ASFNET: JOINT APPEARANCE, STRUCTURE, AND FREQUENCY MODELING FOR ACTION RECOGNITION

YUNING SHI, HAN SHI, QINGWEI SONG, NAOYUKI KUBOTA

Department of Mechanical Systems Engineering, Graduate School of Systems Design, Tokyo Metropolitan University,
6-6 Asahigaoka, Hino-shi, Tokyo 191-0065, Japan
E-MAIL: {shi-yuning, shi-han, qingwei-song}@ed.tmu.ac.jp, kubota@tmu.ac.jp

Abstract:

Deep learning-based action recognition has achieved remarkable progress in understanding human activities from RGB videos and skeleton data. However, most existing methods primarily focus on spatio-temporal modeling in the time domain, often overlooking informative motion patterns present in the frequency domain. In this paper, we propose ASFNet, a novel multimodal framework that jointly models appearance, structural, and frequency representations for action recognition. Specifically, the appearance branch extracts visual features from RGB frames to capture both spatial appearance and temporal context, while the structure branch encodes the temporal dynamics of human poses using 2D skeleton sequences. To further exploit motion characteristics, we introduce a frequency branch that applies temporal differencing followed by Fourier transformation on skeleton trajectories, enabling the model to capture periodic and subtle motion patterns. These complementary features are then effectively integrated through a gated fusion mechanism, allowing for adaptive aggregation of appearance and motion cues. We evaluate our method on a self-collected dataset of daily living actions captured in a simulated home environment. Experimental results show that ASFNet outperforms the method that relies solely on RGB video data or omits frequency-domain features, highlighting the effectiveness of frequency-aware skeletal representations and multimodal fusion for robust real-world action recognition.

Keywords:

Action recognition; Deep learning; Multimodal learning; Feature fusion; Frequency analysis

1. Introduction

Population aging has become a global societal challenge, driving a rapidly growing demand for intelligent healthcare and elderly care systems. Continuous monitoring of daily behaviors is essential for supporting independent living among older adults and preventing functional decline. In this context, activities of daily living (ADLs), including walking, sitting, eating, dressing, and maintaining personal hygiene, serve as key indicators for assessing physical capability, autonomy, and overall quality of life. Monitoring ADL performance enables the early detection of functional changes and facilitates timely interventions and rehabilitation, thereby playing a crucial role in promoting healthy aging and preventing decline.

With the rapid advancement of artificial intelligence (AI), data-driven decision-making has become increasingly important in real-world applications. Consequently, exploring how to effectively integrate AI into practical environments has become a central focus of contemporary studies. For example, Japan's Moonshot Goal 3 aims to develop AI-powered robots capable of autonomous learning, environmental adaptation, and seamless coexistence with humans, with promising applications in homes, hospitals, and long-term care facilities.

To support intelligent elderly care systems, a variety of sensor-based and data-driven approaches have been proposed. Prior work has employed wireless sensor networks in conjunction with platforms such as informationally structured space (ISS) to integrate heterogeneous data for multimodal behavior analysis [1]. Meanwhile, life-log-based studies have utilized frameworks such as ICF, NHK activity classification, and social rhythm metrics (SRM) to evaluate behavioral regularity and associated health conditions [2]. However, these approaches lack a unified framework that can effectively integrate multimodal data while simultaneously capturing both semantic information and the spatio-temporal dynamics of daily activities.

Deep convolutional neural networks (CNNs) [3, 4, 5] have achieved remarkable success in image recognition. By integrating recurrent neural networks (RNNs) [6, 7, 8] or Trans-

former architectures [9] and extending 2D convolutions to 3D [10, 11], these approaches have been further advanced in modeling the spatio-temporal dynamics of human activities. In parallel, skeleton-based methods that leverage keypoint sequences with RNNs or graph neural networks (GNNs) [12] have also been widely adopted for this task.

However, most existing methods primarily focus on spatio-temporal modeling while neglecting the frequency domain. This limitation restricts their ability to capture subtle or periodic motion patterns and to fully exploit the complementarity between appearance and skeletal representations. To address these issues, we propose ASFNet, a multimodal framework for action recognition that integrates appearance, structural, and frequency representations. The framework comprises three specialized branches that extract features from RGB video frames and skeletal sequences. These features are adaptively fused via a gating module that selectively emphasizes complementary information to improve recognition performance.

