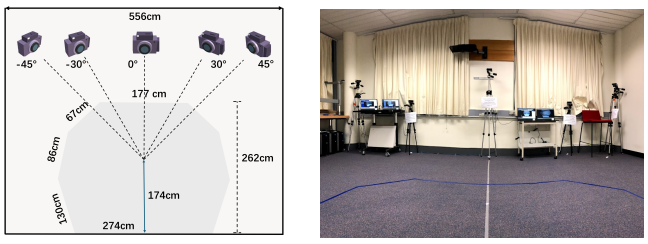
Table A.1: Comprehensive Overview of Skeleton-based Action Recognition Datasets. ANUBIS addresses critical gaps through rear-view coverage, aggressive/security actions, and enhanced 32-joint representation with Azure Kinect.

Dataset	Classes	Views	Subjects	Clips	Sensors	Additional Modalities	Dataset Type
HDM05 (2007)[17]	130	1	5	2,337	-	RGB	Human motion capture
MSRAction3D (2010)[12]	20	1	10	567	Kinect	RGB, Depth	Daily activities
CAD-60 (2011)[24]	12	-	4	68	Kinect	RGB, Depth	Human performing activities
MSRDailyActivity3D (2012)[25]	16	1	10	320	Kinect	RGB, Depth	Daily activities
G3D-Gaming (2012)[2]	20	1	10	-	Kinect	RGB, Depth	Gaming gestures
UTKinect (2012)[28]	10	-	10	200	Kinect	RGB, Depth	Human actions
SBU (2012)[31]	8	-	7	282	Kinect	RGB	Human-human interaction
CAD-120 (2013)[10]	10	-	4	120	Kinect	RGB, Depth	Activity types & object interactions
Berkeley MHAD (2013)[18]	11	4	12	660	Kinect	RGB, Depth, Audio, Accelerometer	Multimodal Capture & Controllable & Synced Data
Florence3D-Action (2013)[22]	9	-	10	215	Kinect	RGB, Depth	Daily Activities
MSRActionPairs3D (2013)[19]	12	-	10	360	Kinect	RGB, Depth	3D Action & Gesture Recognition
UCFKinect (2013)[4]	16	-	16	1,280	-	RGB, Depth	General actions
Northwestern-UCLA (2014)[26]	10	3	10	1,494	Kinect	RGB, Depth	Daily Activities
Multi-View TJU (2014)[13]	20	2	22	7,040	-	RGB, Depth	Multi-view actions
UWA3D Multiview Activity (2014)[21]	30	4	10	701	Kinect	RGB, Depth	Multi-view actions
SYSU 3D HOI (2015)[5]	12	-	40	480	Kinect	RGB, Depth	Human-object interaction
UWA3D Multiview Activity II (2015)[20]	30	4	10	1,070	Kinect	RGB, Depth	Daily activities
NTU-60 (2016)[23]	60	80	40	56,880	Kinect v2	RGB, Depth, Infrared	Large-scale general actions
PKU-MMD I (2017)[14]	51	3	66	1,076	Kinect v2	RGB, Depth, Infrared	Multi-modal actions
Kinetics-skeleton (2018)[29]	400	-	-	260,232	-	-	Based on publicly available RGB videos
RGB-D Varying-View (2018)[8]	40	9	118	25,600	Kinect v2	RGB, Depth	Multi-view actions
NTU-120 (2019)[16]	120	155	106	114,480	Kinect v2	RGB, Depth, Infrared	Large-scale general actions
MMAct (2019)[9]	37	5	20	36,764	-	RGB, Accelerometer, Gyroscope	Multi-modal actions
PKU-MMD II (2020)[15]	41	3	13	1,009	Kinect v2	RGB, Depth, Infrared	Multi-modal actions
ETRI-Activity3D (2020)[7]	55	-	100	112,620	Kinect v2	RGB	Daily activities of the elderly
IKEA ASM (2020)[1]	33	3	48	16,764	Kinect v2	RGB, Depth	Furniture assembly
UAV-Human (2021)[11]	155	-	119	22,476	Azure Kinect	RGB, Infrared, Depth	UAV perspective actions
NCRC (2022)[6]	6	-	8	398	-	-	Nursing care activities
Tai-Chi (2022)[30]	10	-	-	200	Perception Neuron	-	Martial arts
ANUBIS (2025)	102	80	80	66,232	Azure Kinect	RGB, Depth	Large-Scale & Multi-Person & Frontal / Rear-View & In-the-Wild

