Mitigating Context Bias in Action Recognition via Skeleton-Dominated Two-Stream Network

Qiankun Li
HFIPS, Chinese Academy of Sciences
Department of Automation,
University of Science and Technology of China
Hefei, China
qklee@mail.ustc.edu.cn

Xiaolong Huang
School of Artificial Intelligence,
Chongqing University of Technology
Chongqing, China
hirox827@gmail.com

Yuwen Luo
Department of Automation,
University of Science and Technology of China
Hefei, China
liudebin98@mail.ustc.edu.cn

Xiaoyu Hu
HFIPS, Chinese Academy of Sciences
University of Science and Technology of China
Hefei, China
xyhoo@mail.ustc.edu.cn

Xinyu Sun
Hangzhou Hikvision Digital Technology Co., Ltd.
Hangzhou, China
843646876@qq.com

Zengfu Wang*
HFIPS, Chinese Academy of Sciences
Department of Automation,
University of Science and Technology of China
Hefei, China
zfwang@ustc.edu.cn

ABSTRACT

In the realm of intelligent manufacturing and industrial upgrading, sophisticated multimedia computing technologies play a pivotal role in the recognition of video actions. However, most studies suffer from the issue of background bias, where the models excessively focus on the contextual information within the videos rather than concentrating on comprehending the human actions themselves. This could potentially lead to severe misjudgments in industrial applications. In this paper, we propose a Skeleton-Dominated Two-Stream Network (SDTSN), which is a novel two-stream framework that fuses and ensembles the skeleton and RGB modalities for video action recognition. Experimental results on the Mimetics dataset, without any background bias, demonstrate the efficacy of our approach.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

Context bias, Skeleton-dominated, Two-stream, Video action recognition

*Corresponding author.

AMC-SME '23, November 2, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0273-0/23/11…$15.00
https://doi.org/10.1145/3606042.3616458

ACM Reference Format:
Qiankun Li, Xiaolong Huang, Yuwen Luo, Xiaoyu Hu, Xinyu Sun, and Zengfu Wang. 2023. Mitigating Context Bias in Action Recognition via Skeleton-Dominated Two-Stream Network. In *Proceedings of the 2023 Workshop on Advanced Multimedia Computing for Smart Manufacturing and Engineering (AMC-SME '23), November 2, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3606042.3616458

1 INTRODUCTION

With the development of advanced multimedia computing, video action recognition based on deep learning has achieved exceptional performance on existing benchmarks [3, 9, 12, 23], and has gradually been applied to domains such as smart manufacturing and engineering. For instance, it has been utilized for suspicious behavior recognition in the field of security [18, 24], driving action recognition in the realm of transportation [10, 11], and action analysis in the sports domain [17, 19, 27].

Most high-accuracy methods for action recognition rely on RGB-based models, which intuitively develop a 3D Convolution-based [3, 6, 7] or a Vision Transformer (ViT) based backbone [1, 5, 8, 15] to extract the features of the input RGB video data, embed them into the feature space and subsequently employ a classification head to recognize the categories of actions. However, RGB-based methods often infer action categories relying on scene cues rather than focusing on the actions themselves [26]. This can lead to errors, such as actions occurring on a lush green field that tend to be incorrectly assumed to involve playing football, even when that is not true. This issue is called context bias, which poses potential risks in industrial applications.

Skeleton-based video action recognition methods [4, 13, 16, 28] involve using sequences of skeleton graphs as input for the model. These approaches eliminate environmental noise, such as background, light, and objects, allowing action recognition algorithms to concentrate on the essential features of the action. However, these methods have limitations in terms of accuracy because they

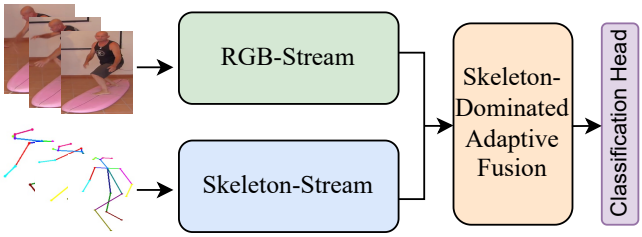

Figure 1: Overview of the SDTSN. SDTSN combines the skeleton and RGB modalities in the two-stream network and uses the skeleton as dominant to fusion, thereby mitigating context bias in mitigating video action recognition.

lose low-level information (e.g., pixel-level details) during the extraction of the skeleton.

