Multimodal Graph Attention Networks for Real-Time Action Prediction and Accident Prevention in Industrial Environments

Guilherme Nunes
Centro Universitário FEI
Email: guilherme.l.nunes@gmail.com

Paulo Sérgio Rodrigues
Centro Universitário FEI
Email: psergio@fei.edu.br

***Abstract*—One of the major challenges faced by the manufacturing industry is the prevention of workplace accidents. In this context, ensuring compliance with safety regulations, such as the use of personal protective equipment (PPE) and the proper execution of specific tasks according to safety protocols, is essential, especially when supervisors or safety personnel are not present on site. To address this issue, we propose a multimodal method for real-time human action prediction in industrial environments, aimed at supporting accident prevention systems. Our approach integrates two parallel Graph Attention Networks (GATs): one based on human skeleton pose estimation, and another built from scene object detection graphs. By combining these two complementary modalities, the model captures both human motion dynamics and contextual environmental information. To the best of our knowledge, this approach has not yet been explored in the literature. The proposed method will be evaluated on two benchmark datasets: *Kinetics-400* (a large-scale video dataset with diverse real-world actions), and *UnsafeNet* (a dataset featuring factory-recorded videos annotated with safe and unsafe behaviors). The expected results aim to demonstrate the feasibility of applying multimodal GAT-based architectures to enhance occupational safety through intelligent action recognition systems.**

I. Introduction

Data released by the *Brazilian Superior Labor Court* [1] show that workplace accidents cause at least one death every three hours and forty-seven minutes in Brazil. Continuous supervision to ensure compliance with safety regulations and correct use of personal protective equipment (PPE) is essential, especially when production supervisors or safety professionals are not physically present. This challenge is even greater for fieldwork, such as that performed by workers in distribution networks, telephony towers, and electrical infrastructure, since these workers operate in geographically dispersed locations, making monitoring and support for safety practices more difficult.

In industrial environments, particularly on the factory floor, ensuring the safety of workers during operations remains a significant challenge for companies. The use of automated systems for human action recognition based on video data has gained prominence in recent years. Studies such as [2], [3], and [4] have made advances in automatic PPE detection but have yet to fully explore the analysis of worker's movements and proper task execution.

As presented in the survey by [5], the most promising methods for human action classification rely on graph neural networks (GNNs) applied to human skeletons extracted through pose estimation algorithms. Since the pioneering method by [6], models such as [7], [8], and [9] have demonstrated excellent results, solidifying the use of GNNs for action recognition tasks.

A natural evolution of this research is real-time human action prediction, enabling early identification of unsafe behaviors and preventive measures. Industrial environments, being highly dynamic, noisy, and visually complex, add additional challenges to this task. Action prediction—also known as early action recognition—aims to classify an action before its completion, based on a partially observed frame sequence. This approach has been explored in domains such as autonomous vehicles, fall prevention in hospitals, and occupational safety.

Specifically in factory settings, many human actions involve interaction with tools and objects in the environment. However, the literature still lacks studies explicitly modeling this interactivity. Incorporating this information can be crucial to improving predictive model accuracy, allowing for more precise detection of deviations in task execution.

Therefore, this research proposes and evaluates a multimodal method for early human action prediction in industrial environments, focused on accident prevention. The proposed model combines two graph attention networks (GATs): one modeling the worker's body dynamics using skeletons estimated by pose algorithms, and another representing objects and tools present in the scene. Both are integrated through a fusion module to capture body-environment interactions, enabling richer and more contextualized action analysis.

II. Related Work

The method proposed by [6] was a pioneer in applying Graph Neural Networks (GNNs) to skeleton graphs, setting a new state of the art for the task of human action recognition based on joint positions. In this approach, the graph is constructed by connecting spatially adjacent joints to capture pose-based spatial relationships. Additionally, corresponding joints across consecutive frames are connected to capture temporal dynamics. This spatiotemporal graph is then processed by a Graph Convolutional Network (GCN), which aggregates and learns the relationships between the joints.