Appendix A. Comparison with Existing Datasets

Table A.1 presents a comparative summary of our dataset against existing benchmarks. Existing datasets exhibit diverse characteristics across devices, modalities, and applications. NTU-60 [23] establishes a multi-view benchmark with 60 indoor actions, while NTU-120 [16] extends to 120 classes. Kinetics-skeleton [29] extracts skeletal information from large-scale RGB videos. For multimodal fusion, PKU-MMD I [14] integrates RGB, depth, and infrared data, while PKU-MMD II [15] adds fine-grained interaction annotations. MMAct [9] combines visual and inertial sensor data for mobile scenarios, and RGB-D Varying-View [8] explores viewpoint robustness through dynamic perspectives.In vertical applications, ETRI-Activity 3D [7] focuses on elderly monitoring, IKEA ASM [1] targets furniture assembly tasks, NCRC-Human [6] analyzes nursing scenarios, and Tai-Chi [30] characterizes Tai Chi kinematics. UAV-Human [11] enriches spatial dimensions through drone perspectives. These datasets serve different objectives including general validation, scenario-specific optimization, and cross-modal learning, forming a diversified research ecosystem.

Appendix B. Dataset Collection



(a) Venue layout. (b) Camera arrangement.

Figure B.1: ANUBIS dataset collection setup overview.

ANUBIS comprises 102 carefully selected actions including both individual behaviors (e.g., drinking water, waving) and multi-person interactions (e.g., handshaking, object exchange, stabbing). The 102 actions are distributed across 40 collection sessions, with each session involving two participants as a group and lasting approximately 1.5 hours. Every 10 sessions incorporates 10-minute breaks to maintain performance quality, and sessions exhibiting substandard action execution are re-recorded to ensure data integrity. The dataset comprises 40 participant groups totaling 80 participants and approximately 60 hours of multi-modal recordings (see Fig. 1 in the main paper).

Multi-view camera setups. The acquisition system uses five Microsoft Azure Kinect devices arranged in symmetric horizontal configuration at 0°, ±30°, and ±45° angles within a standardized 556cm×274cm indoor environment, as illustrated in Fig. B.1. While cameras maintain horizontal symmetry, each device operates at different heights and poses to enhance viewpoint diversity. Camera heights and poses are randomly adjusted every 10 groups to capture more diverse viewpoints.

Participants move freely within marked activity areas and perform each action four times per session to create different viewpoints: facing the cameras, facing away from cameras, switching positions while facing cameras, and switching positions while facing away. This collection protocol results in 20 different camera views for each participant pair performing the same action, ensuring comprehensive coverage from multiple angles, especially challenging rear views that are often missing in existing datasets. For interactive actions, participants also switch their active and passive roles when changing positions to capture both sides of the interaction. To ensure realistic performances while maintaining safety, we use appropriate items for different action types: toy weapons for simulated violence, wigs for hair-pulling actions, tissues for mouth-covering gestures, and soft objects like paper boxes for hitting actions to prevent injury.

Data preprocessing. We developed custom software to man-

age data collection across all five synchronized Azure Kinect cameras. The software records the exact start and end time of each action, ensuring all cameras capture the same actions simultaneously. During data processing, we use these recorded timestamps to extract individual action clips from the complete recordings of each group. Each clip contains three types of data: RGB, depth, and 3D skeleton videos, as shown in Fig. 1 in the main paper. For actions that are naturally short, we extend them to the standard 300-frame length by repeating the action frames.

