PoseIntelliGraph: Detecting Violence in Video Using Human Pose Graph Neural Networks

Marc Anthony B Reyes
Department of Computer Science
University of the Philippines Diliman
P. Velasquez Street, Diliman, Quezon City, 1800 Metro Manila
mbreyes12@up.edu.ph

***Abstract*—We present PoseIntelliGraph, a novel deep learning approach for violence detection in video using human pose estimation. Our method leverages a hybrid architecture combining Graph Neural Networks (GNN) and Transformer models to analyze the spatio-temporal relationships between human joints. The system processes pose keypoints extracted using MMPose, representing each person as a graph where nodes are keypoints and edges represent anatomical connections. PoseIntelliGraph incorporates three specialized graph convolution layers (GCN, GAT, and GIN) followed by a transformer encoder to capture complex pose dynamics indicative of violent behavior. Experiments on a large-scale violence detection dataset demonstrate promising results with an AUC of 0.7255, precision of 0.84, and F1-score of 0.69. The confusion matrix analysis shows strong performance in identifying violent scenes (7,949 true positives) while maintaining reasonable false positive rates. PoseIntelliGraph achieved optimal performance at a classification threshold of 0.696, determined using Youden's J statistic. Our approach offers advantages over RGB-based methods by focusing on human motion patterns rather than visual appearance, potentially improving robustness to lighting conditions and scene variations. This work contributes to the growing field of automated violence detection systems for surveillance and content moderation applications.**

***Index Terms*—violence detection, graph neural networks, pose estimation, transformers, action recognition, computer vision**

I. Introduction

The automatic detection of violent behavior in video content represents a significant challenge in computer vision and artificial intelligence research. With the exponential growth of video data across social media, surveillance systems, and streaming platforms, there is an increasing need for reliable automated methods to identify potentially harmful content [1]. Traditional approaches have primarily relied on appearance-based features extracted from RGB video frames, employing convolutional neural networks (CNNs) to classify violent actions [2]. However, these methods often struggle with generalization across different environments, lighting conditions, and camera angles.

Human pose estimation has emerged as a promising alternative for action recognition tasks, offering a more abstract representation that focuses on the fundamental kinematic patterns of human movement rather than visual appearance [3]. This abstraction potentially enables more robust performance across diverse visual conditions. Recent advances in pose estimation frameworks such as MMPose [4] have made high-quality human keypoint detection increasingly accessible, opening new avenues for downstream applications including violence detection.

Graph Neural Networks (GNNs) have demonstrated remarkable effectiveness in modeling structured data with explicit relational information [5], [6]. The human skeleton naturally lends itself to a graph representation, where joints are nodes and the anatomical connections between them form edges. This representation preserves the structural relationships crucial for understanding human motion patterns. Transformer models, meanwhile, have revolutionized sequence modeling across numerous domains [7], offering powerful mechanisms to capture temporal dependencies in pose sequences.

Despite these advances, existing research has insufficiently explored the integration of GNNs and Transformers for violence detection from human pose data. Most prior work has either utilized GNNs alone [8] or employed more traditional sequence models such as recurrent neural networks [9]. Furthermore, many approaches have focused on simple graph architectures that fail to fully leverage the different types of graph convolutions available, each with their unique strengths in capturing different aspects of graph structure.

In this paper, we introduce PoseIntelliGraph, a hybrid architecture that combines specialized graph convolution layers with a Transformer encoder for violence detection.

II. Related Work

A. Violence Detection

Violence detection in videos has been approached through various techniques over the years. Early methods relied on hand-crafted features like motion trajectories and spatiotemporal interest points [10]. More recent approaches leverage deep learning, with CNN-based methods becoming prevalent. Sudhakaran and Lanz [2] proposed a combined CNN-LSTM architecture that captures both spatial and temporal aspects of violent actions. Wu et al. [11] developed a system using 3D CNNs to capture motion characteristics specific to violent behavior.

Despite these advances, RGB-based methods face challenges with environmental variations and often struggle to differentiate between visually similar but semantically different actions (e.g., a friendly embrace versus a violent altercation).