Subsequent works inspired by this GNN-based skeleton graph approach have attempted to improve it by exploring three main aspects: *Graph Topology*, *Temporal Modeling*, and *Ambiguity Mitigation*.

1) Graph Topology: Several studies, such as [10], [11], [12], and [13], sought to improve skeleton graph topology by treating the adjacency matrix as a learnable parameter. This allows the model to learn the optimal inter-joint relationships instead of relying on a fixed, predefined human-skeleton topology.

Although learning the adjacency matrix led to significant advances over earlier approaches, the incorporation of *attention mechanisms* for discovering dynamic relationships between nodes has yielded even better results. Attention-based models have since become a common component in recent state-of-the-art methods. Notable examples include [14], [8], and [9].

2) Temporal Modeling: The task of human action recognition requires modeling both short- and long-range dependencies across video frames. A robust model must learn how to extract relevant temporal patterns and determine how much of that information should be propagated. To enhance this temporal reasoning, several works have proposed specialized modules, including [15], [16], and [17].

The model proposed in [7] introduces two key components: the *MS-GC (Multi-Scale Spatial Graph Convolution)* to extract spatial relations, and the *MT-GC (Multi-Scale Temporal Graph Convolution)* to capture temporal dependencies. Both modules are based on the *Res2Net* architecture, as introduced in [18]. In the *MT-GC* module, spatial information extracted from a frame via *GCN* is transformed into an embedding and propagated to the next frame as a residual signal through a temporal convolution. This temporal modeling approach has become one of the most effective and commonly used strategies in recent state-of-the-art methods. While RNN-based approaches have been widely used for skeleton-based action recognition (e.g., bi-RNNs, LSTMs, attention-based models), [7] shows that sequential processing can struggle to capture complex spatial relationships between distant joints. The GCN-based approach in MST-GCN avoids these limitations while efficiently modeling both short and long-range dependencies.

3) Ambiguity Mitigation: A common challenge in human action recognition models is the misclassification between similar action classes. For instance, distinguishing between a person running and a person walking can be difficult when analyzing a single frame, as both actions involve similar body postures and motion patterns.

Several studies have proposed strategies to address this ambiguity. In [19], the authors introduce an auxiliary feature refinement module designed to better differentiate highly similar actions such as reading and writing. This module can be integrated at various stages of a GCN, enhancing the model's discriminative capability. In [20], the authors propose a method where the extracted human skeleton is divided into subgraphs representing the head, arms, legs, and hips. Each subgraph generates a descriptive sentence of the body part's position, which is then used as a prompt for GPT-3 to perform the final classification.

Another novel approach is presented in [21], where multiple CNNs are trained, each specialized in distinguishing a specific pair of similar classes. A dedicated CNN processes the initial frames of the video to produce an embedding, which serves as a query to dynamically select the appropriate specialized network for final classification.

4) Human Action Prediction: Another research direction extends human action recognition into the task of human action prediction — that is, classifying actions at an early stage, before they are fully completed, using only the initial frames of a video.

The models proposed in [21] and [22] employ modules trained on specific segments of the video, enabling the system to capture early cues of the ongoing action. In [23], a reinforcement learning-based method is introduced, where the agent is rewarded for making correct early predictions, thus encouraging timely decision-making.

A promising direction also involves synthetic frame generation. The models presented in [24] and [25] achieve strong results by generating missing video frames prior to prediction, increasing the available temporal information and improving the accuracy of early classification.

5) Human Action Recognition for Occupational Safety: Most computer vision techniques applied to occupational safety do not incorporate human action recognition. Typically, they rely on object detection networks to verify whether workers are wearing personal protective equipment (PPE) and operating within designated areas. Notable examples include the methods proposed in [26] and [27]. However, these approaches often overlook the execution of the task itself, which may significantly reduce their effectiveness in scenarios where the manner in which an action is performed is critical to ensuring worker safety.

