow |
| | Composite LSTM[223] | ICML 2015 | Unsupervised | UCF101[221], HMDB51[114] | RGB |
| | Rank Pooling[63] | TPAMI 2016 | Supervised | HMDB51[114], Hollywood2[155], MPII Cooking[193] | RGB+Optical flow |
| | LENN[67] | CVPR 2016 | Supervised | UCF101[221] | RGB |
| | Bilen et al.[14] | TPAMI 2017 | Supervised | UCF101[221], HMDB51[114] | RGB |
| | TSN[264] | TPAMI 2018 | Supervised | HMDB51[114], UCF101[221], Kinetics-400[103], ActivityNet[18], THUMOS14[91] | RGB+RGB differences+Optical flow+Warped optical flow+Audio |
| | Attention-LSTM[152] | CVPR 2018 | Supervised | UCF101[221], HMDB51[114], Kinetics-400[103] | RGB+Optical flow+Audio |
| | PEAR[287] | ICME 2019 | Reinforcement | UCF101[221], Sports-1M[101] | RGB+Optical flow |
| | TSM[134] | ICCV 2019 | Supervised | Something-Something V1[77], Something-Something V2[77], Kinetics-400[103], UCF101[221], HMDB51[114] | RGB |
| | VINCE[73] | arXiv 2020 | Self-supervised | Kinetics-400[103] | RGB |
| | $C^2$LSTM[154] | Neurocomputing 2020 | Supervised | UCF101[221], HMDB51[114] | RGB |
| | MoCo[61] | CVPR 2021 | Self-supervised | Kinetics-400[103], UCF101[221], HMDB51[114] | RGB |
| | TCL[216] | CVPR 2021 | Semi-supervised+Contrastive | Mini-Something-V2[23], Kinetics-400[103], Charades-Ego[213] | RGB |
| | TDN[262] | CVPR 2021 | Supervised | Something-Something V1[77], Something-Something V2[77], Kinetics-400[103] | RGB |
| | DB-LSTM[83] | Neurocomputing 2021 | Supervised | UCF101[221], HMDB51[114] | RGB+Optical flow |
| | SeCo[307] | AAAI 2021 | Self-supervised | Kinetics-400[103], UCF101[221], HMDB51[114], ActivityNet[18] | RGB |
| | Xiao et al.[285] | CVPR 2022 | Semi-supervised+Contrastive | Kinetics-400[103], UCF101[221], HMDB51[114] | RGB |
| | GCSM[309] | ACM MM 2023 | Few-shot | UCF101[221], HMDB51[114], Kinetics-400[103] | RGB |
| | GgHM[289] | ICCV 2023 | Few-shot | HMDB51[114], UCF101[221], Kinetics-400[103], Something-Something V2[77] | RGB |
| 3D-based | C3D[231] | ICCV 2015 | Supervised | UCF101[221] | RGB |
| | I3D[21] | CVPR 2017 | Supervised | Kinetics-400[103], UCF101[221], HMDB51[114] | RGB |
| | P3D[183] | ICCV 2017 | Supervised | Sports-1M[101], UCF101[221], ActivityNet[18] | RGB |
| | ResNet3D[82] | CVPR 2018 | Supervised | Kinetics-400[103], UCF101[221], HMDB51[114], ActivityNet[18] | RGB |
| | S3D[288] | ECCV 2018 | Supervised | Kinetics-400[103], Something-Something V1[77], UCF101[221], HMDB51[114] | RGB+Optical flow |
| | CSN[233] | ICCV 2019 | Supervised | Sports-1M[101], Kinetics-400[103], Something-Something V1[77] | RGB |
| | SlowFast[60] | ICCV 2019 | Supervised | Kinetics-400[103], Kinetics-600[19], Charades[214], AVA[79] | RGB |
| | STM[94] | ICCV 2019 | Supervised | Something-Something V1[77], Something-Something V2[77], Kinetics-400[103], UCF101[221], HMDB51[114] | RGB |
| | DEEP-HAL [258] | ICCV 2019 | Self-supervised | HMDB51[114], Charades[214],MPII Cooking[193] | RGB+Optical flow |
| | Xv et al.[291] | CVPR 2019 | Self-supervised | UCF101[221], HMDB51[114] | RGB |
| | X3D[58] | CVPR 2020 | Supervised | Kinetics-400[103], Kinetics-600[19], Charades[214], AVA[79] | RGB |
| | TPN[297] | CVPR 2020 | Supervised | Kinetics-400[103], Something-Something V1[77], Something-Something V2[77], Epic-Kitchens[38] | RGB |
| | SpeedNet[12] | CVPR 2020 | Self-supervised | Kinetics-400[103], UCF101[221], HMDB51[114], NfS[66] | RGB |
| | CoCLR[81] | NeurIPS 2020 | Self-supervised | UCF101[221], HMDB51[114], Kinetics-400[103] | RGB+Optical flow |
| | VTHCL[296] | arXiv 2020 | Self-supervised | Kinetics-400[103], UCF101[221], HMDB51[114] | RGB |
| | Multi-Transforms[237] | arXiv 2021 | Self-supervised | UCF101[221], HMDB51[114] | RGB |
| | MvPL[290] | ICCV 2021 | Semi-supervised | Kinetics-400[103], UCF101[221], HMDB51[114] | RGB+Optical flow |
| | CVRL[181] | CVPR 2021 | Self-supervised | Kinetics-400[103], Kinetics-600[19], UCF101[221], HMDB51[114] | RGB |
| | Yang et al.[301] | CVPR 2021 | Supervised | Kinetics-400[103], Kinetics-700[20], Charades[214], Something-Something V1[77], AVA[79] | RGB |
| | 3DResNet+ATFR[57] | CVPR 2021 | Supervised | Kinetics-400[103], Kinetics-600[19], UCF101[221], HMDB51[114], Something-Something V2[77] | RGB |
| | MoViNet[108] | CVPR 2021 | Supervised | Kinetics-400[103], Kinetics-600[19], Kinetics-700[20], Something-Something V2[77], Epic-Kitchens-100[39], MiT[165], Charades[214] | RGB |
| | ODF+SDF [253] | ACM MM 2021 | Self-supervised | HMDB51[114], Charades[214],MPII Cooking[193], EPIC-Kitchen[38] | RGB+Optical flow + object / saliency detectors |
| | CLASTER [76] | ECCV 2022 | Reinforcement+Zero-shot | UCF101[221], HMDB51[114], Olympic Sports[171] | RGB+Optical flow+Semantic embeddings |
| | TFCNet[315] | arXiv 2022 | Supervised | Diving48[133], CATER[71] | RGB |
| | HoT [261] | ICASSP 2024 | Supervised | HMDB51[114],MPII Cooking[193] | RGB+Optical flow |
| | Flow corr. [257] | ICASSP 2024 | Supervised | HMDB51[114], Charades[214],MPII Cooking[193] | RGB+Optical flow |
| Two-stream | Two-Stream ConvNet[215] | NeurIPS 2014 | Supervised | UCF101[221], HMDB51[114] | RGB+Optical flow |
| | P-CNN[33] | ICCV 2015 | Supervised | JHMDB[93], MPII Cooking[193] | RGB+Optical Flow+Joint |
| | TDD[260] | CVPR 2015 | Supervised | HMDB51[114], UCF101[221] | RGB+Optical flow |
| | Two-Stream Fusion[62] | CVPR 2016 | Supervised | UCF101[221], HMDB51[114] | RGB+Optical flow |
| | TSN-Two-Stream[263] | ECCV 2016 | Supervised | HMDB51[114], UCF101[221] | RGB+RGB differences+Optical flow+Warped optical flow |
| | DOVF[116] | CVPR 2017 | Supervised | UCF101[221], HMDB51[114] | RGB+Optical flow |
| | TLE[44] | CVPR 2017 | Supervised | UCF101[221], HMDB51[114] | RGB+Optical flow |
| | ActionVLAD[72] | CVPR 2017 | Supervised | HMDB51[114], UCF101[221], Charades[214] | RGB+Optical flow |
| | TRN-Two-Stream[320] | ECCV 2018 | Supervised | Something-Something V1[77], Something-Something V2[77], Charades[214] | RGB |
| | TSM-Two-Stream[134] | ICCV 2019 | Supervised | Something-Something V1[77], Something-Something V2[77], Kinetics-400[103], UCF101[221], HMDB51[114] | RGB+Optical flow |
| | KTSN[146] | arXiv 2020 | Supervised | FSD-10[146] | RGB+Optical flow+Skeleton |
| | MSM-ResNets[325] | IVC 2021 | Supervised | UCF101[221], HMDB51[114] | RGB+Optical Flow+Motion Saliency |
| | MAT-EffNet[319] | MMSys 2023 | Supervised | UCF101[221], HMDB51[114], Kinetics-400[103] | RGB+Optical flow |
| | TTFA[41] | SPL 2024 | Few-shot | Something-Something V2[77], Kinetics-400[103] | RGB+Optical flow |
| (2+1)D-based | R(2+1)D[234] | CVPR 2018 | Supervised | Kinetics-400[103], Sports-1M[101], UCF101[221], HMDB51[114] | RGB+Optical flow |
| | R(2+1)D+BERT[99] | ECCVW 2020 | Supervised | HMDB51[114], UCF101[221] | RGB |
| | XDC[6] | NeurIPS 2020 | Self-supervised | HMDB51[114], UCF101[221] | RGB+Audio |
| | ELo[180] | CVPR 2020 | Self-supervised | Kinetics-400[103], UCF101[221], HMDB51[114] | RGB+Optical flow+Audio |
| | Jin et al.[97] | ICICSP 2021 | Supervised | UCF101[221] | RGB |
| | GDT[176] | arXiv 2021 | Self-supervised | Kinetics-400[103], UCF101[221], HMDB51[114] | RGB+Audio |
| | AVID[166] | CVPR 2021 | Self-supervised | Kinetics-400[103], UCF101[221], HMDB51[114] | RGB+Audio |
| Transformer-based | VTN[170] | ICCV 2021 | Supervised | Kinetics-400[103], MiT[165] | RGB |
| | TimeSformer[13] | ICML 2021 | Supervised | Kinetics-400[103], Kinetics-600[19] | RGB |
| | STAM[207] | arXiv 2021 | Supervised | Kinetics-400[103], UCF101[221], Charades[214] | RGB |
| | ViViT[7] | ICCV 2021 | Supervised | Kinetics-400[103], Kinetics-600[19], Epic-Kitchens-100[39], MiT[165], Something-Something V2[77] | RGB |
| | MViT[54] | ICCV 2021 | Supervised | Kinetics-400[103], Kinetics-600[19], Something-Something V2[77], Charades[214], AVA[79] | RGB |
| | Motionformer[177] | NeurIPS 2021 | Supervised | Kinetics-400[103], Kinetics-600[19], Something-Something V2[77], Epic-Kitchens-100[39] | RGB |
| | X-ViT[17] | NeurIPS 2021 | Supervised | Kinetics-400[103], Kinetics-600[19], Something-Something V2[77], Epic-Kitchens-100[39] | RGB |
| | TallFormer[29] | ECCV 2022 | Supervised | THUMOS14[91], ActivityNet[18] | RGB |
| | VideoSwin[150] | CVPR 2022 | Supervised | Kinetics-400[103], Kinetics-600[19], Something-Something V2[77] | RGB |
| | ORViT[86] | CVPR 2022 | Supervised | Something-Something V2[77], SomethingElse[156], Diving48[133], AVA[79], Epic-Kitchens-100[39] | RGB |
| | BEVT[269] | CVPR 2022 | Self-supervised | Kinetics-400[103], Something-Something V2[77], Diving-48[133] | RGB |
| | MaskFeat[277] | CVPR 2022 | Self-supervised | Kinetics-400[103], Kinetics-600[19], Kinetics-700[20] | RGB |
| | UniFormer[125] | arXiv 2022 | Supervised | Kinetics-400[103], Kinetics-600[19], Something-Something V1[77], Something-Something V2[77] | RGB |
| | VideoMAE[230] | NeurIPS 2022 | Self-supervised | Kinetics-400[103], Something-Something V2[77], UCF101[221], HMDB51[114], AVA[79] | RGB |
| | MTV[293] | CVPR 2022 | Supervised | Kinetics-400[103], Kinetics-600[19], Kinetics-700[20], Something-Something V2[77], Epic-Kitchens-100[39], MiT[165] | RGB |
| | MAE-ST[59] | arXiv 2022 | Self-supervised | Kinetics-400[103], Something-Something V2[77], AVA[79] | RGB |
| | CAST[120] | NeurIPS 2023 | Supervised | Kinetics-400[103], Something-Something V2[77], Epic-Kitchens-100[39] | RGB |
| | UniFormerV2[126] | ICCV 2023 | Supervised+Contrastive | Kinetics-400[103], Kinetics-600[19], Kinetics-700[20], MiT[165], Something-Something V1[77], Something-Something V2[77], ActivityNet[18], HACS[318] | RGB |
| | OmniMAE[70] | CVPR 2023 | Self-supervised | Something-Something V2[77], Epic-Kitchens-100[39], Kinetics-400[103] | RGB |
| | MVD[270] | CVPR 2023 | Self-supervised | Kinetics-400[103], Something-Something V2[77], UCF101[221], HMDB51[114] | RGB |
| | Hiera[196] | ICML 2023 | Self-supervised | Kinetics-400[103], Kinetics-600[19], Kinetics-700[20], Something-Something V2[77], AVA[79] | RGB |
| | VideoMAE V2[251] | CVPR 2023 | Self-supervised | Kinetics-400[103], Something-Something V2[77], UCF101[221], HMDB51[114] | RGB |
| | SOAP[88] | ACM MM 2024 | Few-shot | Something-Something V2[77], Kinetics-400[103], UCF101[221], HMDB51[114] | RGB |
| | C2C[129] | ECCV 2024 | Zero-shot | Sth-com[129] | RGB |
| | VMPs [25] | ACML 2024 | Supervised | HMDB51[114],MPII Cooking 2[194], FineGym [206] | RGB+Motion prompts |
| | TIME Layer [24] | arXiv 2024 | Self-supervised | UCF101[221], HMDB51[114], UWA3D Multiview Activity II[186], NTU RGB+D[205], NTU RGB+D 120[142] | RGB+Depth |