We evaluate ASFNet on a custom dataset collected in a simulated home environment with 16 daily living action classes. Experimental results show that ASFNet achieves superior performance, demonstrating the effectiveness of multimodal fusion for action recognition in complex home settings.

2. Related work

Action recognition aims to automatically identify and classify human activities from video sequences, typically based on RGB video data or skeletal sequences. For RGB video inputs, hybrid architectures are commonly employed. CNN-RNN models, such as LRCN [13], use CNNs [3, 4, 5] for spatial feature extraction and RNNs [6, 7, 8] for temporal modeling. CNN-Transformer approaches, such as Action Transformer [14], replace recurrent modules with Transformers [9] to capture long-range dependencies. Alternatively, 3D CNNs, including C3D [10] and I3D [11], directly learn spatio-temporal representations from video volumes through 3D convolutions.

In skeleton-based action recognition, early methods primarily relied on RNNs to model the temporal dynamics of human joints. However, these approaches often fail to explicitly capture the structural relationships among joints. Graph convolutional networks (GCNs) [15] address this limitation by leveraging the inherent graph structure of the human skeleton. In particular, ST-GCN [16] extends this framework by incorporating both spatial and temporal graph convolutions, enabling joint modeling of inter-joint dependencies and temporal dynamics.

3. Proposed method

To effectively capture complementary information from different modalities and exploit informative motion patterns in both the time and frequency domains, we propose ASFNet, a multimodal framework for action recognition that jointly models appearance, structural, and frequency representations. As illustrated in Figure 1, ASFNet consists of three parallel branches: the appearance branch, which extracts visual features from RGB frames to capture spatial appearance and temporal context; the structure branch, which models spatio-temporal dynamics using skeleton sequences; and the frequency branch, which captures motion patterns by transforming skeleton trajectories into the frequency domain. The features from each branch are integrated through a gated fusion module that adaptively weights and aggregates multimodal features based on their relevance to the action, resulting in a unified representation for action classification.

3.1. Appearance branch

The appearance branch extracts high-level visual features from RGB frames, capturing both scene context and object-level information. Given an input video sequence of T frames, each frame $\mathbf{I}_t$ is processed by a CNN backbone to obtain a d-dimensional frame-level feature:

$$\mathbf{x}_t = f_{\text{CNN}}(\mathbf{I}_t) \in \mathbf{R}^d, \quad t = 1, 2, \ldots, T. \tag{1}$$

To capture temporal dynamics, the frame-level features are fed into an RNN, producing a sequence of hidden states:

$$\mathbf{h}_t = f_{\text{RNN}}(\mathbf{x}_t, \mathbf{h}_{t-1}). \tag{2}$$

A temporal attention mechanism aggregates informative frames with adaptive attention weights computed as:

$$\alpha_t = \frac{\exp(\mathbf{w}^\top \mathbf{h}_t)}{\sum_{k=1}^{T} \exp(\mathbf{w}^\top \mathbf{h}_k)}, \tag{3}$$

where $\mathbf{w}$ is a learnable parameter. The final appearance representation is obtained as a weighted sum of the hidden states:

$$\mathbf{F}_{\text{app}} = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t. \tag{4}$$

3.2. Structure branch

The structure branch models human motion based on 2D skeleton sequences. Given the detected keypoints for each