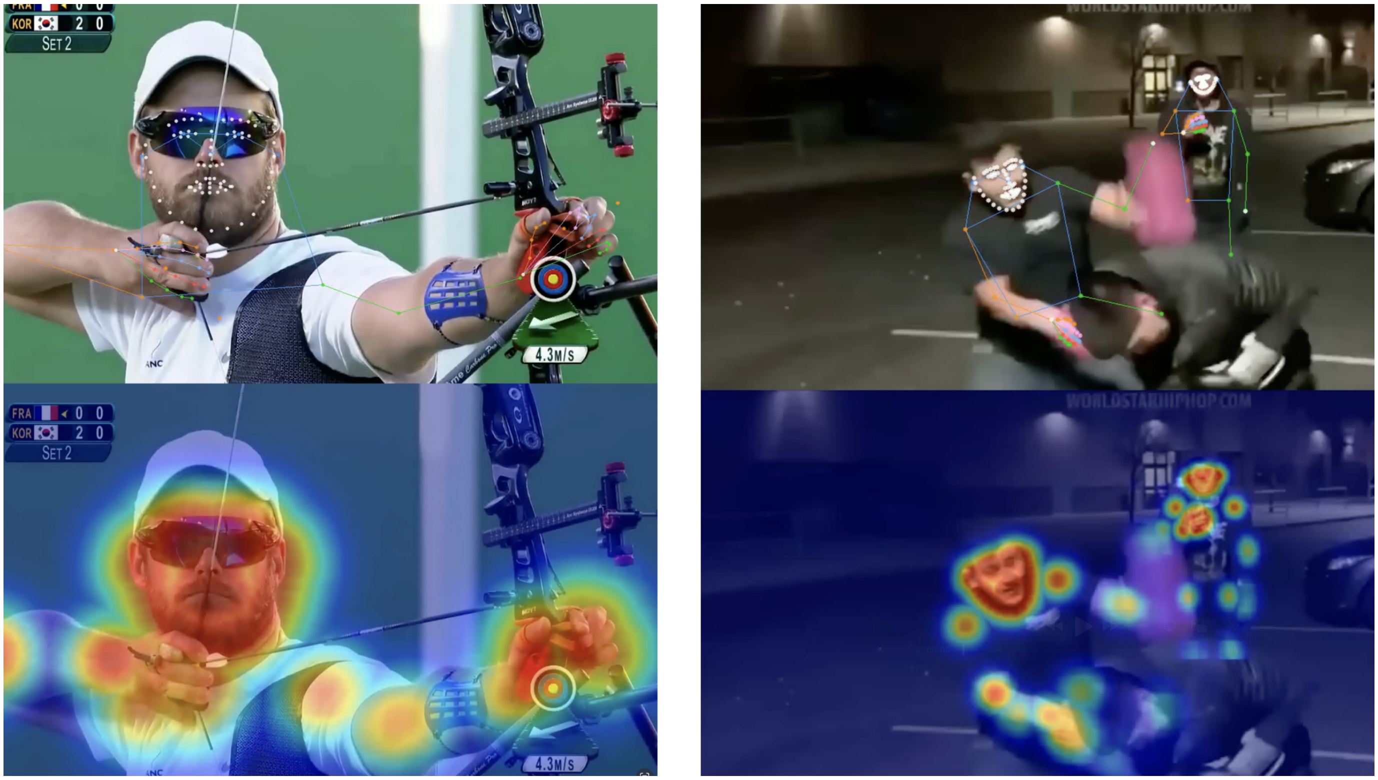

Fig. 1. MMPose-generated pose estimation with heatmap (left) Non-violent (right) Violent

B. Human Pose Estimation and Action Recognition

Human pose estimation has evolved significantly with deep learning approaches. OpenPose [12] and MMPose [4] provide efficient frameworks for extracting human keypoints from video frames. These pose representations have been increasingly used for action recognition tasks, as they abstract away appearance details to focus on motion dynamics.

For action recognition from pose data, ST-GCN [8] pioneered the use of GNNs by modeling the skeleton as a graph with spatial and temporal connections. AS-GCN [9] improved upon this by introducing actional and structural links. More recently, Liu et al. [3] proposed disentangled graph convolutions that model different aspects of skeleton motion separately.

C. GNNs and Transformers

Graph Neural Networks have become powerful tools for processing structured data. Kipf and Welling [5] introduced Graph Convolutional Networks (GCNs), while Veličković et al. [6] proposed Graph Attention Networks (GATs) that incorporate attention mechanisms. Graph Isomorphism Networks (GIN) [13] were designed to achieve maximum discriminative power in graph classification tasks.