## 2 Related Work

Action recognition has garnered substantial attention over the past few decades, resulting in numerous survey papers that examine the evolution of methods, datasets, and models [53, 109, 225, 249, 252]. These surveys provide valuable insights into the historical progression of action recognition, categorize various approaches, and identify the challenges and future directions. However, despite the wealth of surveys, each focuses on different aspects of the problem, and most either emphasize specific models, learning paradigms,

**Table 2: The journey of action recognition (Part 2): Methods using alternative modalities, including skeleton-based, depth-based, infrared-based, point cloud-based, and multi-modal approaches (*e.g.*, text or audio). Columns detail learning paradigms, data modalities, and publication venues (year).**

| | Method | Venue | Learning | Dataset | Modality |
|---|---|---|---|---|---|
| | Dynamic Skeletons [87] | CVPR 2015 | Supervised | MSRDailyActivity[247], CAD-60[227], SYSU 3D HOI[87] | Depth+Joint |
| | HBRNN-L [52] | CVPR 2015 | Supervised | MSRAction3D[132], Berkeley MHAD[173], HDM05[168] | Joint |
| | Part-aware LSTM[205] | CVPR 2016 | Supervised | NTU RGB+D[205] | RGB+Depth+Joint+Infrared |
| | LARP-SO[236] | CVPR 2016 | Supervised | Florence3D-Action[203], MSRActionPairs3D[174], G3D-Gaming[15] | Joint |
| | STA-LSTM [218] | AAAI 2017 | Supervised | NTU RGB+D[205] | Joint |
| | LieNet [90] | CVPR 2017 | Supervised | NTU RGB+D[205], HDM05[168], G3D-Gaming[15] | Joint+Bone |
| | Two-Stream RNN [243] | CVPR 2017 | Supervised | NTU RGB+D[205] | Joint |
| | Ke et al. [104] | CVPR 2017 | Supervised | NTU RGB+D[205] | Joint |
| | VA-LSTM [314] | ICCV 2017 | Supervised | NTU RGB+D[205], SYSU 3D HOI[87] | Joint |
| | View Invariant[145] | Pattern Recognit. 2017 | Supervised | NTU RGB+D[205], Northwestern-UCLA[248], UWA3D Multiview Activity II[186], MSRC-12[64] | Joint |
| | Two-Stream CNN[123] | ICMEW 2017 | Supervised | NTU RGB+D[205], PKU-MMD I[137] | Joint+Skeleton motion |
| | LSTM-CNN[122] | ICMEW 2017 | Supervised | NTU RGB+D[205] | Joint |
| | ST-LSTM+Trust Gate [143] | TPAMI 2018 | Supervised | NTU RGB+D[205], MSRAction3D[132], SYSU 3D HOI[87], Berkeley MHAD[173] | Joint |
| | ST-GCN[294] | AAAI 2018 | Supervised | Kinetics-400[103], NTU RGB+D[205] | Joint |
| | Tang et al. [229] | CVPR 2018 | Reinforcement | NTU RGB+D[205], SYSU 3D HOI[87], UTKinect-Action3D[284] | Joint+Bone |
| | AS-GCN [128] | CVPR 2019 | Supervised | NTU RGB+D[205], Kinetics-400[103] | Joint+Bone |
| | 2s-AGCN[211] | CVPR 2019 | Fully-supervised | NTU RGB+D[205], Kinetics-skeleton[294] | Joint+Bone |
| | DGNN [210] | CVPR 2019 | Supervised | NTU RGB+D[205], Kinetics-skeleton[294] | Joint+Bone |
| | EfficientGCN[219] | ACM MM 2020 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142] | Joint+Velocity+Bone |
| | RA-GCN [220] | TCSVT 2020 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142] | Joint+Bone |
| | Shift-GCN [30] | CVPR 2020 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142], Northwestern-UCLA[248] | Joint+Bone |
| | MS-G3D [151] | CVPR 2020 | Supervised | NTU RGB+D 60[205], NTU RGB+D 120[142], Kinetics-skeleton[294] | Joint+Bone |
| | DSTA-Net [212] | ACCV 2020 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142] | Joint+Bone |
| | SCK+DCK / SCK$\oplus$+DCK$\oplus$ [110] | TPAMI 2020 | Supervised | UTKinect-Action3D[284], Florence3D-Action[203],MSRAction3D[132],NTU RGB+D 60[205], Kinetics-400[103], HMDB51[114], MPII Cooking[193] | Joint |
| | CTR-GCN[27] | ICCV 2021 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142], Northwestern-UCLA[248] | Joint+Bone |
| Skeleton-based | FGCN [300] | TIP 2022 | Supervised | NTU RGB+D[205], NTU RGB+D120[142], Northwestern-UCLA[248] | Joint+Bone |
| | AGE-Ens [182] | TNNLS 2022 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142] | Joint+Bone |
| | PoseConv3D[53] | CVPR 2022 | Supervised | Kinetics-400[103], UCF101[221], HMDB51[114] | Joint+Bone+RGB |
| | InfoGCN [34] | CVPR 2022 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142], Northwestern-UCLA[248] | Joint+Bone |
| | DASTM[153] | ECCV 2022 | Few-shot | NTU RGB+D 120[142], Kinetics-skeleton[294] | Joint+Bone |
| | Uncertainty-DTW [255] | ECCV 2022 | Supervised/Unsupervised few-shot | NTU RGB+D[205], NTU RGB+D 120[142], Kinetics-skeleton[294] | Skeleton sequences |
| | TranSkeleton [139] | TCSVT 2023 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142] | Joint+Bone |
| | HiCo [50] | AAAI 2023 | Unsupervised+Contrastive | NTU RGB+D[205], NTU RGB+D 120[142], PKU-MMD I[144], PKU MMD II[144] | Joint |
| | FR-Head [321] | CVPR 2023 | Supervised+Contrastive | NTU RGB+D[205], NTU RGB+D 120[142], Northwestern-UCLA[248] | Joint+Bone |
| | 3Mformer [256] | CVPR 2023 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142], Kinetics-400[103], Northwestern-UCLA[248] | Joint + Hyper-edge |
| | HYSP [65] | ICLR 2023 | Self-supervised | NTU RGB+D[205], NTU RGB+D 120[142], PKU-MMD I[144] | Joint |
| | PAINet[148] | ICCV 2023 | Few-shot | NTU RGB+D 120[142], Kinetics-skeleton[294] | Joint+Bone |
| | PCM$^3$ [313] | ACM MM 2023 | Self-supervised | NTU RGB+D[205], NTU RGB+D 120[142], PKU-MMD I[144] | Joint+Bone+Motion |
| | Stream-GCN [303] | arXiv 2023 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142], Northwestern-UCLA[248] | Joint+Bone |
| | SkeletonGCL [89] | arXiv 2023 | Self-supervised | NTU RGB+D[205], NTU RGB+D 120[142], Northwestern-UCLA[248] | Joint+Bone |
| | DSCNet [31] | ESWA 2024 | Supervised+Multimodal | NTU RGB+D[205], NTU RGB+D 120[142], PKU-MMD I[144], UAV-Human[127], IKEA ASM[11], Northwestern-UCLA[248] | RGB+Joint+Bone |
| | Skeleton-OOD [292] | Neurocomputing 2024 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142], Kinetics-400[103] | Joint |
| | ViA [299] | IJCV 2024 | Self-supervised | Posetics[298], NTU RGB+D[205], NTU RGB+D 120[142], Toyota Smarthome[40], UAV-Human[127], Penn Action[316] | Joint+Motion |
| | DeGCN [169] | TIP 2024 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142], Northwestern-UCLA[248] | Joint+Bone |
| | Js-SaPR-GCN[121] | TCSVT 2024 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142], Northwestern-UCLA[248] | Joint+Bone+Motion |
| | BlockGCN [322] | CVPR 2024 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142], Northwestern-UCLA[248] | Joint+Bone+Motion |
| | JEANIE | IJCV 2024 | Supervised/Unsupervised few-shot | NTU RGB+D[205], NTU RGB+D 120[142], Kinetics-skeleton[294], MSRAction3D[132], UWA3D Multiview Activity[187] | Skeleton sequences |
| | SA-DVAE[130] | arXiv 2024 | Zero-shot | NTU RGB+D[205], NTU RGB+D 120[142], PKU-MMD[144] | Joint |
| | ProtoGCN [140] | ArXiv 2024 | Self-supervised+Prototype | NTU RGB+D[205], NTU RGB+D 120[142], Kinetics-skeleton[294], FineGYM[206] | Joint |
| | HSIC-based[302] | arXiv 2024 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142], Northwestern-UCLA[248] | Joint+Bone |
| | USDRL[280] | AAAI 2025 | Self-supervised | NTU RGB+D[205], NTU RGB+D 120[142], PKU-MMD I[144], PKU-MMD II[144] | Joint+Bone+Motion |
| | HON4D[174] | CVPR 2013 | Supervised | MSRAction3D[132], MSRDailyActivity3D[246], MSRActionPairs3D[174] | Depth |
| | HOPC[187] | ECCV 2014 | Supervised | MSRAction3D[132], MSRActionPairs3D[174], UWA3D Multiview Activity[187] | Depth+point cloud |
| | Wang et al.[267] | Trans. Human-Mach. Syst. 2016 | Supervised | MSRAction3D[132], MSRDailyActivity3D[246], UTKinect-Action3D[284] | Depth |
| | Rahmani et al.[188] | CVPR 2016 | Supervised | Northwestern-UCLA[248], UWA3D Multiview Activity II[186] | Depth |
| | S$^2$DDI[268] | ICCVW 2017 | Supervised | MSRAction3D[132], G3D-Gaming[15], MSRDailyActivity3D[246], SYSU 3D HOI[87], UTD-MHAD[22] | Depth |
| Depth-based | Wang et al.[266] | TMM 2018 | Supervised | NTU RGB+D[205] | Depth |
| | MVDI[286] | Inf. Sci. 2018 | Supervised | NTU RGB+D[205], Northwestern-UCLA[248], UWA3D Multiview Activity II[186] | Depth |
| | 3DFCNN[197] | Multimed. Tools Appl. 2020 | Supervised | NTU RGB+D[205], Northwestern-UCLA[248], UWA3D Multiview Activity II[186] | Depth |
| | Liu et al.[138] | ICASSP 2017 | Supervised | MSRAction3D[132], DHA[136] | Depth |
| | Dhiman et al.[43] | TIP 2020 | Supervised | NTU RGB-D[205], UWA3D Multiview Activity II[186], Northwestern-UCLA[248] | RGB+Depth |
| | Stateful ConvLSTM[198] | arXiv 2020 | Supervised | NTU RGB+D[205] | Depth |
| | DEAR[189] | arXiv 2024 | Supervised | Something-Something V2[77] | RGB+Depth |
| | Gao et al.[68] | Neurocomputing 2016 | Supervised | InfAR[68] | Infrared+Optical flow |
| | Jiang et al.[96] | CVPRW 2017 | Supervised | InfAR[68] | Infrared+Optical flow |
| | Kawashima et al.[102] | AVSS 2017 | Supervised | Custom Dataset[102] | Infrared |
| | Shah et al.[204] | SPIE 2018 | Supervised | Custom IR Dataset[204] | Infrared |
| Infrared-based | TSTDDs[149] | SPL 2018 | Supervised | InfAR[68], NTU RGB+D[205] | Infrared+Optical flow |
| | Akula et al.[3] | CSR 2018 | Supervised | Custom IR Dataset[3] | Infrared |
| | Imran et al.[92] | Infrared Phys. Technol. 2019 | Supervised | InfAR[68], IITR-IAR[92] | Infrared+Optical flow |
| | Meglouli et al.[157] | CEAI 2019 | Supervised | InfAR[68] | Infrared+Optical flow |
| | Mehta et al.[158] | ICPR 2020 | Adversarial | TSF[235] | Infrared+Optical flow |
| | MeteorNet[147] | ICCV 2019 | Supervised | MSRAction3D[132] | Point cloud |
| | PointLSTM[164] | CVPR 2020 | Supervised | MSRAction3D[132] | Point cloud |
| | 3DV-PointNet++[274] | CVPR 2020 | Supervised | NTU RGB+D[205], NTU RGB+D 120[142], Northwestern-UCLA[248], UWA3D Multiview Activity II[186] | Depth |
| | ASTA3DConv[239] | Trans. Instrum. Meas. 2020 | Supervised | MSRAction3D[132] | Point cloud |
| | Wang et al.[244] | WACV 2021 | Self-supervised | NTU RGB+D[205], NTU-PCL[205], MSRAction3D[132] | Point cloud |
| Point cloud | P4Transformer[55] | CVPR 2021 | Supervised | MSRAction3D[132], NTU RGB+D[205], NTU RGB+D 120[142] | Point cloud |
| | PSTNet[56] | arXiv 2021 | Supervised | MSRAction3D[132], NTU RGB+D[205], NTU RGB+D 120[142] | Point cloud |
| | PST$^2$[278] | WACV 2022 | Supervised | MSRAction3D[132] | Point cloud |
| | MaST-Pre[208] | ICCV 2023 | Self-supervised | MSRAction3D[132], NTU RGB+D[205] | Point cloud |
| | PointCPSC[209] | ICCV 2023 | Self-supervised | MSRAction3D[132], NTU RGB+D[205] | Point cloud |
| | 3DInAction[10] | CVPR 2024 | Supervised | MSRAction3D[132] | Point cloud |
| | KAN-HyperpointNet[28] | arXiv 2024 | Supervised | NTU RGB+D[205], MSRAction3D[132] | Point cloud |
| | CPD[131] | arXiv 2020 | Self-supervised | Kinetics-400[103], HMDB51[114], UCF101[221] | RGB+Text |
| | G-Blend[271] | CVPR 2020 | Multi-task | Kinetics-400[103], Mini-Sports[101], EPIC-Kitchen[38] | RGB+Optical flow+Audio |
| | MIL-NCE [161] | CVPR 2020 | Self-supervised | HowTo100M[162], HMDB51[114], UCF101[221] | RGB+Text |
| | MMV[5] | NeurIPS 2020 | Self-supervised | UCF101[221], HMDB51[114], Kinetics-600[19] | RGB+Audio+Text |
| | VIMPAC[228] | arXiv 2021 | Self-supervised | Something-Something V2[77], Diving48[133], UCF101[221], HMDB51[114] | RGB+Text |
| | InternVideo[273] | CVPR 2023 | Self-supervised | Kinetics-400[103], Kinetics-600[19], Kinetics-700[20], Something-Something V1[77], Something-Something V2[77], ActivityNet[18], HACS[318], HMDB51[114] | RGB+Text |
| Text / Audio | Side4Video[305] | arXiv 2023 | Self-supervised | Something-Something V1[77], Something-Something V2[77], Kinetics-400[103] | RGB+Text |
| | EZ-CLIP[2] | arXiv 2024 | Zero-shot | Kinetics-400[103], HMDB51[114], UCF101[221], Something-Something V2[77] | RGB+Text |
| | SATA[135] | arXiv 2024 | Zero-shot | UCF101[221], HMDB51[114] | RGB+Text |
| | TC-CLIP[106] | ECCV 2024 | Zero-shot/Few-shot/Fully-supervised | HMDB51[114], UCF101[221], Kinetics-400[103], Something-Something V2[77] | RGB+Text |
| | InternVideo2[272] | arXiv 2024 | Self-supervised+Multimodal | Kinetics-400[103], Kinetics-600[19], Kinetics-700[20], MiT[165], Something-Something V2[77], ActivityNet[18], HACS[318], Charades[214], HMDB51[114] | RGB+Audio+Text |
| | OmniViD[245] | CVPR 2024 | Supervised | Kinetics-400[103], Something-Something V2[77], UCF101[221], HMDB51[114] | RGB+Text |
| | LoCATe-GAT[200] | TETCI 2024 | Zero-shot | UCF101[221], HMDB51[114], ActivityNet[18], Kinetics-400[103] | RGB+Text |
| | STDD[310] | arXiv 2024 | Zero-shot | Kinetics-600[19], UCF101[221], HMDB51[114] | RGB+Text |

or datasets, without providing a comprehensive analysis of the interplay between data, learning methods, and models.