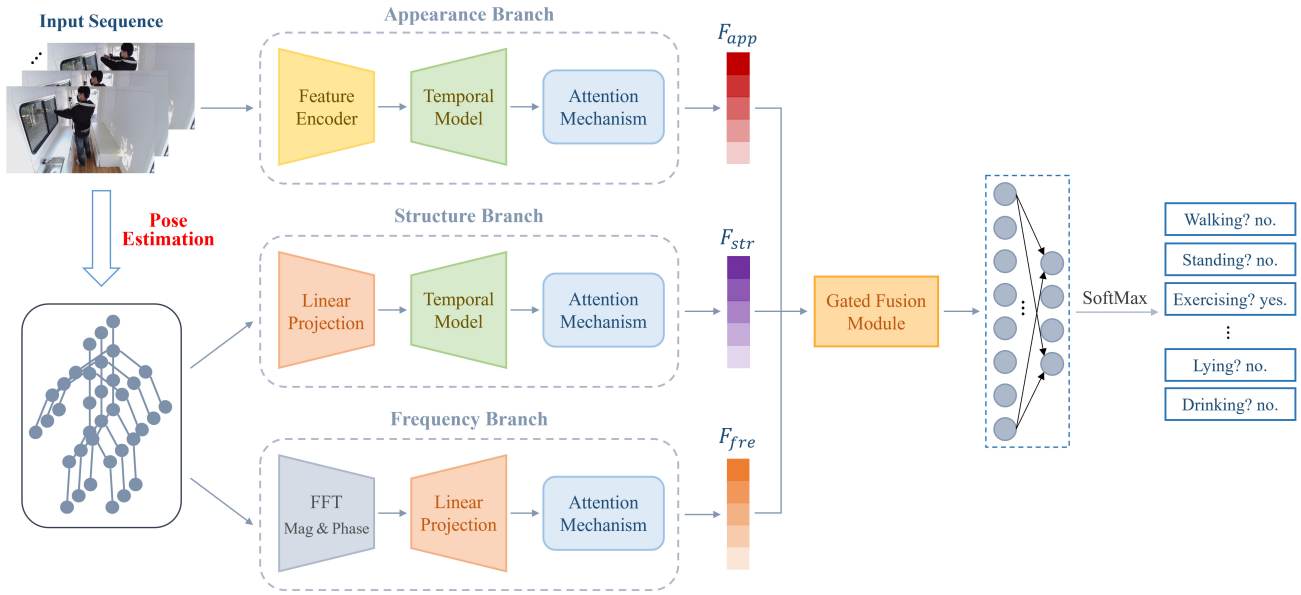


FIGURE 1. Overview of the proposed ASFNet framework. Skeleton points are first detected from RGB frames. The RGB frames are then fed into the appearance branch, while the skeleton points are processed by the structure and frequency branches to extract three complementary feature representations. These features are subsequently adaptively fused through the gated fusion module to generate the final representation for class prediction.

frame, each pose is represented as a set of joint coordinates with confidence scores, denoted as $\mathbf{y}_t \in \mathbf{R}^{N\times 3}$, where N is the number of skeletal joints. The coordinates are first normalized and flattened, then projected into a high-dimensional feature space through a multi-layer perceptron (MLP). Similar to the appearance branch, an RNN is used to model temporal dependencies, followed by an attention mechanism to emphasize informative time steps. The resulting representation $\mathbf{F}_{\text{str}} \in \mathbf{R}^d$ encodes the spatio-temporal dynamics of human motion.

3.3. Frequency branch

Many human actions, such as walking, brushing teeth, and washing hands, exhibit periodic patterns that are difficult to model in the time domain but can be effectively characterized in the frequency domain. The frequency branch captures motion patterns from a spectral perspective, complementing the spatial and temporal representations learned by the appearance and structure branches.

Given the input skeleton sequence $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$, we compute frame-wise differences to model motion dynamics:

$$\Delta\mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}, \quad t = 2, \ldots, T. \tag{5}$$

Due to the inherent noise in pose estimation caused by occlusions, motion blur, and detection errors, the estimated joint coordinates may be unreliable. To address this unreliability, we suppress noisy motion patterns by weighting the motion using the confidence scores:

$$\Delta\tilde{\mathbf{y}}_t = \Delta\mathbf{y}_t \odot (\mathbf{c}_t \odot \mathbf{c}_{t-1}), \tag{6}$$

where $\mathbf{c}_t \in [0, 1]^N$ denotes the confidence score of joints at frame t, and $\odot$ is element-wise multiplication.