To simultaneously consider the detailed information of RGB and the representation of the action itself, two-stream networks have been proposed [13, 20, 22, 25]. These methods [22, 25] mostly utilize two backbones to extract semantic information from RGB and temporal information from optical flow, followed by fusion. With the advancement of skeleton-based approaches, some studies [13, 20] attempt to replace the optical flow stream with a skeleton stream. Although optical flow and skeleton are about the representation of the action, the issue of context bias caused by the RGB modality will still worsen during the collaborative training process of the two-stream networks.

In this paper, we propose a Skeleton-Dominated Two-Stream Network (SDTSN), as shown in Figure 1. Specifically, we first independently train a skeleton-based network that focuses on capturing the motion features, as well as an RGB-based network that captures fine-grained features. Subsequently, we divide the backbones of both networks into two streams and design a skeleton-dominated adaptive fusion module to fuse the features from both modalities, where the skeleton modality takes the lead. In addition, to prevent the loss of knowledge regarding the skeleton information, during the training process of the two-stream network, the parameters of the skeleton stream are frozen and not involved in the training. Experiments are conducted on a Mimetics [26] dataset without context bias and a popular Kinetics-400 [3] dataset. Compared to the skeleton-based methods, RGB-based approaches, and popular two-stream networks, the proposed SDTSN demonstrates state-of-the-art performance on the Mimetics dataset. Moreover, ablation studies demonstrate the effectiveness of each proposed component.

Our main contributions can be summarized as follows:

- We propose SDTSN to combine the skeleton and RGB modalities in the two-stream network, aiming to enhance the focus on action information, thereby mitigating context bias in mitigating video action recognition.
- A skeleton-dominated adaptive fusion module is developed to fuse two-stream features.
- A freeze parameter training strategy is designed to prevent the loss of knowledge regarding the skeleton information.

2 METHOD

The structure of the SDTSN is illustrated in Figure 2. For the skeleton stream, we commence by independently training a comprehensive skeleton-based network, utilizing the popular MS-G3D [16] architecture. Similarly, for the RGB stream, we independently train a complete RGB-based network, employing the widely adopted Video Swin transformer [15]. Subsequently, both networks retain only the backbone and employ the proposed skeleton-dominated adaptive fusion module to fuse the extracted features from both streams. The classification heads of the SDTSN utilize simple linear classification heads. During the training process, we freeze the weight parameters of the skeleton stream to prevent the loss of skeleton action information.

2.1 Skeleton Stream

The skeleton stream is based on MS-G3D [16], which is a powerful network for skeleton-based action recognition that combines multi-scale aggregation and spatial-temporal graph convolutions (G3D) operations. The multi-scale aggregation scheme can effectively capture the graph relationships on the human body skeleton, while the G3D operation can facilitate direct propagation of spatial-temporal information, enabling effective feature learning. By combining these two schemes, MS-G3D achieves multi-scale aggregation with multi-scale receptive fields in both spatial and temporal dimensions, further improving the model performance.

The architecture of MS-G3D contains a series of spatial-temporal graph convolutional (STGC) blocks, which are responsible for feature extraction from skeleton sequences. A global average pooling layer and a softmax classifier head then follow this.

During the training process, the entire MS-G3D is first independently trained on the skeleton dataset. Subsequently, the classification head is removed before incorporating MS-G3D into the skeleton stream of the two-stream network.

2.2 RGB Stream

The RGB stream is based on the Video Swin Transformer [15], which is a popular network that achieves state-of-the-art accuracy on a broad range of benchmarks [3, 9, 12, 23]. The architecture of Video Swin Transformer is upgraded from Swin Transformer [14] designed for the image domain, and adapted to the video domain.