Dataset statistics. ANUBIS comprises 102 action categories collected from 80 participants, generating 66,232 skeleton clips across 80 viewpoints, as presented in Tab.A.1. The viewpoint distribution includes 40 frontal views and 40 rear views from different angles, ensuring balanced coverage between frontal perspectives and challenging posterior orientations. Based on action categories, the dataset contains 45 independent actions (single-person behaviors) and 57 multi-person interactions. Among multi-person actions, we include 17 social interaction behaviors (e.g., handshaking, patting shoulders, object exchange), and 40 aggressive actions (e.g., hitting, stabbing, strangling). The complete statistics of ANUBIS are shown in Fig.B.2.

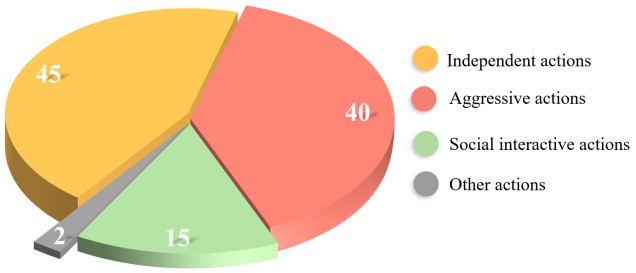


Figure B.2: Distribution of 102 human actions classified into four categories. The pie chart shows: independent actions (45, 44.1%), aggressive actions (40, 39.2%), social interactive actions (15, 14.7%), and other actions (2, 2.0%). Other actions specifically refer to spatial position change behaviors, including walk apart and walk from apart to together.

Appendix C. Experimental Setups

We benchmarked a range of state-of-the-art skeleton-based action recognition methods on our newly collected ANUBIS dataset and evaluated these methods on the NTU datasets for comparative analysis. All experiments were implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU for 50 epochs. Stochastic Gradient Descent (SGD) with momentum 0.9 was used as the optimizer, with an initial learning rate of 0.05, decayed to 10% at epoch 30.

Skeleton data was preprocessed via normalization and translation. All video clips were standardized to 300 frames using action repetition (except SkateFormer [3], which retained its original 64-frame window). Evaluation metrics included Top-1 and Top-5 classification accuracy, as well as model complexity indicators. To show the recognition accuracies of a model for all the action classes, a confusion matrix is used [27].

Appendix D. Additional Analysis and Discussion

Per-class performance analysis: easy vs. hard actions. This section presents a detailed breakdown of action recognition performance across the 102 action categories in ANUBIS, revealing clear patterns in what makes certain actions easy or challenging for current skeleton-based models.

Bone feature integration analysis. This section examines which specific actions benefit from adding bone vector features to joint coordinates, revealing the selective nature of anatomical structure information. The detailed results are presented in Table D.2.

Motion feature integration analysis. This section analyzes the impact of adding motion features (velocity/acceleration) to skeleton representations, revealing highly variable and model-dependent effects. Table D.3 presents the detailed analysis.

Negative impact analysis: when additional features hurt performance. This section identifies actions where adding bone or motion features consistently degrades performance across all models, highlighting potential pitfalls in naive feature fusion. The comprehensive results are shown in Table D.4.

Feature stream effect visualization. This section provides a visual analysis of how different feature combinations affect action recognition performance, illustrating the complex interplay between joint, bone, and motion representations. The results are visualized in Figure D.3.

The visualization reveals several key patterns:

i. Feature interaction effects: Some actions benefit from bone features but are harmed by motion (e.g., Walk Apart), while others show the opposite pattern (e.g., Apply Cream).

ii. Non-additive fusion: The best performance often comes from selective feature combinations rather than using all available features. For instance, "Walk Together" performs best with Joint+Bone but degrades significantly when motion is added.

iii. Action-specific optimization: Different actions require different feature strategies, suggesting the need for adaptive or action-aware fusion mechanisms rather than universal multi-modal approaches.

iv. Complementary vs. competing features: While some feature combinations are complementary (Joint+Bone for spatial actions), others compete or introduce noise (Motion for stable pose actions like Surrender).