6) Fusion of GNN-Based Features: The combination of heterogeneous representations is motivated by findings in multimodal learning. Gómez-Chova et al. [28] show that multimodal fusion effectively captures complementary information in remote sensing and other computer vision tasks. Building on this idea, Liang et al. [29] combined CNNs and GCNs for scene classification, demonstrating improved capture of local patterns and relational dependencies. Following this principle, we fuse skeleton-based and object-based GAT features to jointly model worker dynamics and contextual scene information.

III. Methodology

A. General Idea and Contribution

This work proposes a methodology that integrates human action recognition with object detection to support accident prevention in industrial environments. The underlying hypothesis is that incorporating human action recognition can enhance the detection of hazardous behaviors on the factory floor. Additionally, combining object detection with human action

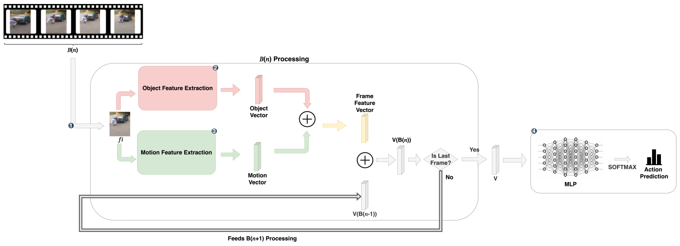

Fig. 1. Proposed Model Overview

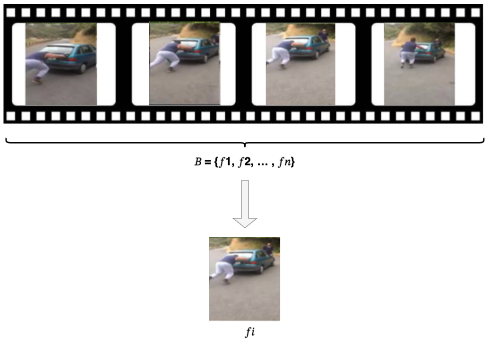

Fig. 2. Frame Extraction

recognition may improve overall performance in recognizing human activities.

The main contribution of this study is the proposal and evaluation of a novel method that fuses human skeletal motion features with object detection data from the scene. To the best of our knowledge, this specific approach has not yet been explored in the literature.

B. Frame Extraction

This work adopts a frame sampling strategy inspired by [17], in which a video with F frames is divided into B blocks, and one randomly selected frame from each block is used for processing.

To enable real-time analysis, instead of feeding the entire video, the proposed method reads only $\frac{F}{B}$ frames—one per block—directly as input. A block is defined as a sequence of n frames, $B = \{f_1, f_2, \ldots, f_n\}$, from which a single frame f_i ($i \in [1, n]$) is randomly selected. Figure 2 illustrates this sampling strategy.

C. Feature Extraction

This module extracts information from each randomly selected frame f_i (as described in Section III-B), feeding two parallel processes: object feature extraction and motion feature extraction. As shown in the general methodology (Figure 1), both types of features are aggregated into a single output vector that accumulates information across all sampled frames.

By combining static object features with human motion features—capturing the interaction between humans and their environment—this approach aims to improve the accuracy of action recognition.

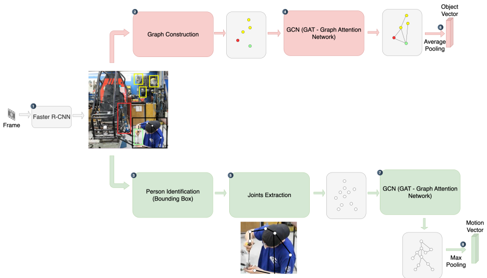

Fig. 3. Feature Extraction

D. Object and Person Identification

To identify static objects in the scene, the *Faster R-CNN* network [30] is employed due to its high accuracy and suitable performance for real-time applications. Unlike *YOLO* [31], *Faster R-CNN* handles small object detection more effectively, achieving processing rates of up to 5 FPS.