Early surveys on action recognition primarily focus on the progress of handcrafted features. These papers, including [252] and [225], provide a thorough examination of early video descriptors and their effectiveness for action recognition in the context of small labeled datasets. They explore how different feature extraction methods contributed to the performance of action recognition models and discuss the limitations of traditional methods when applied to large-scale datasets. While informative, these surveys do not emphasize the evolving role of deep learning, and therefore miss the pivotal transition from handcrafted features to learned representations that would shape future advancements in the field.

**Table 3: The journey of action recognition datasets: An overview of their evolution over time. This table includes detailed statistics, covering key aspects such as sensors, modalities, and characteristics, providing insights into their diversity and scope.**

| Datasets | Year | #Classes | #Subjects | #Views | #Video clips | Sensor | Modalities | Dataset type |
|---|---|---|---|---|---|---|---|---|
| KTH[201] | 2004 | 6 | 25 | 1 | 2391 | Static camera | RGB | Human actions (*e.g.*, walking, jogging) |
| Weizmann[74] | 2005 | 10 | 9 | 1 | 90 | - | RGB | Human actions (*e.g.*, jumping, running) |
| IXMAS[279] | 2006 | 11 | 10 | 5 | 330 | - | RGB | Movie Scenes (*e.g.*, kissing, running) |
| Hollywood[118] | 2008 | 8 | - | - | 1422 | - | RGB | Movie Scenes (*e.g.*, eating, driving) |
| Hollywood2[155] | 2009 | 12 | - | - | 1709 | - | RGB | Movie Scenes (*e.g.*, running, kissing) |
| ADL[160] | 2009 | 10 | 5 | - | 150 | Static camera | RGB | Daily Activities (*e.g.*, brushing teeth, reading) |
| Olympic Sports[171] | 2010 | 16 | - | - | 783 | - | RGB | Sports (*e.g.*, high jumping, diving) |
| MSRAction3D[132] | 2010 | 20 | 10 | 1 | 567 | Kinect v1 | Depth+3DJoints | Daily Activities (*e.g.*, drinking, walking) |
| CAD-60[227] | 2011 | 14 | 4 | - | 68 | Kinect v1 | RGB+Depth+3DJoints | Human performing activities (*e.g.*, cleaning objects) |
| HMDB51[114] | 2011 | 51 | - | - | 6,766 | - | RGB | Human actions (*e.g.*, jumping, running) |
| MSRDailyActivity3D[246] | 2012 | 16 | 10 | 1 | 320 | Kinect v1 | RGB+Depth+3DJoints | Daily Activities (*e.g.*, calling, playing game) |
| UCF101[221] | 2012 | 101 | - | - | 13,320 | - | RGB | Body motion, Human-object interactions, sports *etc.* |
| UTKinect-Action3D[284] | 2012 | 10 | 10 | 1 | 199 | Kinect v1 | RGB+Depth+3DJoints | Human actions (*e.g.*, waving hands, pushing) |
| MPII Cooking[193] | 2012 | 64 | 12 | 1 | 3,748 | - | RGB | Cooking |
| G3D-Gaming[15] | 2012 | 20 | 10 | 1 | - | Kinect v1 | RGB+Depth+3DJoints | Gaming scenario (*e.g.*, defending, climbing) |
| Berkeley MHAD[173] | 2013 | 11 | 12 | 4 | 660 | Multi-baseline stereo cameras | RGB+Depth+3DJoints+Accelerometer+Audio | Human actions (*e.g.*, throwing, clapping hands) |
| CAD-120[112] | 2013 | 10 | 4 | - | 120 | Kinect v1 | RGB+Depth+3DJoints | Human performing activities (*e.g.*, picking objects) |
| UCF50[191] | 2013 | 50 | - | - | 6676 | - | RGB | Body motion, Human-object interactions, sports *etc.* |
| Florence3D-Action[203] | 2013 | 9 | 10 | 1 | 215 | Kinect v1 | RGB+Depth+3DJoints | Human actions (*e.g.*, bowing, drinking) |
| MSRActionPairs3D[174] | 2013 | 12 | 10 | 1 | 360 | Kinect v1 | RGB+Depth+3DJoints | Human actions (*e.g.*, picking up, putting down) |
| Sports-1M[101] | 2014 | 487 | - | - | 1,000,000 | - | RGB | Sports (*e.g.*, swimming, skiing) |
| THUMOS14[91] | 2014 | 101 | - | - | 5,613 | - | RGB | Human Actions (*e.g.*, making up, archery) |
| Northwestern-UCLA[248] | 2014 | 10 | 10 | 3 | 1494 | Kinect v1 | RGB+Depth+3DJoints | Human actions (*e.g.*, dropping trash) |
| UWA3D Multiview Activity[187] | 2014 | 30 | 10 | 1 | 701 | Kinect v1 | RGB+Depth+3DJoints | Daily Activities (*e.g.*, holding head, walking) |
| ActivityNet[18] | 2015 | 203 | - | - | 27,801 | - | RGB | Human actions (*e.g.*, drawing, washing) |
| MPII Cooking 2[194] | 2015 | 67 | 30 | 1 | 273 | Static camera | RGB | Cooking |
| UWA3D Multiview Activity II[186] | 2015 | 30 | 9 | 4 | 1,070 | Kinect v1 | RGB+Depth+3DJoints Daily Activities | (*e.g.*, waving head, jumping) |
| SYSU 3D HOI[87] | 2015 | 12 | 40 | - | 480 | Kinect v1 | RGB+Depth+3DJoints | Human-Object Interactions (*e.g.*, sweeping the floor) |
| NTU RGB+D[205] | 2016 | 60 | 40 | 80 | 56,880 | Kinect v2 | RGB+Depth+3DJoints | Daily actions, health-realted actions *etc.* |
| InfAR[68] | 2016 | 12 | 40 | - | 600 | Infrared camera | Infrared | Human actions (*e.g.*, jogging) |
| TSF[235] | 2016 | 2 | - | 1 | 44 | FLIR ONE | Infrared | Falls and normal activities |
| Charades[214] | 2016 | 157 | - | - | 66,500 | - | RGB+Flow | Indoor activities (*e.g.*, cleaning) |
| PKU-MMD I[137] | 2017 | 51 | 66 | 3 | 1,076 | Kinect v2 | RGB+Depth+Infrared+3DJoints | Human actions (*e.g.*, walking) |
| NfS[66] | 2017 | - | - | - | 100 | 240 FPS camera | RGB | Visual object tracking |
| Kinetics-400[103] | 2017 | 400 | - | - | 306,245 | - | RGB | Human-centered actions (*e.g.*, playing instruments) |
| Something-Something V1[77] | 2017 | 174 | - | - | 108,499 | - | RGB | Human performing actions with everyday objects |
| Kinetics-skeleton[294] | 2017 | 400 | - | - | 260,232 | - | 2DJoints | Human-centered actions |
| HACS[318] | 2017 | 200 | - | - | 1,500,000 | - | RGB+Flow | Human actions (*e.g.*, dancing) |
| Charades-Ego[213] | 2018 | 157 | 112 | 2 | 68,536 | Head-mounted+standard camera | RGB | Egocentric indoor activities |
| AVA[79] | 2018 | 80 | - | - | 211,000 | - | RGB+Flow | Human actions (*e.g.*, talking, sitting) |
| Diving48[133] | 2018 | 48 | - | - | 18,404 | - | RGB+Flow | Diving actions |
| Epic-Kitchens[38] | 2018 | 149 | 32 | - | 39,594 | - | RGB+Flow | Cooking |
| Something-Something V2[77] | 2018 | 174 | - | - | 220,847 | - | RGB | Human performing actions with everyday objects |
| MiT[165] | 2018 | 339 | - | - | 1,000,000+ | - | RGB+Audio+Flow | Dynamic actions (*e.g.*, human, animals) |
| Kinetics-600[19] | 2018 | 600 | - | - | 495,547 | - | RGB | Human-centered actions (*e.g.*, playing instruments) |
| NTU RGB+D 120[142] | 2019 | 120 | 106 | 155 | 114,480 | Kinect v2 | RGB+Depth+3DJoints+Infrared | Daily actions, health-realted actions *etc.* |
| IITR-IAR[92] | 2019 | 21 | 35 | - | 1,470 | FLIR T1020 | Infrared | Human actions (hugging, fighting) |
| Kinetics-700[20] | 2019 | 700 | - | - | 650,317 | - | RGB | Human-centered actions (*e.g.*, playing instruments) |
| HowTo100M[162] | 2019 | 23,611 | - | - | 136,000,000 | - | RGB | Instructional videos (*e.g.*, cooking) |
| CATER[71] | 2019 | 301 | - | - | 5,500 | - | RGB | Compositional actions and temporal reasoning |
| FineGym[206] | 2020 | 530 | - | - | 32,697 | - | RGB | gymnasium videos (*e.g.*, balance beam) |
| PKU-MMD II[144] | 2020 | 41 | 13 | 3 | 1,009 | Kinect v2 | RGB+Depth+Infrared+3DJoints | Human actions (*e.g.*, standing) |
| EPIC-KITCHENS-100[39] | 2020 | 4,053 | 37 | - | 89,977 | GoPro Hero7 Black | RGB+Flow | Cooking |
| UAV-Human[127] | 2021 | 155 | 119 | - | 22,476 | UAV Camera | RGB+3DJoints | Human Actions (*e.g.*, walking, jogging) |

As deep learning began to dominate, surveys such as [111] and [80] explore the impact of convolutional neural networks (CNNs), 3D CNNs, and two-stream architectures on action recognition. These works review the evolution from simple 2D architectures to more complex 3D and temporal models, which incorporate the crucial temporal dimension alongside spatial features. They also focus on datasets like UCF101, HMDB51, and Kinetics, which play a significant role in advancing the field. While these surveys provide in-depth analyses of different architectures, they are often limited in scope, concentrating mainly on model innovations without exploring the full spectrum of learning paradigms, such as unsupervised and self-supervised learning, that would later play a crucial role in overcoming the challenges posed by limited labeled data.

In more recent years, there has been an increasing interest in action recognition models that use large-scale multimodal datasets and emerging learning paradigms, including self-supervised learning, few-shot learning, and transfer learning. Surveys such as [175], [167], and [276] provide insights into these newer approaches and highlight how they use large video datasets, *e.g.*, Kinetics-400, to improve action recognition performance. These works focus on the advantages of training models on large, diverse datasets and discuss the trade-offs between supervised and unsupervised learning. However, they often treat learning paradigms and model architectures as separate topics, without sufficiently considering how they co-evolve in tandem with the increasing complexity of data. A few recent surveys, such as [46, 95, 100, 115], have started to address the role of transformers and vision-language models in action recognition, emphasizing the growing importance of incorporating semantic context into video understanding. These works explore how transformers, with their ability to model long-range dependencies, have become a powerful tool for capturing temporal dynamics in action recognition tasks. While these surveys acknowledge the synergy between models like BERT or GPT and vision models, they generally do not delve deeply into how these models interact with the data and learning paradigms to shape the development of action recognition.

**Differences from existing work.** While existing surveys on action recognition have made significant contributions by examining various aspects of the field, there are key differences in the scope and focus of this work. First, this paper takes a more holistic approach by integrating the perspectives of data, learning paradigms, and model architectures in a unified framework. Rather than treating each component in isolation, we explore how they co-evolve and influence one another, offering a comprehensive understanding of the factors driving advancements in action recognition.

Second, this paper provides a deeper exploration of the evolution of learning paradigms in action recognition, including the transition from supervised learning to unsupervised, self-supervised, and few-shot learning. This is an important distinction, as it highlights the increasing reliance on large-scale unlabeled datasets and the emergence of pretraining techniques, which are essential for handling the complexities of modern video datasets. Unlike many existing surveys that focus primarily on model architectures or specific datasets, this work emphasizes the shifting learning paradigms that are enabling the field to scale and generalize.

Finally, this work incorporates a forward-looking perspective by discussing the integration of vision-language models and transformers in action recognition, which are still underexplored in existing surveys. While other surveys mention these advances, they often fail to address their potential for cross-modal learning and the broader impact on video processing tasks. This paper not only examines the impact of transformers and language models on action recognition, but also explores how these developments contribute to the broader landscape of video understanding, anomaly detection, captioning, and beyond.

## 3 From Actions to Insights

In this section, we systematically explore the evolution of action recognition through the interconnected lenses of data, model architectures, and learning paradigms. We delve into how each perspective has driven advancements in the field, while highlighting their co-evolution, showcasing how innovations in one domain have influenced and been shaped by progress in the others. Tables 1 and 2 show the progression of action recognition methods, while Table 3 highlights the evolution of action recognition datasets.