References

[1] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. *WACV*, pages 847–859, 2021.

[2] Victoria Bloom, Dimitrios Makris, and Vasileios Argyriou. G3d: A gaming action dataset and real time action recognition evaluation framework. *CVPRW*, pages 7–12, 2012.

Table D.2: Actions with consistent accuracy gains across all three models (LA-GCN, STTFormer, DeGCN) when adding the Bone feature type. Only 7 out of 102 classes show universal benefit, indicating that bone vectors selectively help actions where limb geometry and joint relationships are key (e.g., arm-hand positioning).

Rank	Action Name	LA-GCN			STTFormer			DeGCN			Avg. Improve
		Joint	Joint+Bone	Improve	Joint	Joint+Bone	Improve	Joint	Joint+Bone	Improve	
1	sneeze	0.5310	0.5693	+0.0383	0.3628	0.5192	+0.1564	0.4779	0.5310	+0.0531	**+0.0826**
2	thumb up	0.4074	0.4175	+0.0101	0.2828	0.4141	+0.1313	0.2525	0.3468	+0.0943	**+0.0786**
3	follow person	0.8395	0.8696	+0.0301	0.7559	0.8462	+0.0903	0.7057	0.7559	+0.0502	**+0.0569**
4	running	0.7883	0.7948	+0.0065	0.7362	0.8241	+0.0879	0.7557	0.8013	+0.0456	**+0.0467**
5	pull collar	0.4202	0.4448	+0.0246	0.3834	0.4141	+0.0307	0.3006	0.3804	+0.0798	**+0.0450**
6	self-cutting with knife	0.3323	0.3867	+0.0544	0.3021	0.3263	+0.0242	0.2779	0.2870	+0.0091	**+0.0292**
7	punch to face	0.4787	0.5488	+0.0701	0.4909	0.4939	+0.0030	0.3994	0.4116	+0.0122	**+0.0284**

Table D.3: Actions improved in at least two of three models when adding the Motion feature type. No action achieved consistent gains across all models, underscoring the model-dependent and unstable nature of motion integration, beneficial for certain temporally distinctive actions but harmful in others. Action abbreviations: "support old people walking" refers to "support with arms for old people walking".

Rank	Action Name	LA-GCN			STTFormer			DeGCN			Avg. Improve
		Joint	Joint+Motion	Improve	Joint	Joint+Motion	Improve	Joint	Joint+Motion	Improve	
1	stand up	0.7964	0.7725	-0.0239	0.4820	0.6617	+0.1797	0.5719	0.7605	+0.1886	**+0.1148**
2	play magic cube	0.3237	0.2596	-0.0641	0.2276	0.3558	+0.1282	0.1282	0.3750	+0.2468	**+0.1036**
3	take off a hat	0.5761	0.7104	+0.1343	0.5493	0.5284	-0.0209	0.3851	0.4567	+0.0716	**+0.0617**
4	play a phone	0.4149	0.5403	+0.1254	0.4239	0.3403	-0.0836	0.2657	0.3791	+0.1134	**+0.0517**
5	cutting paper	0.2038	0.3726	+0.1688	0.3312	0.3121	-0.0191	0.2261	0.2261	+0.0000	**+0.0499**
6	running	0.7883	0.8241	+0.0358	0.7362	0.8143	+0.0781	0.7557	0.7362	-0.0195	**+0.0315**
7	support old people walking	0.7800	0.6967	-0.0833	0.6500	0.7067	+0.0567	0.5700	0.6700	+0.1000	**+0.0245**
8	squat down	0.8155	0.8452	+0.0297	0.8423	0.7887	-0.0536	0.7113	0.8065	+0.0952	**+0.0238**
9	jump up	0.7545	0.7455	-0.0090	0.6108	0.7156	+0.1048	0.7126	0.6826	-0.0300	**+0.0219**
10	pull collar	0.4202	0.3957	-0.0245	0.3834	0.4325	+0.0491	0.3006	0.3712	+0.0706	**+0.0317**

Table D.4: Actions with consistent accuracy drops across all three models when adding either Bone or Motion features (worst-affected feature type reported). All top declines are linked to Motion, with some drops exceeding 40%, highlighting the risk of unfiltered motion cues overwhelming stable joint-based representations. Action abbreviations: "walk apart together" refers to "walk form apart to together" and "throw object to person" refers to "pick and throw an object to person".