The motion sequence is then transformed into the frequency domain by applying the fast Fourier transform (FFT) along the temporal dimension:

$$\mathbf{Z} = \mathbf{FFT}(\Delta\tilde{\mathbf{y}}). \tag{7}$$

From the complex-valued spectrum, we extract magnitude and phase and then concatenate them to form frequency features:

$$\mathbf{Z}_{\text{feat}} = [|\mathbf{Z}|; \angle\mathbf{Z}]. \tag{8}$$

The resulting features are projected into a high-dimensional space via a MLP:

$$\mathbf{H} = \phi(\mathbf{F}_{\text{feat}}). \tag{9}$$

To obtain a compact representation, we first apply average pooling to aggregate temporal information, followed by the same attention mechanism as in the appearance and structure branches to emphasize informative joints, producing the final frequency features $\mathbf{F}_{\text{fre}}$.

3.4. Gated fusion module

To effectively combine complementary information across modalities, we design a gated fusion module. First, the structure and frequency features are concatenated and projected to form a motion representation:

$$\mathbf{F}_{\text{mot}} = \phi([\mathbf{F}_{\text{str}}; \mathbf{F}_{\text{fre}}]) \in \mathbf{R}^d, \tag{10}$$

where $\phi(\cdot)$ denotes a linear transformation followed by a nonlinear activation. The motion feature $\mathbf{F}_{\text{mot}}$ is then fused with the appearance feature $\mathbf{F}_{\text{app}}$ via a gating mechanism. The gating weights are computed as:

$$[\mathbf{g}_{\text{app}}, \mathbf{g}_{\text{mot}}] = \sigma(\mathbf{W} \cdot [\mathbf{F}_{\text{app}}; \mathbf{F}_{\text{mot}}]), \tag{11}$$

where $\mathbf{g}_{\text{app}}$ and $\mathbf{g}_{\text{mot}} \in [0, 1]^d$ are adaptive weights for the appearance and motion features, $\mathbf{W}$ is a learnable weight matrix, and $\sigma(\cdot)$ is the sigmoid function.

The final fused representation is then used for final prediction and is obtained as:

$$\mathbf{F}_{\text{fused}} = \mathbf{g}_{\text{app}} \odot \mathbf{F}_{\text{app}} + \mathbf{g}_{\text{mot}} \odot \mathbf{F}_{\text{mot}}. \tag{12}$$

4. Experiments

4.1. Dataset

In this study, we created a custom dataset collected from eight participants in a simulated home environment. Videos were recorded from both front and back perspectives to capture multiple viewpoints of each activity. The dataset comprises 7,309 video clips covering 16 common daily activities, including Walking, Dressing, Eating, Sitting, Stand Up, and Washing Hands. Each clip has an approximate duration of five seconds. For model validation, data from five participants were used for training, one participant for validation, and two participants for testing. Table 1 summarizes the sample distribution.

TABLE 1. Per-category video counts in train, validation, and test sets

Category	Train	Validation	Test
BlowDryHair	327	66	98
BrushTeeth	346	55	128
Cleaning	348	48	93
CombHair	318	38	91
Dressing	268	60	82
Drinking	314	49	125
Eating	278	47	119
Exercise	318	43	134
Lying	331	29	120
SitDown	237	70	89
Sitting	256	24	117
StandUp	242	70	91
Standing	293	34	130
Walking	331	56	123
WashFace	281	39	90
WashHands	291	58	94
Total	4779	786	1724

4.2. Experimental setup

We compare ASFNet with two baselines that utilize different input modalities, including RGB-only and RGB+Skeleton. All experiments were implemented using the PyTorch framework. Experiments were conducted on a workstation equipped with an NVIDIA RTX 6000 Ada GPU, an Intel Xeon W7-3555 CPU, and 256GB system memory.

The input sequence length was set to 16, and the RGB video frames were resized to 224×224 pixels. We used YOLO11-Pose [17] to extract 17 skeletal keypoints per frame. Additionally, ResNet-18 [5] was employed as the backbone network, and a single-layer GRU [8] was used for sequence modeling. The output dimension of each branch was set to 256.

During training, the batch size was set to 128, and the model was trained for 50 epochs. We used the AdamW optimizer with an initial learning rate of 0.001 and a weight decay of 0.0001. The learning rate was reduced by a factor of 0.5 if the validation accuracy did not improve for five consecutive epochs. The model parameters that achieved the highest validation accuracy were saved and used for final evaluation.