Each randomly selected frame f_i from a block B is processed by *Faster R-CNN* (Figure 3, step 1), which detects the objects of interest and returns their class labels, bounding boxes, and positions.

E. Scene Object Feature Extraction

In step 2 (Figure 3), the centroid of each bounding box detected in step 1 is computed. Combined with the object class information, this forms a data structure (class + centroid) for each object, which is then passed as input to a Graph Attention Network (GAT) [30] (Figure 3, step 4). The GAT produces a graph $G_{objects}(V, A)$, where the vertex set $V_{objects} = \{v_i \mid i \in \{1, ..., N\}\}$ represents the detected objects, and the edge set $A_{objects} = \{a_{i,j} \mid \forall i, j \in \{1, ..., N\}\}$ captures semantic relationships among them.

Finally, the scene object feature vector $v_{objects} \in R^n$ (Figure 3, step 6) is obtained by applying average pooling over the vertex set $V_{objects}$. Following the findings in [6], average pooling is used here due to its effectiveness in capturing node distribution, which aligns with our objective of modeling the quantity and types of objects in the scene.

F. Motion Feature Extraction

Using the bounding boxes and class labels from the object and person detection step (Section III-D), the region corresponding to the person is selected (Figure 3, step 3). This region is fed into an *HRNet* model [32], which estimates the 2D positions of body joints (Figure 3, step 5).

Following [33], the choice of pose estimation method has a significant impact on model accuracy. Since 2D top-down estimators outperform 3D and bottom-up approaches—especially on benchmarks like *COCO-keypoints* [34]—this work adopts a top-down 2D estimation strategy.

The extracted joints are modeled as a graph, which is then processed by a Graph Attention Network (GAT) (Figure 3, step 7). The resulting graph $G_{joints}(V, A)$ has a vertex set $V_{joints} = \{v_i \mid i \in \{1, ..., N\}\}$ representing body joints, and

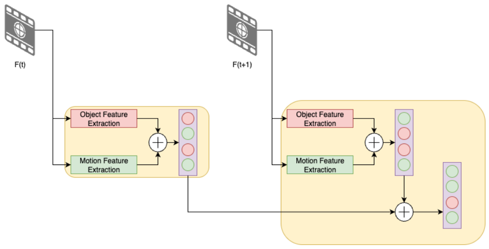
F(t)
F(t+1)
Object Feature Extraction
Motion Feature Extraction
Object Feature Extraction
Motion Feature Extraction

Fig. 4. Temporal Modeling

an edge set $A_{joints} = \{a_{i,j} \mid \forall i, j \in \{1, ..., N\}\}$ representing semantic relationships between them.

The final motion feature vector $v_{motion} \in R^n$ (Figure 3, step 8) is obtained by applying max pooling over the vertex set V_{joints}. Based on [6], max pooling is preferred here since capturing node distribution is less relevant for representing human joint configurations.

G. Scene and Motion Feature Vector Construction

The vectors $v_{objects}$ and v_{motion}, obtained as described in Sections III-E and III-F, are combined to form a single feature vector $v_{f_i} \in R^n$ representing the characteristics of frame f_i, as shown in Equation 1.

$$v_{f_i} = v_{objects} \oplus v_{motion} \tag{1}$$

H. Temporal Modeling

To model the temporal dimension of the features extracted in Section III-C, this work accumulates information across a sequence of frames. For each block of B frames, we apply the *MT-GC* technique proposed in [7] (see Section II-2). *MT-GC* is based on residual networks and is widely adopted in recent state-of-the-art approaches.