Rank	Action Name	Feature	LA-GCN			STTFormer			DeGCN			Avg. Decline
			Joint	Added feature	Decline	Joint	Added feature	Decline	Joint	Added feature	Decline	
1	walk apart together	Motion	0.9497	0.8365	-0.1132	0.9340	0.2799	-0.6541	0.9057	0.3491	-0.5566	**-0.4413**
2	walk apart	Motion	0.9579	0.3042	-0.6537	0.9709	0.8123	-0.1586	0.8382	0.5761	-0.2621	**-0.3581**
3	surrender	Motion	0.7212	0.6154	-0.1058	0.7308	0.4872	-0.2436	0.7372	0.4423	-0.2949	**-0.2148**
4	bite person	Motion	0.6592	0.5732	-0.0860	0.6369	0.3949	-0.2420	0.6561	0.3631	-0.2930	**-0.2070**
5	fist bumping	Motion	0.8190	0.6499	-0.1691	0.7774	0.5905	-0.1869	0.5816	0.3917	-0.1899	**-0.1820**
6	back pain	Motion	0.5131	0.4739	-0.0392	0.6176	0.4216	-0.1960	0.5817	0.2745	-0.3072	**-0.1808**
7	throw object to person	Motion	0.6730	0.5143	-0.1587	0.7143	0.5143	-0.2000	0.5556	0.3778	-0.1778	**-0.1788**
8	open bottle	Motion	0.3735	0.2018	-0.1717	0.2289	0.1506	-0.0783	0.3916	0.1325	-0.2591	**-0.1697**
9	thumb down	Motion	0.5623	0.4815	-0.0808	0.6364	0.3939	-0.2425	0.5320	0.3906	-0.1414	**-0.1549**
10	strangling neck	Motion	0.5666	0.3746	-0.1920	0.4799	0.3715	-0.1084	0.4458	0.2817	-0.1641	**-0.1548**