### 3.1 From a Data Perspective

The development of action recognition has been fundamentally shaped by the evolution of datasets, which act as the foundation for learning paradigms and model architectures. This journey shows a dynamic interplay between the characteristics of data and technological advancements in extracting meaningful patterns, leading to a continuous refinement of methods.

**Data evolution and paradigm shifts.** Early datasets like KTH [201], Hollywood2 [155], and Olympic Sports [171] mark the initial phase of action recognition research. These datasets, collected in controlled environments, feature a limited number of subjects and simple actions such as walking, waving, or running. Their simplicity inspires researchers to focus on handcrafted features [159, 190, 242], such as Histogram of Oriented Gradients (HOG) [37] and dense trajectories [241]. These manually crafted descriptors, combined with traditional classifiers like Support Vector Machines (SVMs), excel in recognizing these straightforward actions.

The next wave of datasets, such as HMDB51 [114], UCF101 [221], and Sports-1M [101], introduce more diversity in terms of actions, scenes, and contexts. The increased scale and variety requires a paradigm shift towards data-driven methods [152, 154, 223]. These datasets facilitate the adoption of deep learning, as convolutional neural networks (CNNs) could now exploit the broader representation power of larger and more complex datasets [304].

Larger-scale datasets like the Kinetics family [19, 20, 103], Something-Something V1 and V2[77], and Moments in Time [165] further push the field towards supervised learning. These datasets, with millions of labeled videos, provide the necessary foundation for deep models to achieve state-of-the-art results [58, 108, 297]. However, the high cost of annotating video data leads to innovations in unsupervised and self-supervised learning. For instance, unlabeled datasets like HowTo100M [162] spur progress in contrastive learning approaches [61, 73, 81], while multimodal datasets, such as video-text pairs from ActivityNet Captions [113] and WebVid [8], enable breakthroughs in vision-language models like CLIP [184] and Flamingo [4]. These advancements demonstrate how the evolution of datasets directly influences paradigm shifts, from supervised learning to unsupervised, self-supervised, and multimodal approaches. Each paradigm addresses the growing complexity and scale of modern video data.

**Learning paradigms driven by data.** The nature of datasets plays a pivotal role in determining the choice of learning paradigms. Supervised learning thrives on large, labeled datasets, where explicit annotations like action labels provide clear supervision signals. However, challenges such as noisy labels and class imbalance in real-world datasets can degrade performance, necessitating robust loss functions and data augmentation techniques [36, 75, 306].

Unsupervised learning, by contrast, eliminates the reliance on labels and aims to learn generalizable representations. For example, methods like MoCo [61] and BYOL [78] use contrastive learning to distinguish video instances based on their spatiotemporal features. These methods benefit from diverse datasets with varied contexts, enabling the model to capture a broad range of patterns [69, 195]. However, the lack of labels complicates evaluation, as metrics often depend on downstream tasks [50, 223]. Few-shot and zero-shot learning paradigms address the scarcity of labeled examples [129, 276]. Few-shot methods, such as prototypical networks [217], rely on curated support sets to generalize across classes. Zero-shot approaches [2, 106], powered by vision-language models, use textual descriptions to infer knowledge about unseen actions. For example, CLIP [184] can recognize actions like "playing guitar" by aligning visual features with corresponding textual embeddings, even when such actions are absent in the training data [305]. Self-supervised learning builds on unlabeled data through pretext tasks, such as temporal order prediction or video masking [59, 230]. These tasks encourage the model to learn useful features without explicit supervision. However, the design of pretext tasks must align with downstream objectives; otherwise, the learned representations may not generalize effectively.

**Architectural innovation.** Video datasets, unlike static image datasets such as ImageNet [195], introduce temporal complexity, requiring specialized architectures. The sequential nature of video data drives innovation in model design to capture both spatial and temporal dependencies. Early attempts [63, 67, 101, 311] to adapt 2D CNNs for video processing fall short, as they are ill-equipped to handle temporal relationships. This limitation leads to the development of 3D CNNs and two-stream networks, such as C3D[231] and I3D[21], which either extend convolutional operations into the temporal dimension to capture motion dynamics or model spatial and temporal information separately. More recently, transformers [7, 54, 170, 177] have emerged as a powerful alternative. Models like TimeSformer[13] and Video Swin Transformer [150] use attention mechanisms to capture long-range temporal dependencies, making them particularly effective for large-scale and complex datasets. These architectures outperform earlier methods in tasks requiring fine-grained temporal reasoning [88, 196, 251].

Multimodal datasets [137, 205, 214, 284] have further driven the design of architectures that integrate multiple data types. For example, models like CLIP [184] and Flamingo [4] fuse video and textual information, enabling cross-modal reasoning. Similarly, methods tailored for RGB-D data (*e.g.*, combining RGB frames with depth maps) use specialized components to process the complementary modalities effectively. Data augmentation and preprocessing also influence architectural choices. For instance, datasets with high variability in lighting, viewpoint, or action dynamics require architectures with robust components like dropout layers or attention mechanisms [39, 77, 142, 205]. Self-supervised models [5, 273] benefit from contrastive augmentation techniques, where diverse crops or temporal shifts enhance the model's ability to learn invariant spatiotemporal features. Finally, the scale of datasets dictates the complexity of models. Large datasets enable the training of deeper architectures with millions of parameters, while smaller datasets necessitate simpler models or the use of transfer learning [1]. Pretrained models on large visual datasets (*e.g.*, Kinetics [103] or ImageNet [195]) can be fine-tuned to smaller, domain-specific datasets, demonstrating how data availability shapes model design [42]. Representative models include [59, 120, 196, 230, 251, 272, 273, 305].

The journey of action recognition datasets underscores their central role in shaping the field. From early handcrafted descriptors [283] to cutting-edge transformers [13] and multimodal models [271], the evolution of datasets has driven progress in both learning paradigms and architectures. As datasets become increasingly diverse and complex, they will continue to inspire innovations in action recognition, pushing the boundaries of what machines can learn from video data.

## 3.2 From a Model Perspective

The journey of action recognition models has been shaped by the interplay between data characteristics and the demand for capturing spatiotemporal relationships. Early approaches, intermediate innovations, and the latest breakthroughs all reflect how the challenges and opportunities in data have driven model evolution.

**Early models: handcrafted descriptors and motion-aware designs.** Initial attempts at action recognition rely heavily on handcrafted descriptors tailored to conventional RGB videos [118, 190]. These methods focus on extracting spatiotemporal and motion information. For example, spatiotemporal features like 3D-SIFT [202], extended SURF [281], HOG3D [107], and local trinary patterns [308] are developed to analyze relationships across frames. These descriptors effectively capture the dynamics of simple actions (*e.g.*, walking, waving) in controlled settings. However, they struggle with the complexity of real-world videos, particularly when camera motion introduces noise [323]. To address these challenges, dense trajectories [240] and improved dense trajectories [242] emerge as robust solutions. By tracking local features through video frames, these methods mitigate the impact of camera motion and enabled better representation of dynamic actions. Bag-of-visual-words [178] and Fisher vector embeddings [119] further enhance their effectiveness, allowing these descriptors to achieve significant success despite limited training data.

**Deep learning revolution: spatiotemporal feature learning.** The advent of large-scale datasets like Sports-1M and Kinetics-400 catalyzes a paradigm shift toward learned feature representations [233]. Inspired by the success of 2D CNNs in image recognition, researchers initially explore 2D networks with temporal aggregation, such as CNN-LSTM[311] and TSN[264], which fuse spatial features across frames. However, these methods lack the capacity to fully capture temporal dynamics [275].

To overcome these limitations, models like two-stream ConvNets [215] and 3D CNNs (*e.g.*, C3D[231] and I3D[21] are introduced. Two-stream architectures use separate branches for spatial and motion information, often using optical flow [116, 260] for the motion stream. Meanwhile, 3D CNNs extend convolutional operations into the temporal dimension, directly modeling spatiotemporal features [183]. Despite their success, both approaches face challenges: two-stream models incur high computational costs [134], while 3D CNNs require extensive data and computational resources [108].

Innovations like (2+1)D convolution decompose 3D operations into separate spatial and temporal components, balancing efficiency and performance [99, 238]. Examples include R(2+1)D networks[234] and their integration with transformers [97], which enhance the ability to model long-range temporal dependencies.

**Transformer era and multimodal integration.** Transformers have redefined action recognition by introducing global attention mechanisms [222]. Vision transformers (ViTs) initially demonstrate the potential for spatial feature extraction in videos [51]. Subsequent transformer-based video models, such as TimeSformer[13] and Motionformer [177], extend this approach to capture complex spatiotemporal relationships. These models excel at handling diverse data distributions and variability in lighting, scale, and viewpoint [88, 129, 251].

Recent advancements include video masked autoencoders (*e.g.*, VideoMAE [230] and VideoMAE V2 [251]), which use self-supervised learning to extract spatial and temporal representations. These architectures, inspired by masked autoencoders in image tasks [84], have set new benchmarks in efficiency and performance for video analysis. Simultaneously, multimodal models such as CLIP [184] and BLIP [124] have integrated video and text data, unlocking new capabilities in action recognition. By aligning video frames with textual descriptions, these models facilitate tasks like zero-shot action recognition and general-purpose video understanding [106, 135, 200]. This integration has paved the way for applications extending beyond action recognition, including video captioning and anomaly detection [45, 47, 295, 324].

**Expanding modalities: depth, skeleton, and large foundation models.** The introduction of depth videos and skeleton sequences through devices like the Microsoft Kinect expands the scope of action recognition [185]. Depth-based models, such as HON4D [174] and HOPC [187], effectively segment human subjects in cluttered scenes, while skeleton-based models capitalize on the structural and temporal continuity of 3D joint movements [53, 299, 321]. Handcrafted skeleton features (*e.g.*, LARP-SO [236]) evolve into learned representations like ST-GCN [294] and its successors [34, 121, 140, 169, 219, 220, 254, 259, 300, 303, 322], including ShiftGCN [30] and CTR-GCN [27]. These graph-based models advance the field by using human pose information for more accurate action recognition. Point cloud-based methods include [10, 164, 274].

Large foundation models like InternVideo2 [272] represent the latest milestone in action recognition. Trained on vast, multimodal datasets, these models demonstrate exceptional versatility across video processing tasks [245, 271, 273]. They exemplify how increased data volume and multimodal integration enable the development of deeper, more powerful architectures, bridging the gap between specialized tasks and general video understanding [5, 46].

**Insights.** The evolution of action recognition models underscores a recurring theme: data characteristics dictate model design. Early handcrafted methods prioritize robustness to motion noise, while deep learning models embrace scale and diversity [175]. Transformers and multimodal architectures have further transformed the field, emphasizing the importance of flexibility and scalability [271, 272]. As video data continues to grow in complexity and volume, future models must navigate challenges such as motion diversity, temporal resolution, and ethical considerations in data use. This journey, driven by both data availability and computational advances, highlights the symbiotic relationship between datasets and model architectures in shaping the trajectory of action recognition.

## 3.3 From a Learning Perspective

The evolution of action recognition models is closely tied to the development of learning paradigms, each offering unique insights and solutions to the challenges posed by video data. From supervised methods relying on large labeled datasets to emerging paradigms like self-supervised and zero-shot learning, the journey reflects a dynamic interplay between data availability, model architecture, and task complexity.

**The supervised learning era.** Supervised learning has been the dominant paradigm in action recognition for decades [67, 101, 231, 325]. Early models rely on fully labeled datasets, where each video is paired with a specific label, such as an action category or bounding box. This explicit mapping between inputs and outputs, guided by loss functions like cross-entropy, enables models to learn spatiotemporal patterns effectively [317]. However, the reliance on high-quality labeled datasets introduces limitations [105]. Labeling video data is costly, time-consuming, and prone to biases, such as noisy labels or skewed class distributions, which degrade model performance. Despite these challenges, supervised learning establishes foundational architectures, including convolutional neural networks (CNNs) [49, 288, 311] and two-stream networks [134, 263, 320], that excel in tasks requiring spatial and motion analysis. Pretraining on large-scale datasets like Kinetics [19, 20, 103] allows models to capture diverse motion patterns, reducing the need for task-specific data through transfer learning [21]. This paradigm demonstrates how large labeled datasets can accelerate progress but also highlights the necessity for alternative approaches to address scalability and diversity challenges.

**The rise of self-supervised and semi-supervised learning.** To overcome the dependence on labeled data, self-supervised learning emerges as a powerful alternative [12, 81]. In this paradigm, models generate pseudo-labels from the data itself, using auxiliary tasks such as predicting motion trajectories [291], solving spatiotemporal puzzles [237], or reconstructing masked regions [230]. Methods like contrastive learning (*e.g.*, SimCLR [26], MoCo [61] and video masked autoencoders (*e.g.*, VideoMAE [230]) demonstrate the ability to learn high-quality spatiotemporal features without explicit supervision [251]. These approaches use data augmentation to create positive and negative pairs, enabling models to distinguish between similar and dissimilar samples [307].

Self-supervised learning has proven particularly effective for pretraining on large-scale unlabeled datasets, significantly enhancing performance on downstream tasks like action recognition. For instance, VideoMAE models, pretrained on small datasets like HMDB51, achieve competitive results, showcasing the paradigm's efficiency in using limited data [230, 251]. Semi-supervised learning bridges the gap between supervised and self-supervised approaches by combining small amounts of labeled data with large volumes of unlabeled data [98, 216, 285, 290]. This paradigm reduces the reliance on extensive labeling efforts, using labeled examples to guide the learning of representations from unlabeled data. Semi-supervised techniques have proven valuable in scenarios where labeled video data is scarce or expensive to obtain.

**Emerging paradigms: few-shot, zero-shot, and unified learning.** Recent advancements [2, 200, 289, 309] have focused on making action recognition models more flexible and adaptable. Few-shot learning enables models to generalize to new action categories using only a handful of labeled examples. Architectures like prototypical networks [217] and relation networks [226] are designed to perform well under limited data conditions, using meta-learning principles. Zero-shot learning goes a step further, enabling models to classify unseen action categories using multimodal inputs, such as textual descriptions or video-text pairs [106]. Models like CLIP [184] demonstrate the effectiveness of vision-language pretraining in achieving generalization across tasks.

Transformers have been instrumental in advancing these paradigms [13, 170]. Originally developed for natural language processing [282], transformers excel in multimodal and unified learning settings. Their attention mechanisms capture long-range dependencies, enabling robust temporal dynamics modeling [35, 172]. By integrating vision and text modalities, transformers facilitate cross-domain learning, paving the way for unified multimodal frameworks capable of handling diverse tasks, from action recognition to video question answering [312].