4.3. Results

As shown in Table 2, the RGB-only baseline achieves 83.00% accuracy, 85.78% precision, 83.33% recall, and 82.43% F1-score. Incorporating skeleton data alongside RGB inputs leads to slightly lower performance, yielding 79.18% accuracy, 82.29% precision, 79.77% recall, and 78.55% F1-score. In contrast, ASFNet achieves the best overall performance, with 84.45% accuracy, 85.95% precision, 84.47% recall, and 84.46% F1-score.

Figure 2 presents the confusion matrices of the RGB-only baseline and ASFNet. Compared with the RGB-only baseline, ASFNet produces a higher number of correct predictions for categories such as BrushTeeth, CombHair, and Drinking. However, its performance noticeably degrades on the WashHands and WashFace categories.

4.4. Discussion

The RGB-only baseline demonstrates strong performance, indicating that visual features extracted from RGB frames are

TABLE 2. Performance comparison of different methods on the test set

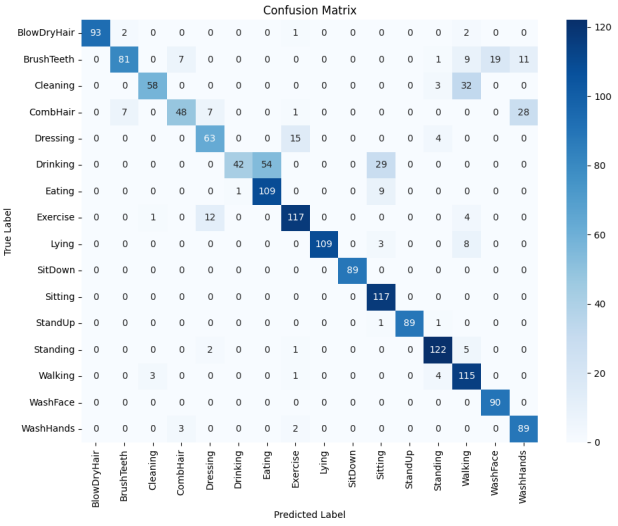
Method	Accuracy	Precision	Recall	F1-score
RGB-only	83.00	85.78	83.33	82.43
RGB + Skeleton	79.18	82.29	79.77	78.55
ASFNet	**84.45**	**85.95**	**84.47**	**84.46**

highly informative for this task. In contrast, the method combining RGB frames with skeleton points exhibits certain limitations. Although skeleton points capture structural information about human poses, they often fail to adequately represent fine-grained human-object interactions, such as brushing teeth, dressing, eating and drinking, which involve tool use or precise hand movements. Moreover, skeleton points extracted using YOLO11-pose in this study may contain noise or inaccuracies, potentially interfering with the already informative RGB features and slightly reducing overall performance.

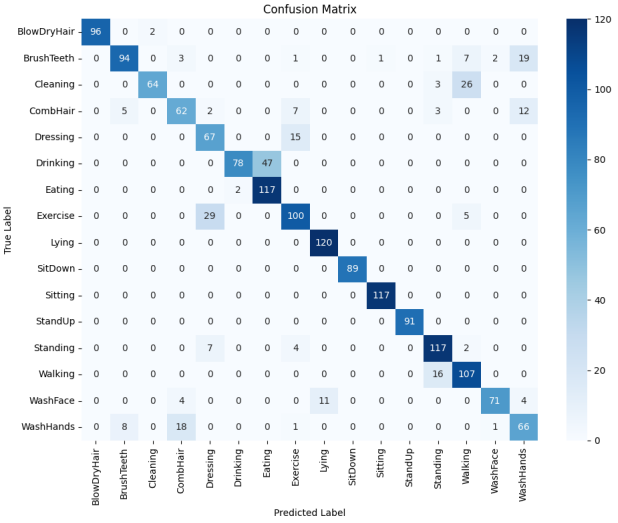
By comparison, ASFNet outperforms both baselines by adaptively fusing complementary features, including appearance, structural, and frequency information. It also employs confidence weighting to capture time-frequency dynamics, effectively reducing the influence of noisy skeleton points and enhancing the quality of spatio-temporal features. This approach enables ASFNet to achieve higher recall and F1 scores, produce more consistent predictions across diverse action categories, and demonstrate robustness in real-world scenarios.