At the end of each frame processing step (Section III-C), the feature vector v_{f_i} is generated by combining the current frame's objects and motion features with the feature vector from the previous frame $v_{f_{i-1}}$, thus capturing temporal dependencies across blocks. This fusion process is illustrated in Figure 4 and formalized in Equation 2.

$$v_{f_i} = \begin{cases} v_{objects} \oplus v_{motion} & \text{for } i = 0 \\ (v_{objects} \oplus v_{motion}) \oplus v_{f_{i-1}} & \text{for } i > 0 \end{cases} \tag{2}$$

I. MLP Network

The vector v, obtained by aggregating all feature vectors v_{f_i} extracted from the frames (as described in Section III-H), is used as input to a Multi-Layer Perceptron (*MLP*). A *Softmax* function is applied at the output layer to compute the probability distribution over possible actions. The action with the highest probability is selected as the predicted class.

IV. Expected Results

A. Datasets

TABLE I
Baseline Models Used for Comparison

Paper	Model	Dataset
[17]	Temporal Difference Network	Kinetics 400
[33]	PoseConv3D	Kinetics 400
[35]	Unsafe-Net	Unsafe Net

The *Kinetics-400* dataset was selected based on the nature of the video content. Unlike other datasets composed of scripted actions recorded in controlled environments, *Kinetics-400* consists of user-generated YouTube videos, where actions often lack clearly defined movements. However, the presence of rich contextual information from surrounding objects makes it well-suited for evaluating the proposed methodology.

The *UnsafeNet* dataset stands out as the only one considered in this work that features factory-floor videos labeled as either safe or unsafe. While it is highly aligned with the objectives of this study, it is a recent dataset that has not yet been widely explored by other methods, providing a limited basis for comparison—currently restricted to [35].

B. Expected Quantitative Results

The quantitative metrics used to evaluate the performance of the human action prediction models are *Accuracy* and *AUC (Area Under the Curve)*. Unlike traditional action recognition models, action prediction models are not evaluated using the full video as input. Instead, the concept of observation ratio is applied, in which only portions of the video are provided to the model. This approach allows for assessing the model's ability to predict actions based on partial information, simulating real-world scenarios where the action is still in progress. Testing across multiple observation ratios allows us to evaluate how well the model predicts actions as they unfold, providing insight into its performance under different levels of partial information.

Prior works used for comparison with the proposed methodology adopt a 20% observation ratio for evaluation. Accordingly, each video will be divided into five segments, and the model will be tested using $20\%, 40\%, 60\%, 80\%$, and 100% of the original video. Note that using 100% of the video transforms the task from action prediction into standard action recognition. Table I presents the models selected for comparative evaluation.

C. Expected Qualitative Results

For the qualitative evaluation, an embedding analysis will be conducted. A balanced subset of videos will be randomly selected to represent all action classes in the dataset. These videos will be processed by the trained network with optimized parameters. The values produced by the perceptrons in the last hidden layer of the MLP model (Section III-I) will be used as embeddings representing each processed video. To enable visual analysis, the dimensionality of the embeddings will be reduced to three using Principal Component Analysis (PCA), allowing for the inspection of action clusters in a three-dimensional space.

References

[1] N. Pianegonda. (2023) Acidentes de trabalho matam ao menos uma pessoa a cada 3h47min no brasil. [Online]. Available: https://tst.jus.br/-/acidentes-de-trabalho-matam-ao-menos-uma-pessoa-a-cada-3h47min-no-brasil-1

[2] S. Adhikesaven, "An industrial workplace alerting and monitoring platform to prevent workplace injury and accidents," *ArXiv*, vol. abs/2210.17414, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253237851

[3] N. D. Nath, A. H. Behzadan, and S. G. Paal, "Automation in construction," 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:213081043

[4] M. M. Saudi, A. Hakim, A. Ahmad, A. Shakir, M. Hanafi, A. Narzullaev, and M. Ifwat, "Image detection model for construction worker safety conditions using faster r-cnn," *International Journal of Advanced Computer Science and Applications*, vol. 11, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221349861

[5] C. ling Wang and J. Yan, "A comprehensive survey of rgb-based and skeleton-based human action recognition," *IEEE Access*, vol. 11, pp. 53 880–53 898, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259058367