**Insights.** The trajectory of action recognition learning paradigms underscores the evolving role of data. Labeled datasets have driven supervised learning, while unlabeled and multimodal datasets fuel the rise of self-supervised, semi-supervised, and zero-shot approaches [175]. The interplay between data characteristics and learning methods has shaped models, from CNNs to vision transformers [7, 13, 49]. Future innovations will likely focus on unified learning paradigms that integrate multimodal data and use pretrained video foundation models for broader generalization across tasks.

## 4 Future Directions

In this section, we highlight three key areas poised to shape the future of action recognition: multimodal integration, transformer-based architectures, and vision-language models (VLMs). These directions not only aim to enhance model performance but also tackle some of the most pressing challenges in video understanding.

**Integration of multimodal data.** As video data alone often fails to capture the full complexity of actions, integrating multimodal data (visual, auditory, and textual) has become a critical focus in advancing action recognition. This integration enables models to use complementary information, such as speech, environmental sounds, or contextual text, to better understand actions in diverse and noisy settings. For example, recognizing an action like "talking on the phone" becomes more accurate when the auditory signal (speech) is paired with visual information (body language). The ability to simultaneously process multiple data streams presents new challenges in synchronizing and aligning heterogeneous modalities, but the potential for more robust and nuanced action recognition is vast. This shift to multimodal systems may help models understand actions with greater contextual awareness, reducing ambiguity and improving performance in real-world applications where visual cues alone are often insufficient.

**Transformer-based architectures.** The rise of transformer-based architectures represents a monumental shift in how temporal dependencies are modeled in action recognition. Unlike traditional CNNs, which rely on local spatial filters, transformers excel at capturing long-range dependencies across sequences, making them ideal for video data where context over time is crucial. Transformers enable better modeling of complex temporal dynamics, such as long-range interactions between frames or global motion patterns that span the entire video. By using self-attention mechanisms, transformers can selectively focus on relevant parts of the video sequence, allowing for more accurate action classification, even in the presence of noise or occlusions. This ability to handle long-range dependencies also opens the door to more sophisticated methods for action recognition in dynamic and highly variable environments, such as sports or surveillance footage, where actions are often interdependent and occur over extended periods. While transformer models are computationally intensive, their increasing efficiency and scalability make them a promising avenue for the next generation of action recognition systems.

**Vision-language models.** Another transformative trend in action recognition is the integration of vision-language models (VLMs), which combine the understanding of visual content with linguistic representations. These models have the potential to overcome one of the biggest challenges in action recognition: understanding ambiguous or context-dependent actions. By incorporating natural language processing (NLP) techniques, VLMs can infer the meaning behind a sequence of actions in a video based on textual descriptions or situational context. For instance, the action of "grabbing a cup" could be interpreted differently based on the surrounding environment or verbal cues, such as "grabbing a cup of coffee" versus "grabbing a cup to throw". This alignment between vision and language facilitates more comprehensive reasoning about actions and allows models to handle complex, abstract tasks like action sequencing, goal recognition, and activity prediction. Furthermore, VLMs enable the development of systems that can interact with users or adapt to specific contexts, making them highly applicable for interactive media, autonomous systems, and personalized healthcare applications.

**Potential for cross-domain advancements.** The integration of these emerging trends also opens new opportunities for cross-domain advancements in action recognition. Multimodal data and transformer architectures, for instance, can be combined to tackle complex video datasets where both long-range temporal dependencies and multimodal context are essential. Similarly, VLMs can be enhanced with transformer-based architectures to refine the attention mechanisms, improving both the understanding of temporal dynamics and the contextual alignment between visual and linguistic data. These hybrid approaches not only promise to address current challenges but also pave the way for a new generation of action recognition systems that are adaptable, context-aware, and capable of reasoning about actions in a human-like manner.

The future of action recognition lies in the intersection of multimodal data integration, transformer-based architectures, and VLMs. By addressing the challenges of temporal complexity, contextual ambiguity, and cross-domain generalization, these trends have the potential to revolutionize the field, making action recognition more accurate, adaptable, and robust across diverse real-world applications. As these technologies mature, we anticipate a significant leap forward in how video content is understood and processed, leading to more intelligent systems that can interpret, predict, and interact with the world in ways previously imagined only in science fiction.

## 5 Conclusion

Action recognition has evolved significantly, driven by advancements in data, model architectures, and learning paradigms. Initially relying on handcrafted features and small labeled datasets, the field shifted with the advent of large-scale video datasets and learned representations, using models like 2D, 3D, and (2+1)D CNNs, and GCNs. As video data grow more complex, innovative learning paradigms, such as self-supervised, few-shot, and contrastive learning, help harness the power of large, unlabeled datasets. The introduction of transformer-based models marks a key milestone, enhancing the ability to capture temporal dynamics. Masked autoencoders improve the balance between spatial and temporal features, while the integration of language models enriched action recognition with semantic context. The rise of video foundation models, combining image, video, and language data, has expanded the scope of action recognition to include broader video processing tasks, such as anomaly detection and video captioning. Ultimately, the evolution of action recognition has transformed it into a core element of general video processing, offering insights for future challenges and opportunities in video analysis and beyond.

## Acknowledgments

Xi Ding, a Research Assistant with the Temporal Intelligence and Motion Extraction (TIME) Lab at ANU, contributed to this work. TIME Lab is a dynamic research team comprising master's and honours students focused on advancing video processing and motion analysis. This research was conducted under the supervision of Lei Wang.


## References


[1] Yousry M Abdulazeem, Hossam Magdy Balaha, Waleed M. Bahgat, and Mahmoud Badawy. 2021. Human Action Recognition Based on Transfer Learning Approach. *IEEE Access* 9 (2021), 82058–82069. https://api.semanticscholar.org/CorpusID:235406558

[2] Shahzad Ahmad, Sukalpa Chanda, and Yogesh Singh Rawat. 2023. EZ-CLIP: Efficient Zeroshot Video Action Recognition. *ArXiv* abs/2312.08010 (2023). https://api.semanticscholar.org/CorpusID:266191106

[3] Aparna Akula, Anuj K Shah, and Ripul Ghosh. 2018. Deep learning approach for human action recognition in infrared images. *Cognitive Systems Research* 50 (2018), 146–154.

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *ArXiv* abs/2204.14198 (2022). https://api.semanticscholar.org/CorpusID:248476411

[5] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. *Advances in neural information processing systems* 33 (2020), 25–37.

[6] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2020. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems* 33 (2020), 9758–9770.

[7] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6836–6846.

[8] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 1708–1718. https://api.semanticscholar.org/CorpusID:232478955

[9] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*. Springer, 404–417.

[10] Yizhak Ben-Shabat, Oren Shrout, and Stephen Gould. 2024. 3dinaction: Understanding human actions in 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19978–19987.

[11] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. 2021. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 847–859.

[12] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T. Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. 2020. SpeedNet: Learning the Speediness in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding?. In *ICML*, Vol. 2. 4.

[14] Hakan Bilen, Basura Fernando, Efstratios Gavves, and Andrea Vedaldi. 2017. Action recognition with dynamic image networks. *IEEE transactions on pattern analysis and machine intelligence* 40, 12 (2017), 2799–2813.

[15] Victoria Bloom, Dimitrios Makris, and Vasileios Argyriou. 2012. G3D: A gaming action dataset and real time action recognition evaluation framework. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 7–12. https://doi.org/10.1109/CVPRW.2012.6239175

[16] Matteo Bregonzio, Shaogang Gong, and Tao Xiang. 2009. Recognising action as clouds of space-time interest points. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 1948–1955.

[17] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. 2021. Space-time mixing attention for video transformer. *Advances in neural information processing systems* 34 (2021), 19594–19607.

[18] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 961–970.

[19] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340* (2018).

[20] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987* (2019).

[21] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[22] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. 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)*. 168–172. https://doi.org/10.1109/ICIP.2015.7350781

[23] Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. 2021. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6165–6175.

[24] Huilin Chen, Lei Wang, Yifan Chen, Tom Gedeon, and Piotr Koniusz. 2024. When spatial meets temporal in action recognition. *arXiv preprint arXiv:2411.15284* (2024).

[25] Qixiang Chen, Lei Wang, Piotr Koniusz, and Tom Gedeon. [n. d.]. Motion meets attention: Video motion prompts. In *The 16th Asian Conference on Machine Learning (Conference Track)*.

[26] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[27] Yuxin Chen, Ziqi Zhang, Chunfen Yuan, Bing Li, Ying Deng, and Weiming Hu. 2021. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 13339–13348. https://api.semanticscholar.org/CorpusID:236428765

[28] Zhaoyu Chen, Xing Li, Qian Huang, Qiang Geng, Tianjin Yang, and Shihao Han. 2024. KAN-HyperpointNet for Point Cloud Sequence-Based 3D Human Action Recognition. *arXiv preprint arXiv:2409.09444* (2024).

[29] Feng Cheng and Gedas Bertasius. 2022. Tallformer: Temporal action localization with a long-memory transformer. In *European Conference on Computer Vision*. Springer, 503–521.

[30] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. 2020. Skeleton-Based Action Recognition With Shift Graph Convolutional Network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 180–189. https://doi.org/10.1109/CVPR42600.2020.00026

[31] Qin Cheng, Jun Cheng, Zhen Liu, Ziliang Ren, and Jianming Liu. 2024. A Dense-Sparse Complementary Network for Human Action Recognition based on RGB and Skeleton Modalities. *Expert Systems with Applications* 244 (2024), 123061.

[32] Anoop Cherian, Basura Fernando, Mehrtash Harandi, and Stephen Gould. 2017. Generalized rank pooling for activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3222–3231.

[33] Guilhem Cheron, Ivan Laptev, and Cordelia Schmid. 2015. P-CNN: Pose-Based CNN Features for Action Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[34] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. 2022. InfoGCN: Representation Learning for Human Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20186–20196.

[35] Sanghyuk Roy Choi and Minhyeok Lee. 2023. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology* 12, 7 (2023), 1033.

[36] Mickael Cormier, Yannik Schmid, and Jürgen Beyerer. 2024. Enhancing Skeleton-Based Action Recognition in Real-World Scenarios Through Realistic Data Augmentation. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* (2024), 300–309. https://api.semanticscholar.org/CorpusID:269191024

[37] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. 886–893 vol. 1. https://doi.org/10.1109/CVPR.2005.177

[38] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*. 720–736.

[39] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision* (2022),

1–23.
[40] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. 2019. Toyota Smarthome: Real-World Activities of Daily Living. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
[41] Long Deng, Ao Li, Bingxin Zhou, and Yongxin Ge. 2024. Two-Stream Temporal Feature Aggregation Based on Clustering for Few-Shot Action Recognition. *IEEE Signal Processing Letters* 31 (2024), 2435–2439. https://doi.org/10.1109/LSP.2024.3456670
[42] Sourish Gunesh Dhekane and Thomas Ploetz. 2024. Transfer Learning in Human Activity Recognition: A Survey. *ArXiv* abs/2401.10185 (2024). https://api.semanticscholar.org/CorpusID:267034857
[43] Chhavi Dhiman and Dinesh Kumar Vishwakarma. 2020. View-Invariant Deep Architecture for Human Action Recognition Using Two-Stream Motion and Shape Temporal Dynamics. *IEEE Transactions on Image Processing* 29 (2020), 3835–3844. https://doi.org/10.1109/TIP.2020.2965299
[44] Ali Diba, Vivek Sharma, and Luc Van Gool. 2016. Deep Temporal Linear Encoding Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 1541–1550. https://api.semanticscholar.org/CorpusID:6709077
[45] Dexuan Ding, Lei Wang, Liyun Zhu, Tom Gedeon, and Piotr Koniusz. 2024. Lego: Learnable expansion of graph operators for multi-modal feature fusion. *arXiv preprint arXiv:2410.01506* (2024).
[46] Xi Ding and Lei Wang. 2024. Do Language Models Understand Time? *arXiv preprint arXiv:2412.13845* (2024).
[47] Xi Ding and Lei Wang. 2024. Quo Vadis, Anomaly Detection? LLMs and VLMs in the Spotlight. *arXiv preprint arXiv:2412.18298* (2024).
[48] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. 2005. Behavior recognition via sparse spatio-temporal features. In *2005 IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance*. IEEE, 65–72.
[49] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.
[50] Jianfeng Dong, Shengkai Sun, Zhonglin Liu, Shujie Chen, Baolong Liu, and Xun Wang. 2023. Hierarchical contrast for unsupervised skeleton-based action representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 525–533.
[51] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv* abs/2010.11929 (2020). https://api.semanticscholar.org/CorpusID:225039882
[52] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1110–1118.
[53] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. 2022. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2969–2978.
[54] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6824–6835.
[55] Hehe Fan, Yi Yang, and Mohan Kankanhalli. 2021. Point 4D Transformer Networks for Spatio-Temporal Modeling in Point Cloud Videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14199–14208. https://doi.org/10.1109/CVPR46437.2021.01398
[56] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and M. Kankanhalli. 2022. PSTNet: Point Spatio-Temporal Convolution on Point Cloud Sequences. *ArXiv* abs/2205.13713 (2022). https://api.semanticscholar.org/CorpusID:235613642
[57] Mohsen Fayyaz, Emad Bahrami, Ali Diba, Mehdi Noroozi, Ehsan Adeli, Luc Van Gool, and Jurgen Gall. 2021. 3D CNNs With Adaptive Temporal Feature Resolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4731–4740.
[58] Christoph Feichtenhofer. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 203–213.
[59] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. 2022. Masked Autoencoders As Spatiotemporal Learners. *ArXiv* abs/2205.09113 (2022). https://api.semanticscholar.org/CorpusID:248863181
[60] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6202–6211.
[61] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. 2021. A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3299–3309.
[62] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1933–1941.
[63] Basura Fernando, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars. 2016. Rank pooling for action recognition. *IEEE transactions on pattern analysis and machine intelligence* 39, 4 (2016), 773–787.
[64] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. 2012. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1737–1746.
[65] Luca Franco, Paolo Mandica, Bharti Munjal, and Fabio Galasso. 2023. Hyperbolic self-paced learning for self-supervised skeleton-based action representations. *arXiv preprint arXiv:2303.06242* (2023).
[66] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. 2017. Need for Speed: A Benchmark for Higher Frame Rate Object Tracking. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 1134–1143. https://api.semanticscholar.org/CorpusID:9857301
[67] Chuang Gan, Ting Yao, Kuiyuan Yang, Yi Yang, and Tao Mei. 2016. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 923–932.
[68] Chenqiang Gao, Yinhe Du, Jiang Liu, Jing Lv, Luyu Yang, Deyu Meng, and Alexander G Hauptmann. 2016. Infar dataset: Infrared action recognition at different times. *Neurocomputing* 212 (2016), 36–47.
[69] Deepti Ghadiyaram, Matt Feiszli, Du Tran, Xueting Yan, Heng Wang, and Dhruv Kumar Mahajan. 2019. Large-Scale Weakly-Supervised Pre-Training for Video Action Recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 12038–12047. https://api.semanticscholar.org/CorpusID:143423501
[70] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10406–10417.
[71] Rohit Girdhar and Deva Ramanan. 2019. CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. *arXiv preprint arXiv:1910.04744* (2019).
[72] Rohit Girdhar, Deva Ramanan, Abhinav Kumar Gupta, Josef Sivic, and Bryan C. Russell. 2017. ActionVLAD: Learning Spatio-Temporal Aggregation for Action Classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 3165–3174. https://api.semanticscholar.org/CorpusID:16091693
[73] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. 2020. Watching the World Go By: Representation Learning from Unlabeled Videos. *ArXiv* abs/2003.07990 (2020). https://api.semanticscholar.org/CorpusID:212747934
[74] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. 2007. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence* 29, 12 (2007), 2247–2253.
[75] Shreyank N. Gowda, Marcus Rohrbach, Frank Keller, and Laura Sevilla-Lara. 2022. Learn2Augment: Learning to Composite Videos for Data Augmentation in Action Recognition. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:249605450
[76] Shreyank N Gowda, Laura Sevilla-Lara, Frank Keller, and Marcus Rohrbach. 2022. Claster: clustering with reinforcement learning for zero-shot action recognition. In *European Conference on Computer Vision*. Springer, 187–203.
[77] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*. 5842–5850.
[78] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 1786, 14 pages.
[79] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6047–6056.
[80] Fuqiang Gu, Mu-Huan Chung, Mark H. Chignell, Shahrokh Valaee, Baoding Zhou, and Xue Liu. 2021. A Survey on Deep Learning for Human Activity Recognition. *ACM Computing Surveys (CSUR)* 54 (2021), 1 – 34. https://api.semanticscholar.org/CorpusID:238260765
[81] Tengda Han, Weidi Xie, and Andrew Zisserman. 2020. Self-supervised co-training for video representation learning. *Advances in neural information processing systems* 33 (2020), 5679–5690.
[82] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE*