The confusion matrix indicates that ASFNet achieves superior performance on actions such as BrushTeeth, CombHair, and Drinking. This can be attributed to the complementary cues provided by appearance, structural, and frequency modalities, which capture distinctive object features, rhythmic motions, and periodic patterns. However, performance declines for actions such as WashHands and WashFace due to their high similarity in hand positions, motion patterns, and frequency characteristics. The lack of sufficiently discriminative cues makes these actions inherently ambiguous, posing a challenge even for multimodal fusion.

Overall, ASFNet demonstrates advantages in recognizing actions with distinctive appearance, structural, and frequency characteristics, particularly those involving tool use. However, actions with highly similar characteristics across these modalities remain challenging to distinguish.



(a) RGB-only baseline



(b) ASFNet

FIGURE 2. Confusion matrices comparison.

5. Conclusion and future work

In this paper, we propose ASFNet, a multimodal framework that integrates appearance, structural, and frequency information for human daily action recognition. The framework consists of three independent branches, each extracting complementary information from a specific modality. The appearance branch processes RGB video frames to capture spatial and texture cues, the structure branch analyzes skeletal sequences to model body dynamics and motion patterns, and the frequency branch captures temporal frequency characteristics, providing an additional perspective on action dynamics. Features from all three branches are then adaptively fused using a gating module that selectively emphasizes the most informative signals from

each modality. This strategy improves recognition robustness by effectively leveraging complementary multimodal cues.

To evaluate our method, we collected a dataset containing 16 categories of daily living actions in a simulated home environment. Experimental results show that ASFNet outperforms methods that rely solely on RGB video data or omit frequency-domain features. These results confirm the effectiveness of frequency-aware skeletal representations and multimodal information fusion for accurate and robust action recognition.

In future work, we plan to validate ASFNet on standard benchmark datasets for action recognition and compare it with SOTA models to further evaluate and enhance its performance.

Acknowledgements

This work was partially supported by the Japan Science and Technology Agency (JST) under Moonshot R&D (Grant Number JPMJMS2034).

References

[1] Tang, D., Yoshihara, Y., Takeda, T., Botzheim, J., and Kubota, N., "Informationally Structured Space for Life Log Monitoring in Elderly Care", 2015 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1421–1426, 2015.

[2] Tang, D., Yoshihara, Y., Obo, T., Takeda, T., Botzheim, J., and Kubota, N., "Social rhythm management support system based on Informationally Structured Space", 2016 9th International Conference on Human System Interactions (HSI), pp. 429–434, 2016.

[3] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks", Advances in Neural Information Processing Systems, vol. 25, 2012.

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

[5] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[6] Elman, J. L., "Finding structure in time", Cognitive Science, vol. 14, no. 2, pp. 179–211, 1990.

[7] Hochreiter, S. and Schmidhuber, J., "Long short-term memory", Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y., "Empirical evaluation of gated recurrent neural networks on sequence modeling", arXiv:1412.3555, 2014.

[9] Vaswani, A., et al, "Attention is all you need", Advances in Neural Information Processing Systems, vol. 30, 2017.

[10] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M., "Learning spatiotemporal features with 3D convolutional networks", Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497, 2015.

[11] Carreira, J. and Zisserman, A., "Quo vadis, action recognition? A new model and the Kinetics dataset", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308, 2017.

[12] Gori, M., Monfardini, G., and Scarselli, F., "A new model for learning in graph domains", Proceedings of the IEEE International Joint Conference on Neural Networks, vol. 2, pp. 729–734, 2005.

[13] Donahue, J., et al. , "Long-term recurrent convolutional networks for visual recognition and description", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634, 2015.

[14] Girdhar, R., Carreira, J., Doersch, C., and Zisserman, A., "Video action transformer network", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 244–253, 2019.

[15] Kipf, T. N. and Welling, M., "Semi-supervised classification with graph convolutional networks", arXiv:1609.02907, 2016.

[16] Yan, S., Xiong, Y., and Lin, D., "Spatial temporal graph convolutional networks for skeleton-based action recognition", Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[17] Khanam, R. and Hussain, M., "YOLOv11: An overview of the key architectural enhancements", arXiv:2410.17725, 2024.