[6] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:19167105

[7] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235306189

[8] Y. Zhou, C. Li, Z.-Q. Cheng, Y. Geng, X. Xie, and M. Keuper, "Hypergraph transformer for skeleton-based action recognition," *ArXiv*, vol. abs/2211.09590, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253581243

[9] H. gun Chi, M. H. Ha, S. geun Chi, S. W. Lee, Q.-X. Huang, and K. Ramani, "Infogcn: Representation learning for human skeleton-based action recognition," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20 154–20 164, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250286899

[10] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 018–12 027, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:195440283

[11] ——, "Skeleton-based action recognition with directed graph neural networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7904–7913, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:195446390

[12] F. Ye, S. Pu, Q. Zhong, C. Li, D. Xie, and H. Tang, "Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220845919

[13] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13 339–13 348, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236428765

[14] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition," in *Asian Conference on Computer Vision*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:229622392

[15] H. Liu, J. Tu, and M. Liu, "Two-stream 3d convolutional neural network for skeleton-based action recognition," *ArXiv*, vol. abs/1705.08106, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3733902

[16] M. Korban and X. Li, "Ddgcn: A dynamic directed graph convolutional network for action recognition," in *European Conference on Computer Vision*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:226308363

[17] L. Wang, Z. Tong, B. Ji, and G. Wu, "Tdn: Temporal difference networks for efficient action recognition," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1895–1904, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:229331798

[18] S. Gao, M.-M. Cheng, K. Zhao, X. Zhang, M.-H. Yang, and P. H. S. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 652–662, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:91184391

[19] H. Zhou, Q. Liu, and Y. Wang, "Learning discriminative representations for skeleton based action recognition," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 608–10 617, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257378316

[20] W. Xiang, C. Li, Y. Zhou, B. Wang, and L. Zhang, "Generative action description prompts for skeleton-based action recognition," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 242–10 251, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251467851

[21] L. G. Foo, T. Li, H. Rahmani, Q. Ke, and J. Liu, "Era: Expert retrieval and assembly for early action prediction," *ArXiv*, vol. abs/2207.09675, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250698984

[22] X. Wang, J. Hu, J. Lai, J. Zhang, and W. Zheng, "Progressive teacher-student learning for early action prediction," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3551–3560, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:198918427

[23] J. Weng, X. Jiang, W.-L. Zheng, and J. Yuan, "Early action recognition with category exclusion using policy-based reinforcement learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, pp. 4626–4638, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:213992081

[24] G. Pang, X. Wang, J. Hu, Q. Zhang, and W. Zheng, "Dbdnet: Learning bi-directional dynamics for early action prediction," in *International Joint Conference on Artificial Intelligence*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:199465926

[25] S. geun Chi, H. gun Chi, Q. Huang, and K. Ramani, "Infogcn++: Learning representation by predicting the future for online human skeleton-based action recognition," *ArXiv*, vol. abs/2310.10547, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264146685

[26] O. Elharrouss, N. Almaadeed, S. A. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A combined multiple action recognition and summarization for surveillance video sequences," *Applied Intelligence*, vol. 51, pp. 690 – 712, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221310912

[27] O. Önal and E. Dandıl, "Video dataset for the detection of safe and unsafe behaviours in workplaces," *Data in Brief*, vol. 56, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271698999

[28] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1560–1584, 2015.

[29] J. Liang, Y. Deng, and D. Zeng, "A deep neural network combined cnn and gcn for remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4325–4338, 2020.

[30] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:10328909

[31] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:206594738

[32] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 3349–3364, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:201124533

[33] H. Duan, Y. Zhao, K. Chen, D. Shao, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2959–2968, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233423473

[34] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:14113767

[35] O. Önal and E. Dandıl, "Unsafe-net: Yolo v4 and convlstm based computer vision system for real-time detection of unsafe behaviours in workplace," *Multimedia Tools and Applications*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269595736