*conference on Computer Vision and Pattern Recognition*. 6546–6555.

[83] Jun-Yan He, Xiao Wu, Zhi-Qi Cheng, Zhaoquan Yuan, and Yu-Gang Jiang. 2020. DB-LSTM: Densely-connected Bi-directional LSTM for human action recognition. *Neurocomputing* 444 (2020), 319–331. https://api.semanticscholar.org/CorpusID:229448897

[84] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 15979–15988. https://api.semanticscholar.org/CorpusID:243985980

[85] Samitha Herath, Mehrtash Tafazzoli Harandi, and Fatih Murat Porikli. 2016. Going deeper into action recognition: A survey. *Image Vis. Comput.* 60 (2016), 4–21. https://api.semanticscholar.org/CorpusID:14814753

[86] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. 2021. Object-Region Video Transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 3138–3149. https://api.semanticscholar.org/CorpusID:238744000

[87] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. 2015. Jointly learning heterogeneous features for RGB-D activity recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5344–5352. https://doi.org/10.1109/CVPR.2015.7299172

[88] Wenbo Huang, Jinghui Zhang, Xuwei Qian, Zhen Wu, Meng Wang, and Lei Zhang. 2024. SOAP: Enhancing Spatio-Temporal Relation and Motion Information Capturing for Few-Shot Action Recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 4572–4580.

[89] Xiaohu Huang, Hao Zhou, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jingdong Wang, Xinggang Wang, Wenyu Liu, and Bin Feng. 2023. Graph contrastive learning for skeleton-based action recognition. *arXiv preprint arXiv:2301.10900* (2023).

[90] Zhiwu Huang, Chengde Wan, Thomas Probst, and Luc Van Gool. 2017. Deep learning on lie groups for skeleton-based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6099–6108.

[91] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. 2017. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding* 155 (2017), 1–23.

[92] Javed Imran and Balasubramanian Raman. 2019. Deep residual infrared action recognition by integrating local and global spatio-temporal cues. *Infrared Physics & Technology* 102 (2019), 103014.

[93] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. 2013. Towards Understanding Action Recognition. In *2013 IEEE International Conference on Computer Vision*. 3192–3199. https://doi.org/10.1109/ICCV.2013.396

[94] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. 2019. Stm: Spatiotemporal and motion encoding for action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2000–2009.

[95] Zihao Jiang and Yunxiang Liu. 2024. A Review of Human Action Recognition Based On Deep Learning. *2024 9th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)* 9 (2024), 78–83. https://api.semanticscholar.org/CorpusID:274827033

[96] Zhuolin Jiang, Viktor Rozgic, and Sancar Adali. 2017. Learning Spatiotemporal Features for Infrared Action Recognition with 3D Convolutional Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 309–317. https://doi.org/10.1109/CVPRW.2017.44

[97] Hao Jin, Jianming Yang, and Sheng Zhang. 2021. Efficient Action Recognition with Introducing R(2+1)D Convolution to Improved Transformer. In *2021 4th International Conference on Information Communication and Signal Processing (ICICSP)*. 379–383. https://doi.org/10.1109/ICICSP54369.2021.9611970

[98] Longlong Jing, Toufiq Parag, Zhe Wu, Yingli Tian, and Hongcheng Wang. 2021. Videossl: Semi-supervised learning for video classification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1110–1119.

[99] M Esat Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. 2020. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 731–747.

[100] Misha Karim, Shah Khalid, Aliya Aleryani, Jawad Khan, Irfan Ullah, and Z. Ali. 2024. Human Action Recognition Systems: A Review of the Trends and State-of-the-Art. *IEEE Access* 12 (2024), 36372–36390. https://api.semanticscholar.org/CorpusID:268345466

[101] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[102] Takayuki Kawashima, Yasutomo Kawanishi, Ichiro Ide, Hiroshi Murase, Daisuke Deguchi, Tomoyoshi Aizawa, and Masato Kawade. 2017. Action recognition from extremely low-resolution thermal image sequence. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 1–6. https://doi.org/10.1109/AVSS.2017.8078497

[103] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[104] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. 2017. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3288–3297.

[105] Muhammad Attique Khan, Mamta Mittal, Lalit Mohan Goyal, and Sudipta Roy. 2021. A deep survey on supervised learning based human detection and activity classification methods. *Multimedia Tools and Applications* 80, 18 (2021), 27867–27923.

[106] Minji Kim, Dongyoon Han, Taekyung Kim, and Bohyung Han. 2025. Leveraging temporal contextualization for video action recognition. In *European Conference on Computer Vision*. Springer, 74–91.

[107] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. 2008. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 275–1.

[108] D. I. Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew A. Brown, and Boqing Gong. 2021. MoViNets: Mobile Video Networks for Efficient Video Recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 16015–16025. https://api.semanticscholar.org/CorpusID:232307534

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

[110] Piotr Koniusz, Lei Wang, and Anoop Cherian. 2020. Tensor Representations for Action Recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE.

[111] Maryam Koohzadi and Nasrollah Moghaddam Charkari. 2017. Survey on deep learning methods in human action recognition. *IET Comput. Vis.* 11 (2017), 623–632. https://api.semanticscholar.org/CorpusID:29473256

[112] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. 2013. Learning Human Activities and Object Affordances from RGB-D Videos. *IJRR* (1 2013).

[113] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 706–715. https://api.semanticscholar.org/CorpusID:1026139

[114] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*. IEEE, 2556–2563.

[115] Rahul Kumar and Shailender Kumar. 2024. A survey on intelligent human action recognition techniques. *Multim. Tools Appl.* 83 (2024), 52653–52709. https://api.semanticscholar.org/CorpusID:269930971

[116] Zhenzhong Lan, Yi Zhu, Alexander G. Hauptmann, and S. Newsam. 2017. Deep Local Video Feature for Action Recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017), 1219–1225. https://api.semanticscholar.org/CorpusID:11599090

[117] Ivan Laptev. 2005. On space-time interest points. *International journal of computer vision* 64 (2005), 107–123.

[118] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning realistic human actions from movies. In *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.

[119] Sebastian Lapuschkin, Alexander Binder, Gregoire Montavon, Klaus-Robert Muller, and Wojciech Samek. 2016. Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[120] Dongho Lee, Jongseo Lee, and Jinwoo Choi. 2024. CAST: cross-attention in space and time for video action recognition. *Advances in Neural Information Processing Systems* 36 (2024).

[121] Chang Li, Yingchi Mao, Qian Huang, Xiaowei Zhu, and Jie Wu. 2023. Scale-Aware Graph Convolutional Network with Part-Level Refinement for Skeleton-Based Human Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[122] Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. 2017. Skeleton-based action recognition using LSTM and CNN. In *2017 IEEE International conference on multimedia & expo workshops (ICMEW)*. IEEE, 585–590.

[123] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. 2017. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. 597–600. https://doi.org/10.1109/ICMEW.2017.8026285

[124] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:246411402

[125] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. 2022. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676* (2022).

[126] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. 2023. Uniformerv2: Unlocking the potential of image vits for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1632–1643.

[127] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 2021. 3D Human Action Representation Learning via Cross-View Consistency Pursuit. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4741–4750.

[128] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2019. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3590–3598. https://doi.org/10.1109/CVPR.2019.00371

[129] Rongchang Li, Zhenhua Feng, Tianyang Xu, Linze Li, Xiao-Jun Wu, Muhammad Awais, Sara Atito, and Josef Kittler. 2025. C2c: Component-to-composition learning for zero-shot compositional action recognition. In *European Conference on Computer Vision*. Springer, 369–388.

[130] Sheng-Wei Li, Zi-Xiang Wei, Wei-Jie Chen, Yi-Hsin Yu, Chih-Yuan Yang, and Jane Yung-jen Hsu. 2025. Sa-dvae: Improving zero-shot skeleton-based action recognition by disentangled variational autoencoders. In *European Conference on Computer Vision*. Springer, 447–462.

[131] Tianhao Li and Limin Wang. 2020. Learning Spatiotemporal Features via Video and Text Pair Discrimination. *ArXiv* abs/2001.05691 (2020). https://api.semanticscholar.org/CorpusID:210698572

[132] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. 2010. Action recognition based on a bag of 3D points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 9–14. https://doi.org/10.1109/CVPRW.2010.5543273

[133] Yingwei Li, Yi Li, and Nuno Vasconcelos. 2018. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 513–528.

[134] Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7083–7093.

[135] Kun-Yu Lin, Henghui Ding, Jiaming Zhou, Yu-Ming Tang, Yi-Xing Peng, Zhilin Zhao, Chen Change Loy, and Wei-Shi Zheng. 2024. Rethinking clip-based video learners in cross-domain open-vocabulary action recognition. *arXiv preprint arXiv:2403.01560* (2024).

[136] Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, Yung-Huan Hsieh, and Hong-Ming Chen. 2012. Human action recognition and retrieval using sole depth information. In *Proceedings of the 20th ACM International Conference on Multimedia* (Nara, Japan) *(MM '12)*. Association for Computing Machinery, New York, NY, USA, 1053–1056. https://doi.org/10.1145/2393347.2396381

[137] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. 2017. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475* (2017).

[138] Hong Liu, Qinqin He, and Mengyuan Liu. 2017. Human action recognition using Adaptive Hierarchical Depth Motion Maps and Gabor filter. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1432–1436. https://doi.org/10.1109/ICASSP.2017.7952393

[139] Haowei Liu, Yongcheng Liu, Yuxin Chen, Chunfeng Yuan, Bing Li, and Weiming Hu. 2023. TranSkeleton: Hierarchical Spatial–Temporal Transformer for Skeleton-Based Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 8 (2023), 4137–4148. https://doi.org/10.1109/TCSVT.2023.3240472

[140] Hongda Liu, Yunfan Liu, Min Ren, Hao Wang, Yunlong Wang, and Zhenan Sun. 2024. Revealing Key Details to See Differences: A Novel Prototypical Perspective for Skeleton-based Action Recognition. *arXiv preprint arXiv:2411.18941* (2024).

[141] Jingen Liu, Jiebo Luo, and Mubarak Shah. 2009. Recognizing realistic actions from videos "in the wild". In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 1996–2003. https://doi.org/10.1109/CVPR.2009.5206744

[142] Jun Liu, Amir Shahroudy, Mauricio Perez, G. Wang, Ling yu Duan, and Alex Chichung Kot. 2019. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2019), 2684–2701. https://api.semanticscholar.org/CorpusID:152282878

[143] Jun Liu, Amir Shahroudy, Dong Xu, Alex C. Kot, and Gang Wang. 2018. Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 12 (2018), 3007–3021. https://doi.org/10.1109/TPAMI.2017.2771306

[144] Jiaying Liu, Sijie Song, Chunhui Liu, Yanghao Li, and Yueyu Hu. 2020. A Benchmark Dataset and Comparison Study for Multi-modal Human Action Analytics. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 2, Article 41 (May 2020), 24 pages. https://doi.org/10.1145/3365212

[145] Mengyuan Liu, Hong Liu, and Chen Chen. 2017. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition* 68 (2017), 346–362.

[146] Shenlan Liu, Xiang Liu, Gao Huang, Lin Feng, Lianyu Hu, Dong Jiang, Aibin Zhang, Yang Liu, and Hong Qiao. 2020. FSD-10: a dataset for competitive sports content analysis. *arXiv preprint arXiv:2002.03312* (2020).

[147] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. 2019. MeteorNet: Deep Learning on Dynamic 3D Point Cloud Sequences. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 9245–9254. https://api.semanticscholar.org/CorpusID:204800955

[148] Xingyu Liu, Sanping Zhou, Le Wang, and Gang Hua. 2023. Parallel attention interaction network for few-shot skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1379–1388.