The overall architecture of the Video Swin Transformer consists of multiple Video Swin Transformer Blocks, each including a Linear Embedding layer, a Patch Merging layer, and a 3D Patch Partition layer. The classification head is I3D [3]. SDTSN employs the base version of Video Swin Transformer, referred to as Swin-B, which has four stages, including multiple Video Swin Transformer Blocks.

Similar to the skeleton stream, the video swin transformer is first trained independently, then removed classification head and integrated into the two-stream network.

2.3 Skeleton-Dominated Adaptive Fusion

In the SDTSN framework, the skeleton stream and RGB stream output skeleton features x_s and RGB features x_r, respectively. To better fuse the two and alleviate information bias in the RGB features, we design a skeleton-dominated adaptive fusion module.

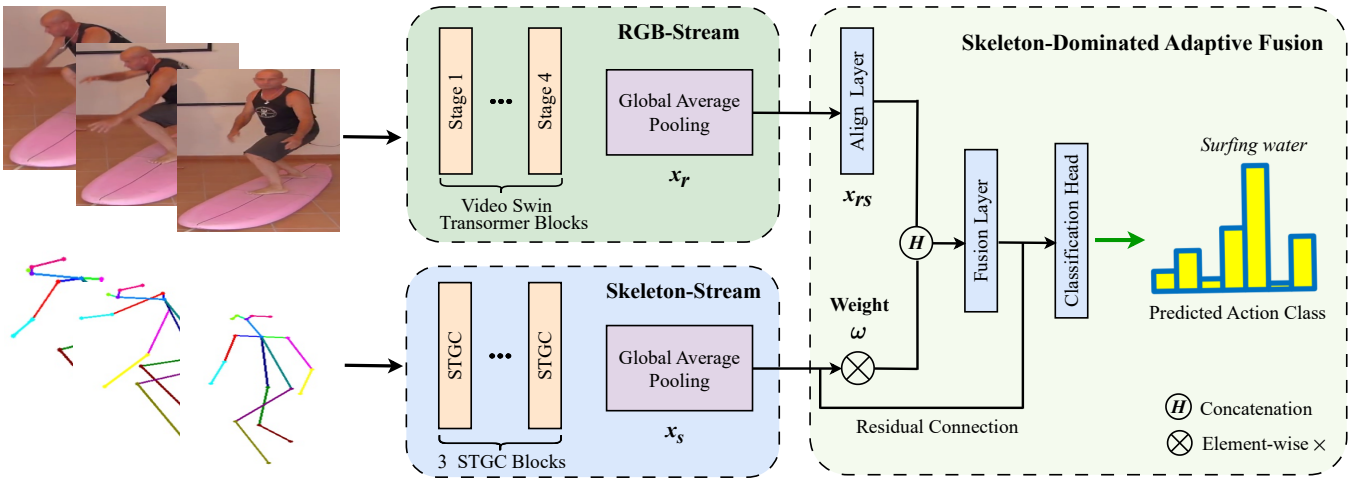

Figure 2: The overall architecture of the Skeleton-Dominated Two-Stream Network. SDTSN combines the skeleton and RGB modalities in the two-stream network, and designs a skeleton-dominated adaptive fusion module to fuse two-stream features.

Specifically, we first linearly transform the RGB features x_r, which have a relatively larger feature dimension, to match the feature dimension of the skeleton features, resulting in x_{rs}.

$$x_{rs} = W_{rs}x_r + b_r, \tag{1}$$

where, W_r is the weight matrix and b_r is the bias term for the linear transformation.

Next, we define the weight parameter ω required for feature fusion, which is allocated to the skeleton and set as a learnable parameter to achieve adaptivity. The range of this parameter is set to $[2, 5]$, with an initial value of 2, to emphasize the dominant role of the skeleton features in the fusion process. The fusion process entails concatenating the transformed RGB features x_{rs} with the skeleton features x_s, which have been multiplied by the adaptive weight parameter ω. This is followed by a linear transformation to squeeze the original dimensions. The concatenation results x_{cat} and squeeze result x_{squ} are defined as

$$x_{cat} = H(\cdot)(x_{rs}, \omega x_s), \tag{2}$$

$$x_{squ} = W_{squ}x_{cat} + b_{squ}, \tag{3}$$

where $H(\cdot)$ represents the concatenation, $\omega \in [2, 5]$ and ω with an inital value of 2.