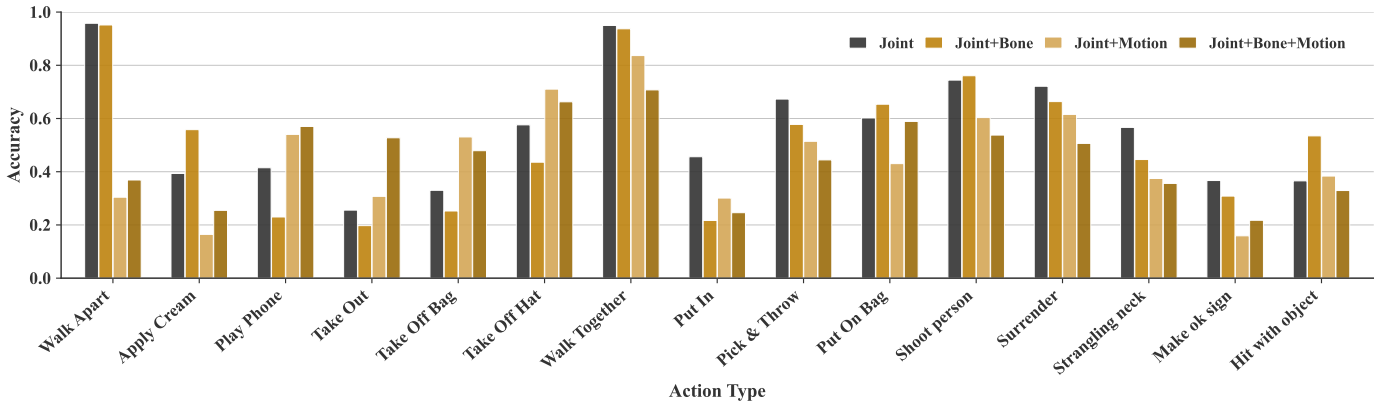


Figure D.3: Analysis of Joint, Bone, Motion data stream effects on action recognition using LA-GCN. Results from 15 actions with significant recognition accuracy fluctuations are shown. Action abbreviations: Apply Cream (apply cream on hand), Take Out (take object out of bag), Walk Together (walk form apart to together), Put In (put object into bag), Pick & Throw (pick and throw an object to person), Support Walk (support with arms for old people walking).

[3] Jeonghyeok Do and Munchurl Kim. Skateformer: skeletal-temporal transformer for human action recognition. *ECCV*, pages 401–420, 2025.

[4] Chris Ellis, Syed Zain Masood, Marshall F Tappen, Joseph J Laviola Jr, and Rahul Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *IJCV*, 101(3):420–436, 2013.

[5] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. *CVPR*, pages 5344–5352, 2015.

[6] Momal Ijaz, Renato Diaz, and Chen Chen. Multimodal transformer for nursing activity recognition. *CVPR*, pages 2065–2074, 2022.

[7] Jinhyeok Jang, Dohyung Kim, Cheonshu Park, Minsu Jang, Jaeyeon Lee, and Jaehong Kim. Etri-activity3d: A large-scale rgb-d dataset for robots to recognize daily activities of the elderly. *IROS*, pages 10990–10997, 2020.

[8] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. *ACM MM*, pages 1510–1518, 2018.

[9] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. *ICCV*, pages 8658–8667, 2019.

[10] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8):951–970, 2013.

[11] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. *CVPR*, pages 16266–16275, 2021.

[12] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. *CVPRW*, pages 9–14, 2010.

[13] An-An Liu, Yu-Ting Su, Ping-Ping Jia, Zan Gao, Tong Hao, and Zhao-Xuan Yang. Multiple/single-view human action recognition via part-induced multitask structural learning. *transactions on cybernetics*, 45(6):1194–1208, 2014.

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

[15] Jiaying Liu, Sijie Song, Chunhui Liu, Yanghao Li, and Yueyu Hu. A benchmark dataset and comparison study for multi-modal human action analytics. *TOMM*, 16(2):1–24, 2020.

[16] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *TPAMI*, 42(10):2684–2701, 2019.

[17] Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. Mocap database hdm05. *Institut für Informatik II, Universität Bonn*, 2(7), 2007.

[18] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. *WACV*, pages 53–60, 2013.

[19] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. *CVPR*, pages 716–723, 2013.

[20] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. *TPAMI*, 38(12):2430–2443, 2016.

[21] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. *ECCV*, pages 742–757, 2014.

[22] Lorenzo Seidenari, Vincenzo Varano, Stefano Berretti, Alberto Bimbo, and Pietro Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. *CVPRW*, pages 479–485, 2013.

[23] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. *CVPR*, pages 1010–1019, 2016.

[24] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Human activity detection from rgbd images. *plan, activity, and intent recognition*, 64, 2011.

[25] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. *CVPR*, pages 1290–1297, 2012.

[26] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. *CVPR*, pages 2649–2656, 2014.

[27] Lei Wang, Du Q. Huynh, and Piotr Koniusz. A comparative review of recent kinect-based action recognition algorithms. *TIP*, 29(1):15–28, 2019.

[28] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. View invariant human action recognition using histograms of 3d joints. *CVPRW*, pages 20–27, 2012.

[29] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI*, 2018.

[30] Lin Yuan, Zhen He, Qiang Wang, Leiyang Xu, and Xiang Ma. Spatial transformer network with transfer learning for small-scale fine-grained skeleton-based tai chi action recognition. *IECON*, pages 1–6, 2022.

[31] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. *CVPRW*, pages 28–35, 2012.