[149] Yang Liu, Zhaoyang Lu, Jing Li, Tao Yang, and Chao Yao. 2018. Global temporal representation based cnns for infrared action recognition. *IEEE Signal Processing Letters* 25, 6 (2018), 848–852.

[150] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3202–3211.

[151] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 143–152.

[152] Xiang Long, Chuang Gan, Gerard De Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. 2018. Attention clusters: Purely attention based local feature integration for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7834–7843.

[153] Ning Ma, Hongyi Zhang, Xuhui Li, Sheng Zhou, Zhen Zhang, Jun Wen, Haifeng Li, Jingjun Gu, and Jiajun Bu. 2022. Learning spatial-preserved skeleton representations for few-shot action recognition. In *European Conference on Computer Vision*. Springer, 174–191.

[154] Mahshid Majd and Reza Safabakhsh. 2020. Correlational convolutional LSTM for human action recognition. *Neurocomputing* 396 (2020), 224–229.

[155] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. 2009. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2929–2936.

[156] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. 2019. Something-Else: Compositional Action Recognition With Spatial-Temporal Interaction Networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 1046–1056. https://api.semanticscholar.org/CorpusID:209439709

[157] Hocine Meglouli, Layachi Bentabet, Mohamed Airouche, et al. 2019. A new technique based on 3D convolutional neural networks and filtering optical flow maps for action classification in infrared video. *Journal of Control Engineering and Applied Informatics* 21, 4 (2019), 43–50.

[158] Vineet Mehta, Abhinav Dhall, Sujata Pal, and Shehroz S. Khan. 2021. Motion and Region Aware Adversarial Learning for Fall Detection with Thermal Imaging. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 6321–6328. https://doi.org/10.1109/ICPR48806.2021.9412632

[159] Ross Messing, Chris Pal, and Henry Kautz. 2009. Activity recognition using the velocity histories of tracked keypoints. In *2009 IEEE 12th International Conference on Computer Vision*. 104–111. https://doi.org/10.1109/ICCV.2009.5459154

[160] Ross Messing, Chris Pal, and Henry Kautz. 2009. Activity recognition using the velocity histories of tracked keypoints. In *2009 IEEE 12th international conference on computer vision*. IEEE, 104–111.

[161] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations From Uncurated Instructional Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[162] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 2630–2640. https://api.semanticscholar.org/CorpusID:182952863

[163] Krystian Mikolajczyk and Cordelia Schmid. 2005. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence* 27, 10 (2005), 1615–1630.

[164] Yuecong Min, Yanxiao Zhang, Xiujuan Chai, and Xilin Chen. 2020. An Efficient PointLSTM for Point Clouds Based Gesture Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5760–5769. https://doi.org/10.1109/CVPR42600.2020.00580

[165] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. 2019. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence* 42, 2 (2019), 502–508.

[166] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. 2020. Audio-Visual Instance Discrimination with Cross-Modal Agreement. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 12470–12481. https://api.semanticscholar.org/CorpusID:216553230

[167] Md Golam Morshed, Tangina Sultana, Aftab Alam, and Young-Koo Lee. 2023. Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities. *Sensors (Basel, Switzerland)* 23 (2023). https://api.semanticscholar.org/

CorpusID:256936214
[168] Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. 2007. Mocap database hdm05. *Institut für Informatik II, Universität Bonn* 2, 7 (2007).
[169] Woomin Myung, Nan Su, Jing-Hao Xue, and Guijin Wang. 2024. DeGCN: Deformable Graph Convolutional Networks for Skeleton-Based Action Recognition. *IEEE Transactions on Image Processing* 33 (2024), 2477–2490.
[170] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. 2021. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3163–3172.
[171] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. 2010. Modeling temporal structure of decomposable motion segments for activity classification. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11*. Springer, 392–405.
[172] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A review on the attention mechanism of deep learning. *Neurocomputing* 452 (2021), 48–62.
[173] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*. 53–60. https://doi.org/10.1109/WACV.2013.6474999
[174] Omar Oreifej and Zicheng Liu. 2013. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 716–723. https://doi.org/10.1109/CVPR.2013.98
[175] Preksha Pareek and Ankit Thakkar. 2020. A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review* 54 (2020), 2259–2322. https://api.semanticscholar.org/CorpusID:224901816
[176] Mandela Patrick, Yuki Asano, Polina Kuznetsova, Ruth Fong, Joao F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. 2021. Multi-modal Self-Supervision from Generalized Data Transformations. https://openreview.net/forum?id=mgVbI13p96
[177] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. 2021. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems* 34 (2021), 12493–12506.
[178] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. 2016. Bag of visual words and fusion methods for action recognition. *Comput. Vis. Image Underst.* 150, C (Sept. 2016), 109–125. https://doi.org/10.1016/j.cviu.2016.03.013
[179] James Philbin, Anna Bosch, Ondrej Chum, Jan-Mark Geusebroek, Josef Sivic, Andrew Zisserman, et al. 2006. Oxford TRECVID 2006-Notebook paper.. In *TRECVID*.
[180] A. J. Piergiovanni, Anelia Angelova, and Michael S. Ryoo. 2020. Evolving Losses for Unsupervised Video Representation Learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 130–139. https://api.semanticscholar.org/CorpusID:211532320
[181] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, H. Wang, Serge J. Belongie, and Yin Cui. 2020. Spatiotemporal Contrastive Video Representation Learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 6960–6970. https://api.semanticscholar.org/CorpusID:221090567
[182] Zhenyue Qin, Yang Liu, Pan Ji, Dongwoo Kim, Lei Wang, Bob McKay, Saeed Anwar, and Tom Gedeon. 2022. Fusing Higher-Order Features in Graph Neural Networks for Skeleton-based Action Recognition. *IEEE TNNLS* (2022).
[183] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*. 5533–5541.
[184] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:231591445
[185] H. Rahmani and Bennamoun. 2017. Learning Action Recognition Model from Depth and Skeleton Videos. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 5833–5842. https://api.semanticscholar.org/CorpusID:24999668
[186] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. 2016. Histogram of Oriented Principal Components for Cross-View Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 12 (2016), 2430–2443. https://doi.org/10.1109/TPAMI.2016.2533389
[187] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. 2014. HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*. Springer, 742–757.
[188] Hossein Rahmani and Ajmal Mian. 2016. 3D Action Recognition From Novel Viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[189] Sadegh Rahmaniboldaji, Filip Rybansky, Quoc Vuong, Frank Guerin, and Andrew Gilbert. 2024. DEAR: Depth-Enhanced Action Recognition. *arXiv preprint arXiv:2408.15679* (2024).
[190] Michalis Raptis and Stefano Soatto. 2010. Tracklet descriptors for action modeling and video analysis. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*. Springer, 577–590.
[191] Kishore K Reddy and Mubarak Shah. 2013. Recognizing 50 human action categories of web videos. *Machine vision and applications* 24, 5 (2013), 971–981.
[192] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. 2008. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8. https://doi.org/10.1109/CVPR.2008.4587727
[193] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012. A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 1194–1201. https://doi.org/10.1109/CVPR.2012.6247801
[194] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. 2016. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision* 119 (2016), 346–373.
[195] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115 (2014), 211 – 252. https://api.semanticscholar.org/CorpusID:2930547
[196] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. 2023. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*. PMLR, 29441–29454.
[197] Adrián Sánchez-Caballero, Sergio de López-Diz, David Fuentes-Jiménez, Cristina Losada-Gutiérrez, Marta Marrón-Romera, David Casillas-Pérez, and Mohammad Ibrahim Sarker. 2020. 3DFCNN: real-time action recognition using 3D deep neural networks with raw depth information. *Multimedia Tools and Applications* 81 (2020), 24119 – 24143. https://api.semanticscholar.org/CorpusID:219686945
[198] Adrián Sánchez-Caballero, David Fuentes-Jiménez, and Cristina Losada-Gutiérrez. 2020. Exploiting the ConvLSTM: Human Action Recognition using Raw Depth Video-Based Recurrent Neural Networks. *ArXiv* abs/2006.07744 (2020). https://api.semanticscholar.org/CorpusID:219687182
[199] Allah Bux Sargano, Plamen P. Angelov, and Zulfiqar Habib. 2017. A Comprehensive Review on Handcrafted and Learning-Based Action Representation Approaches for Human Activity Recognition. *Applied Sciences* 7 (2017), 110. https://api.semanticscholar.org/CorpusID:63948114
[200] Sandipan Sarma, Divyam Singal, and Arijit Sur. 2024. LoCATe-GAT: Modeling Multi-Scale Local Context and Action Relationships for Zero-Shot Action Recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2024), 1–13. https://doi.org/10.1109/TETCI.2024.3499995
[201] Christian Schuldt, Ivan Laptev, and Barbara Caputo. 2004. Recognizing human actions: a local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Vol. 3. IEEE, 32–36.
[202] Paul Scovanner, Saad Ali, and Mubarak Shah. 2007. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*. 357–360.
[203] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala. 2013. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. *CVPR Workshop* (2013).
[204] Anuj K Shah, Ripul Ghosh, and Aparna Akula. 2018. A spatio-temporal deep learning approach for human action recognition in infrared videos. In *Optics and Photonics for Information Processing XII*, Vol. 10751. SPIE, 249–257.
[205] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and G. Wang. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 1010–1019. https://api.semanticscholar.org/CorpusID:15928602
[206] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2616–2625.
[207] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. 2021. An image is worth 16x16 words, what is a video worth? *arXiv preprint arXiv:2103.13915* (2021).
[208] Zhiqiang Shen, Xiaoxiao Sheng, Hehe Fan, Longguang Wang, Yulan Guo, Qiong Liu, Hao Wen, and Xi Zhou. 2023. Masked spatio-temporal structure prediction for self-supervised learning on point cloud videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16580–16589.
[209] Xiaoxiao Sheng, Zhiqiang Shen, Gang Xiao, Longguang Wang, Yulan Guo, and Hehe Fan. 2023. Point contrastive prediction with semantic clustering for self-supervised learning on point cloud videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16515–16524.

[210] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Skeleton-Based Action Recognition With Directed Graph Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7904–7913. https://doi.org/10.1109/CVPR.2019.00810

[211] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12026–12035.

[212] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2020. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *Proceedings of the Asian conference on computer vision*.

[213] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. 2018. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626* (2018).

[214] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 510–526.

[215] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems* 27 (2014).

[216] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. 2021. Semi-Supervised Action Recognition With Temporal Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10389–10399.

[217] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems* 30 (2017).

[218] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[219] Yisheng Song, Zhang Zhang, Caifeng Shan, and Liang Wang. 2020. Stronger, Faster and More Explainable: A Graph Convolutional Baseline for Skeleton-based Action Recognition. *Proceedings of the 28th ACM International Conference on Multimedia* (2020). https://api.semanticscholar.org/CorpusID:222278406

[220] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. 2020. Richly activated graph convolutional network for robust skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 5 (2020), 1915–1925.

[221] K Soomro. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[222] Derya Soydaner. 2022. Attention mechanism in neural networks: where it comes and where it goes. *Neural Computing and Applications* 34 (2022), 13371 – 13385. https://api.semanticscholar.org/CorpusID:248427085

[223] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. 2015. Unsupervised learning of video representations using lstms. In *International conference on machine learning*. PMLR, 843–852.

[224] Ju Sun, Yadong Mu, Shuicheng Yan, and Loong-Fah Cheong. 2010. Activity recognition using dense long-duration trajectories. In *2010 IEEE International Conference on Multimedia and Expo*. 322–327. https://doi.org/10.1109/ICME.2010.5583046

[225] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. 2022. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence* 45, 3 (2022), 3200–3225.

[226] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1199–1208.

[227] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. 2011. Human activity detection from RGBD images. In *Workshops at the twenty-fifth AAAI conference on artificial intelligence*.

[228] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. 2021. VIMPAC: Video Pre-Training via Masked Token Prediction and Contrastive Learning. *ArXiv* abs/2106.11250 (2021). https://api.semanticscholar.org/CorpusID:235489838

[229] Yansong Tang, Yi Tian, Jiwen Lu, Peiyang Li, and Jie Zhou. 2018. Deep Progressive Reinforcement Learning for Skeleton-Based Action Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5323–5332. https://doi.org/10.1109/CVPR.2018.00558

[230] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.

[231] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[232] Du Tran and Alexander Sorokin. 2008. Human Activity Recognition with Metric Learning. In *European Conference on Computer Vision*. https://api.semanticscholar.org/CorpusID:14160859

[233] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. 2019. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5552–5561.

[234] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.

[235] Somasundaram Vadivelu, Sudakshin Ganesan, OV Ramana Murthy, and Abhinav Dhall. 2017. Thermal imaging based elderly fall detection. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*. Springer, 541–553.

[236] Raviteja Vemulapalli and Rama Chellappa. 2016. Rolling Rotations for Recognizing Human Actions from 3D Skeletal Data. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4471–4479. https://doi.org/10.1109/CVPR.2016.484

[237] Duc-Quang Vu, Ngan Le, and Jia-Ching Wang. 2024. Self-supervised learning via multi-transformation classification for action recognition. In *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 1–6.

[238] Duc-Quang Vu, Ngan T.H. Le, and Jia-Ching Wang. 2022. (2+1)D Distilled ShuffleNet: A Lightweight Unsupervised Distillation Network for Human Action Recognition. In *2022 26th International Conference on Pattern Recognition (ICPR)*. 3197–3203. https://doi.org/10.1109/ICPR56361.2022.9956634

[239] Guangming Wang, Hanwen Liu, Muyao Chen, Yehui Yang, Zhe Liu, and Hesheng Wang. 2020. Anchor-Based Spatio-Temporal Attention 3-D Convolutional Networks for Dynamic 3-D Point Cloud Sequences. *IEEE Transactions on Instrumentation and Measurement* 70 (2020), 1–11. https://api.semanticscholar.org/CorpusID:229340177

[240] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2013. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *International Journal of Computer Vision* 103 (2013), 60 – 79. https://api.semanticscholar.org/CorpusID:5401052

[241] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2011. Action recognition by dense trajectories. In *CVPR 2011*. 3169–3176. https://doi.org/10.1109/CVPR.2011.5995407

[242] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*. 3551–3558.

[243] Hongsong Wang and Liang Wang. 2017. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 499–508.