To further enhance the dominance of the skeleton features, we introduce a residual connection that adds the skeleton features x_s to the compressed fusion features to form the final skeleton-dominated fusion results x_{fused}. This operation can be expressed as

$$x_{fused} = x_s + x_{squ}. \tag{4}$$

2.4 Training Strategy for Two-stream Network

Given the notable occurrence of context bias in the RGB stream trained independently, there is a potential for the skeleton stream to experience knowledge decay while training the two-stream network. To tackle this problem, we use a strategy that entails training only a portion of the parameters. Specifically, during the joint training of the two-stream network, we freeze the backbone parameters of the skeleton stream's MS-G3D to preserve its ability to focus on extracting action features.

2.5 Extra Disscussion on Ensemble Learning

Generally, skeleton-based networks have higher network efficiency due to the process of smaller-scale skeleton data. In this direction, ensembling skeleton-based network to other networks only costs a little, making it feasible for industrial applications. Taking our SDTSN as an example, the ensemble can be achieved simply by predicting the confidence level and combining it with the original MS-G3D network in the skeleton stream. This operation is able to further enhance the focus on action information, possibly improving accuracy for action recognization.

3 EXPERIMENT

3.1 Experiment Configuration

Kinetics-400 dataset. Kinetics-400 [3] is a popular large-scale video dataset containing 400 human action classes, each with at least 400 video clips. The videos are collected from YouTube and are between 10 and 30 seconds long. The official split of the dataset is 234,619 instances for the training set and 19,761 instances for the validation set. During our experiment, the MS-G3D skeleton network and the video swin transformer RGB network are independently trained on the Kinetics-400 dataset before integrating into the two-stream network.

Mimetics dataset. Mimetics [26] is a no-context bias dataset (As shown in Figure 3) of short YouTube video clips featuring mimed human actions, such as playing sports, performing daily activities, and playing musical instruments. It contains 713 video clips for 50 human action classes, which are subset classes of kinetics-400. Unlike other benchmarks, the video clips in this dataset are devoid of contextual biases, and some even lack interacting objects. Therefore, Mimetics focuses on out-of-context action recognition, making it suitable for evaluating the effectiveness of our method. Considering

Table 1: Results of the unimodal methods.

Method	Modality	Kinetics-400	Mimetics
MS-G3D [16]	Skeleton	38.0	**23.9**
Video Swin-B [15]	RGB	**95.5**	17.4

Table 2: Results of the two-stream methods.

Method	Modality	Mimetics
TSN [22]	Optical-Flow + RGB	18.7
SDTSN	Skeleton + RGB	21.7
SDTSN-Ensemble	Skeleton + RGB	**25.4**

the tiny scale of Mimetics, we randomly select two samples from each class to build the training set for few-shot learning scenarios.

Make skeleton data. To obtain the skeleton data from Kinetics-400 and Mimetics datasets, we first resize all videos to the resolution of 340×256 and convert the frame rate to 30 fps. Then, we extract skeletons from each frame in Kinetics by Openpose. Specifically, each skeleton graph contains 18 body joints, along with their 2D spatial coordinates and the prediction confidence score from OpenPose [2, 2] as the initial joint features. It is important to note that the process of generating these skeleton data aligns with the prior work [16, 21, 28].

Evaluation metric. To evaluate the performance of the proposed method, we use the Top-1 accuracy as the metric for action recognization. The calculation formulas as

$$\text{Top-1 accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100\%, \tag{5}$$

where the N_{correct} refers to the count of actions correctly classified by the proposed method, and the N_{total} represents the overall number of actions in the dataset being evaluated.

Implementation details. During the independent training phase, we use pre-trained weights from the official release of the Kinetics-400 dataset for both the skeleton-based and RGB-based networks. For the training phase, we use the Mimetics dataset. The batch size for the skeleton-based network is set to 64, with 65 epochs, a learning rate of 0.1, SGD optimizer, and weight decay set to 3e-4. The batch size for the RGB-based network is set to 6, with 50 epochs, a learning rate of 5e-5, Adam optimizer, and weight decay set to 3e-5. After incorporating our two-stream network SDTSN, we keep the configuration consistent with the RGB-based network. We conduct all experiments using the PyTorch framework on 4 Nvidia GeForce RTX 3090 GPUs with 24 GB memory.

3.2 Main Results and discussions

The results of the unimodal methods. Kinetics-400 [3] is currently the most extensively utilized large dataset for action recognition research, possessing the shared characteristic of substantial context bias inherent in popular datasets. Mimetics [26], on the other hand, is a tiny dataset that lacks context bias. Therefore, both these datasets are suitable for exploring the disparities between skeleton-based and RGB-based methods, thereby exemplifying the degree of emphasis placed on motion information. The results are listed in Table 1.

Specifically, on the Kinetics-400 dataset, the RGB-based Video Swin-B achieves a top-1 accuracy of 95.5%, which significantly outperforms the skeleton-based MS-G3D with a top-1 accuracy of 38.0%. This indicates the importance of fine-grained information contained in the RGB modality, and solely from the skeleton modality is not satisfactory. However, it is noteworthy that on the Mimetics dataset, MS-G3D achieves a top-1 accuracy of 23.9, which is 6.5% higher than Video Swin-B's performance on Kinetics-400. This suggests that the RGB-based approach may downplay the network's understanding of the actions themselves, which is inconsistent with human perception of actions. These results highlight the performance differences of different modalities on different datasets and provide valuable insights for further research in action recognition.

The results of the two-stream methods. The classical two-stream network posits that optical flow can capture greater dynamic information. TSN [22] represents a classic network that incorporates both RGB and multi-flow modalitie. In the previous section, the importance of the skeleton modality in action comprehension has already been demonstrated through unimodal experiments, while the RGB modality offers more fine-grained feature information. Therefore, our two-stream SDTSN is designed to use skeleton modality as dominated to combine RGB modality. In addition, SDTSN-Ensemble, as mentioned in the Method section, combines our two-stream network with the unimodal skeleton stream to further enhance action information. The results on the Mimetics dataset for the above three networks are listed in Table 2.

It is evident that our SDTSN successfully combines the skeleton modality and the RGB modality, achieving a Top-1 accuracy of 21.7%. Compared with RGB-based Video Swin-B in Table 1, the integration of the skeleton stream in SDTSN results in a 4.3% improvement in accuracy. Furthermore, when compared to TSN's Top-1 accuracy of 18.7%, SDTSN demonstrates a significant advantage, surpassing it by 3.0%. This indicates that our skeleton-dominated two-stream approach has a substantial advantage over the optical flow and RGB two-stream methods in action comprehension.

Notably, SDTSN-Ensemble achieves the best performance with a Top-1 accuracy of 25.4%. This highlights that the further ensemble of the skeleton stream in SDTSN leads to the most significant enhancement in action comprehension, surpassing the single skeleton stream in Table 1 by 1.5%. Additionally, in terms of efficiency, the model size of the skeleton-based MS-G3D is only 2.8M, making the integration into SDTSN a cost little, while improving the accuracy by approximately 17% compared to the original SDTSN. Therefore, it is feasible for practical applications.

3.3 Ablation Study

Most current RGB-based methods suffer from the issue of context bias, and our SDTSN primarily leverages the skeleton modality to enhance the model's understanding of actions. Therefore, we take the RGB-based Video Swin-B as the baseline to conduct ablation

Figure 3: Visualization of the Mimetics dataset. Mimetics is a no-context bias dataset, such as the first row depicting a basketball player emulating the action of archery on a basketball court. The second row showcases a basketball player performing a slam dunk without the physical presence of a basketball. The last row portrays a man executing the action of surfing water, albeit in an indoor setting.