[244] Haiyan Wang, Liang Yang, Xuejian Rong, Jinglun Feng, and Yingli Tian. 2021. Self-supervised 4D Spatio-temporal Feature Learning via Order Prediction of Sequential Point Cloud Clips. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 3761–3770. https://doi.org/10.1109/WACV48630.2021.00381

[245] Junke Wang, Dongdong Chen, Chong Luo, Bo He, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2024. OmniViD: A Generative Framework for Universal Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18209–18220.

[246] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 1290–1297. https://doi.org/10.1109/CVPR.2012.6247813

[247] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. 2014. Learning Actionlet Ensemble for 3D Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 5 (2014), 914–927. https://doi.org/10.1109/TPAMI.2013.198

[248] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. 2014. Cross-View Action Modeling, Learning, and Recognition. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), 2649–2656. https://api.semanticscholar.org/CorpusID:2239612

[249] Lei Wang. 2017. *Analysis and Evaluation of Kinect-based Action Recognition Algorithms*. Master's thesis. School of the Computer Science and Software Engineering, The University of Western Australia.

[250] Lei Wang. 2023. *Robust human action modelling*. Ph. D. Dissertation. The Australian National University (Australia).

[251] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14549–14560.

[252] Lei Wang, Du Q. Huynh, and Piotr Koniusz. 2019. A Comparative Review of Recent Kinect-based Action Recognition Algorithms. *TIP* (2019). https://doi.org/10.1109/TIP.2019.2925285

[253] Lei Wang and Piotr Koniusz. 2021. *Self-Supervising Action Recognition by Statistical Moment and Subspace Descriptors*. Association for Computing Machinery,

New York, NY, USA, 4324–4333. https://doi.org/10.1145/3474085.3475572
[254] Lei Wang and Piotr Koniusz. 2022. Temporal-Viewpoint Transportation Plan for Skeletal Few-shot Action Recognition. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. 4176–4193.
[255] Lei Wang and Piotr Koniusz. 2022. Uncertainty-DTW for time series and sequences. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*. Springer, 176–195.
[256] Lei Wang and Piotr Koniusz. 2023. 3Mformer: Multi-Order Multi-Mode Transformer for Skeletal Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5620–5631.
[257] Lei Wang and Piotr Koniusz. 2024. Flow dynamics correction for action recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3795–3799.
[258] Lei Wang, Piotr Koniusz, and Du Q. Huynh. 2019. Hallucinating IDT Descriptors and I3D Optical Flow Features for Action Recognition With CNNs. In *ICCV*.
[259] Lei Wang, Jun Liu, and Piotr Koniusz. 2021. 3D Skeleton-based Few-shot Action Recognition with JEANIE is not so Naïve. *arXiv preprint arXiv:2112.12668* (2021).
[260] Limin Wang, Yu Qiao, and Xiaoou Tang. 2015. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4305–4314.
[261] Lei Wang, Ke Sun, and Piotr Koniusz. 2024. High-order tensor pooling with attention for action recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3885–3889.
[262] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. 2020. TDN: Temporal Difference Networks for Efficient Action Recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 1895–1904. https://api.semanticscholar.org/CorpusID:229331798
[263] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.
[264] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2018. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence* 41, 11 (2018), 2740–2755.
[265] Lei Wang, Xiuyuan Yuan, Tom Gedeon, and Liang Zheng. [n. d.]. Taylor Videos for Action Recognition. In *Forty-first International Conference on Machine Learning*.
[266] Pichao Wang, Wanqing Li, Zhimin Gao, Chang Tang, and Philip O. Ogunbona. 2018. Depth Pooling Based Large-Scale 3-D Action Recognition With Convolutional Neural Networks. *IEEE Transactions on Multimedia* 20, 5 (2018), 1051–1061. https://doi.org/10.1109/TMM.2018.2818329
[267] Pichao Wang, Wanqing Li, Zhimin Gao, Jing Zhang, Chang Tang, and Philip O. Ogunbona. 2016. Action Recognition From Depth Maps Using Deep Convolutional Neural Networks. *IEEE Transactions on Human-Machine Systems* 46, 4 (2016), 498–509. https://doi.org/10.1109/THMS.2015.2504550
[268] Pichao Wang, Shuang Wang, Zhimin Gao, Yonghong Hou, and Wanqing Li. 2017. Structured Images for RGB-D Action Recognition. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 1005–1014. https://doi.org/10.1109/ICCVW.2017.123
[269] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. 2021. BEVT: BERT Pretraining of Video Transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 14713–14723. https://api.semanticscholar.org/CorpusID:244799265
[270] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. 2023. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6312–6322.
[271] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12695–12705.
[272] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. 2024. Internvideo2: Scaling video foundation models for multimodal video understanding. *Arxiv e-prints* (2024), arXiv–2403.
[273] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. 2022. InternVideo: General Video Foundation Models via Generative and Discriminative Learning. *arXiv preprint arXiv:2212.03191* (2022).
[274] Yancheng Wang, Yang Xiao, Fu Xiong, Wenxiang Jiang, ZHIGUO CAO, Joey Tianyi Zhou, and Junsong Yuan. 2020. 3DV: 3D Dynamic Voxel for Action Recognition in Depth Video. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 508–517. https://api.semanticscholar.org/CorpusID:213194487
[275] Zihan Wang, Yang Yang, Zhi Liu, and Y. Zheng. 2023. Deep Neural Networks in Video Human Action Recognition: A Review. *ArXiv* abs/2305.15692 (2023). https://api.semanticscholar.org/CorpusID:258887932
[276] Yuyang Wanyan, Xiaoshan Yang, Weiming Dong, and Changsheng Xu. 2024. A Comprehensive Review of Few-shot Action Recognition. *ArXiv* abs/2407.14744 (2024). https://api.semanticscholar.org/CorpusID:271329302
[277] Chen Wei, Haoqi Fan, Saining Xie, Chaoxia Wu, Alan Loddon Yuille, and Christoph Feichtenhofer. 2021. Masked Feature Prediction for Self-Supervised Visual Pre-Training. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 14648–14658. https://api.semanticscholar.org/CorpusID:245218767
[278] Yimin Wei, Hao Liu, Tingting Xie, Qiuhong Ke, and Yulan Guo. 2021. Spatial-Temporal Transformer for 3D Point Cloud Sequences. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2021), 657–666. https://api.semanticscholar.org/CorpusID:239024595
[279] Daniel Weinland, Remi Ronfard, and Edmond Boyer. 2006. Free viewpoint action recognition using motion history volumes. *Computer vision and image understanding* 104, 2-3 (2006), 249–257.
[280] Wanjiang Weng, Hongsong Wang, Junbo He, Lei He, and Guosen Xie. 2024. USDRL: Unified Skeleton-Based Dense Representation Learning with Multi-Grained Feature Decorrelation. *arXiv preprint arXiv:2412.09220* (2024).
[281] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. 2008. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II 10*. Springer, 650–663.
[282] Thomas Wolf. 2020. Transformers: State-of-the-Art Natural Language Processing. *arXiv preprint arXiv:1910.03771* (2020).
[283] Shu-Fai Wong and Roberto Cipolla. 2007. Extracting Spatiotemporal Interest Points using Global Information. In *2007 IEEE 11th International Conference on Computer Vision*. 1–8. https://doi.org/10.1109/ICCV.2007.4408923
[284] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. 2012. View invariant human action recognition using histograms of 3D joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 20–27. https://doi.org/10.1109/CVPRW.2012.6239233
[285] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. 2022. Learning From Temporal Gradient for Semi-Supervised Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3252–3262.
[286] Yang Xiao, Jun Chen, ZHIGUO CAO, Joey Tianyi Zhou, and Xiang Bai. 2018. Action Recognition for Depth Video using Multi-view Dynamic Images. *Inf. Sci.* 480 (2018), 287–304. https://api.semanticscholar.org/CorpusID:49552527
[287] Chen Xiaokai, Gao Ke, and Cao Juan. 2019. Predictability Analyzing: Deep Reinforcement Learning for Early Action Recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. 958–963. https://doi.org/10.1109/ICME.2019.00169
[288] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*. 305–321.
[289] Jiazheng Xing, Mengmeng Wang, Yudi Ruan, Bofan Chen, Yaowei Guo, Boyu Mu, Guang Dai, Jingdong Wang, and Yong Liu. 2023. Boosting few-shot action recognition with graph-guided hybrid matching. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1740–1750.
[290] Bo Xiong, Haoqi Fan, Kristen Grauman, and Christoph Feichtenhofer. 2021. Multiview Pseudo-Labeling for Semi-Supervised Learning From Video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7209–7219.
[291] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. 2019. Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10326–10335. https://doi.org/10.1109/CVPR.2019.01058
[292] Jing Xu, Anqi Zhu, Jingyu Lin, Qiuhong Ke, and Cunjian Chen. 2024. Action-OOD: An End-to-End Skeleton-Based Model for Robust Out-of-Distribution Human Action Detection. *arXiv preprint arXiv:2405.20633* (2024).
[293] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. 2022. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3333–3343.
[294] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *AAAI*.
[295] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 10714–10726. https://api.semanticscholar.org/CorpusID:257232853
[296] Ceyuan Yang, Yinghao Xu, Bo Dai, and Bolei Zhou. 2020. Video Representation Learning with Visual Tempo Consistency. *ArXiv* abs/2006.15489 (2020). https://api.semanticscholar.org/CorpusID:220250229

[297] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. 2020. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 591–600.

[298] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. 2021. Unik: A unified framework for real-world skeleton-based action recognition. *arXiv preprint arXiv:2107.08580* (2021).

[299] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. 2024. View-Invariant Skeleton Action Representation Learning via Motion Retargeting. *Int. J. Comput. Vision* 132, 7 (Jan. 2024), 2351–2366. https://doi.org/10.1007/s11263-023-01967-8

[300] Hao Yang, Dan Yan, Li Zhang, Yunda Sun, Dong Li, and Stephen J. Maybank. 2022. Feedback Graph Convolutional Network for Skeleton-Based Action Recognition. *IEEE Transactions on Image Processing* 31 (2022), 164–175. https://doi.org/10.1109/TIP.2021.3129117

[301] Xitong Yang, Haoqi Fan, Lorenzo Torresani, Larry S. Davis, and Heng Wang. 2021. Beyond Short Clips: End-to-End Video-Level Learning With Collaborative Memories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7567–7576.

[302] Yuheng Yang. 2024. Skeleton-based Action Recognition with Non-linear Dependency Modeling and Hilbert-Schmidt Independence Criterion. arXiv:2412.18780 [cs.CV] https://arxiv.org/abs/2412.18780

[303] Yuheng Yang, Haipeng Chen, Zhenguang Liu, Yingda Lyu, Beibei Zhang, Shuang Wu, Zhibo Wang, and Kui Ren. 2023. Action recognition with multi-stream motion modeling and mutual information maximization. *arXiv preprint arXiv:2306.07576* (2023).

[304] Guangle Yao, Tao Lei, and Jiandan Zhong. 2019. A review of Convolutional-Neural-Network-based action recognition. *Pattern Recognit. Lett.* 118 (2019), 14–22. https://api.semanticscholar.org/CorpusID:85445143

[305] Huanjin Yao, Wenhao Wu, and Zhiheng Li. 2023. Side4video: Spatial-temporal side network for memory-efficient image-to-video transfer learning. *arXiv preprint arXiv:2311.15769* (2023).

[306] Leiyue Yao, Wei Yang, and Wei Huang. 2020. A data augmentation method for human action recognition using dense joint motion images. *Appl. Soft Comput.* 97 (2020), 106713. https://api.semanticscholar.org/CorpusID:225238356

[307] Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. 2020. SeCo: Exploring Sequence Supervision for Unsupervised Representation Learning. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:220935968

[308] Lahav Yeffet and Lior Wolf. 2009. Local trinary patterns for human action recognition. In *2009 IEEE 12th international conference on computer vision*. IEEE, 492–497.

[309] Tianwei Yu, Peng Chen, Yuanjie Dang, Ruohong Huan, and Ronghua Liang. 2023. Multi-Speed Global Contextual Subspace Matching for Few-Shot Action Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 2344–2352. https://doi.org/10.1145/3581783.3612380

[310] Yating Yu, Congqi Cao, Yueran Zhang, Qinyi Lv, Lingtong Min, and Yanning Zhang. 2024. Building a Multi-modal Spatiotemporal Expert for Zero-shot Action Recognition with CLIP. *arXiv preprint arXiv:2412.09895* (2024).

[311] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4694–4702.

[312] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:259075356

[313] Jiahang Zhang, Lilang Lin, and Jiaying Liu. 2023. Prompted contrast with masked motion modeling: Towards versatile 3d action representation learning. In *Proceedings of the 31st ACM International Conference on Multimedia*. 7175–7183.

[314] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. 2017. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE international conference on computer vision*. 2117–2126.

[315] Shiwen Zhang. 2022. Tfcnet: Temporal fully connected networks for static unbiased temporal reasoning. *arXiv preprint arXiv:2203.05928* (2022).

[316] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. 2013. From Actemes to Action: A Strongly-Supervised Representation for Detailed Action Understanding. In *2013 IEEE International Conference on Computer Vision*. 2248–2255. https://doi.org/10.1109/ICCV.2013.280

[317] Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems* 31 (2018).

[318] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. 2019. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8668–8678.

[319] Aihua Zhou, Yujun Ma, Wanting Ji, Ming Zong, Pei Yang, Min Wu, and Mingzhe Liu. 2022. Multi-head attention-based two-stream EfficientNet for action recognition. *Multimedia Syst.* 29, 2 (June 2022), 487–498. https://doi.org/10.1007/s00530-022-00961-3

[320] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. 2018. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*. 803–818.

[321] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. 2023. Learning discriminative representations for skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10608–10617.

[322] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. 2024. BlockGCN: Redefine Topology Awareness for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2049–2058.

[323] Fan Zhu, Ling Shao, J. Xie, and Yi Fang. 2016. From handcrafted to learned representations for human action recognition: A survey. *Image Vis. Comput.* 55 (2016), 42–52. https://api.semanticscholar.org/CorpusID:46817082

[324] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. [n. d.]. Advancing Video Anomaly Detection: A Concise Review and a New Dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[325] Ming Zong, Ruili Wang, Xiubo Chen, Zhe Chen, and Yuanhao Gong. 2021. Motion saliency based multi-stream multiplier ResNets for action recognition. *Image Vis. Comput.* 107 (2021), 104108. https://api.semanticscholar.org/CorpusID:233261013