Table 3: Ablation study of proposed components.

Baseline	Two-stream	Skeleton-dominated	Freeze skeleton stream	Ensemble	Top-1
✓					17.4
	✓				19.1
	✓	✓			21.3
	✓	✓	✓		21.7
	✓	✓	✓	✓	25.4

studies. We explore the effects of each proposed component, including the two-stream, skeleton-dominated adaptive fusion module, training strategy (freeze skeleton stream), and ensemble. The results of the ablation experiments are listed in Table 3.

The introduction of the skeleton stream has significantly enhanced the performance of the model, as evident from the comparison of the first two rows. Furthermore, comparing the first three rows demonstrates the proposed skeleton-dominated adaptive fusion module plays a crucial role. The comparison between the third and fourth rows reveals the positive impact of our training strategy. Lastly, the ensemble with the skeleton-based network further strengthened the utilization of action information and achieved the best accuracy.

4 CONCLUSION

In this paper, we propose a Skeleton-Dominated Two-Stream Network (SDTSN) for solving the context bias problem in most action recognition methods. Our SDTSN combines the skeleton and RGB modalities in the two-stream network, and designs a skeleton-dominated adaptive fusion module to fuse two-stream features. Additionally, we employ a training strategy, wherein the freeze parameter approach is implemented to prevent any loss of skeleton information. Moreover, we discuss ensemble learning that considers the cost factor. The Mimetics dataset is mainly used in this work, which is devoid of context bias. Extensive experiments and ablation studies demonstrate the effectiveness of our method.

REFERENCES

[1] 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.

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[3] 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.

[4] 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.

[5] 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.

[6] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 445–450.

[7] 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.

[8] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. 2019. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 244–253.

[9] 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.

[10] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. 2018. Safe driving: Driver action recognition using surf keypoints. In *2018 30th International Conference on Microelectronics (ICM)*. IEEE, 60–63.

[11] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. 2020. Soft spatial attention-based multimodal driver action recognition using deep learning. *IEEE Sensors Journal* 21, 2 (2020), 1918–1925.

[12] 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.

[13] Guiyu Liu, Jiuchao Qian, Fei Wen, Xiaoguang Zhu, Rendong Ying, and Peilin Liu. 2019. Action recognition based on 3d skeleton and rgb frame fusion. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 258–264.

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[15] 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.

[16] 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.

[17] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2018. Sport action recognition with siamese spatio-temporal cnns: Application to table tennis. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 1–6.

[18] Chundi Mu, Jianbin Xie, Wei Yan, Tong Liu, and Peiqin Li. 2016. A fast recognition algorithm for suspicious behavior in high definition videos. *Multimedia Systems* 22 (2016), 275–285.

[19] Hideo Saito, Thomas B Moeslund, and Rainer Lienhart. 2022. MMSports' 22: 5th International ACM Workshop on Multimedia Content Analysis in Sports. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7386–7388.

[20] Jing Shi, Yuanyuan Zhang, Weihang Wang, Bin Xing, Dasha Hu, and Liangyin Chen. 2023. A Novel Two-Stream Transformer-Based Framework for Multi-Modality Human Action Recognition. *Applied Sciences* 13, 4 (2023), 2058.

[21] 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.

[22] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems* 27 (2014).

[23] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *ArXiv Preprint ArXiv:1212.0402* (2012).

[24] Kamal Kant Verma, Brij Mohan Singh, and Amit Dixit. 2019. A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. *International Journal of Information Technology* (2019), 1–14.

[25] 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.

[26] Philippe Weinzaepfel and Grégory Rogez. 2021. Mimetics: Towards understanding human actions out of context. *International Journal of Computer Vision* 129, 5 (2021), 1675–1690.

[27] Fei Wu, Qingzhong Wang, Jiang Bian, Ning Ding, Feixiang Lu, Jun Cheng, Dejing Dou, and Haoyi Xiong. 2022. A survey on video action recognition in sports: Datasets, methods and applications. *IEEE Transactions on Multimedia* (2022).

[28] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.