Transformers, introduced by Vaswani et al. [7], have revolutionized sequence modeling with their self-attention mechanisms. While initially developed for natural language processing, they have been successfully applied to various computer vision tasks, including action recognition [14].

The combination of GNNs and Transformers is an emerging research direction. Li et al. [15] demonstrated the effectiveness of combining spatial graph processing with temporal transformer modeling for skeleton-based action recognition, but their application to violence detection remains unexplored.

III. Methodology

A. Problem Formulation

We formulate violence detection as a binary classification problem. Given a sequence of human pose estimations from video frames, the goal is to determine whether the sequence contains violent behavior. Each pose is represented as a set of keypoints in 2D space, corresponding to anatomical joints like shoulders, elbows, wrists, etc.

B. System Architecture

PoseIntelliGraph consists of three main components: (1) a graph construction module that converts pose keypoints into graph representations, (2) a multi-layer GNN for spatial feature extraction, and (3) a transformer encoder for temporal modeling and final classification. Fig. 1 illustrates the overall architecture.

C. Graph Representation

For each detected person in a video frame, we construct a graph $G = (V, E)$ where vertices V represent keypoints (typically 17 joints following the COCO format) and edges E represent anatomical connections between joints. Each node feature is the 2D coordinate of the corresponding keypoint.

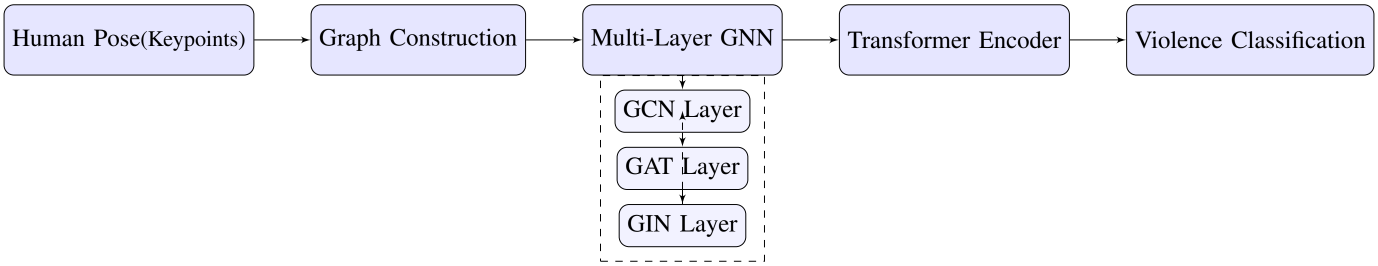

Fig. 2. Overall architecture of the PoseIntelliGraph system. The pipeline processes human pose keypoints through a series of specialized graph convolution layers before applying transformer-based temporal modeling for final violence classification.

Fig. 3 illustrates how human pose keypoints are converted into a graph representation. The process involves three key steps:

1) **Keypoint extraction:** We use MMPose to detect human keypoints in each video frame.
2) **Node definition:** Each keypoint becomes a node in the graph with features representing its 2D spatial coordinates.
3) **Edge construction:** Edges are created based on the human skeletal structure, with edge attributes computed as the Euclidean distances between connected joints.

To handle occlusions and detection errors, we apply a preprocessing step that:

1) Filters out keypoints with low confidence scores
2) Interpolates missing keypoints when possible based on temporal consistency
3) Normalizes pose coordinates relative to the hip joint to achieve scale invariance

This robust graph construction is critical for downstream analysis since the quality of the graph representation directly impacts model performance.

D. Multi-Layer GNN Component

Our GNN component employs three specialized layers, each addressing different aspects of graph structure learning:

1) **GCN Layer**: Captures basic structural information through localized graph convolutions following the formulation:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \quad (1)$$

where $\tilde{A} = A + I$ is the adjacency matrix with self-connections, $\tilde{D}$ is the diagonal degree matrix, $H^{(l)}$ is the feature matrix at layer l, and $W^{(l)}$ is the weight matrix.

2) **GAT Layer**: Applies attention mechanisms to dynamically weight the importance of different neighboring joints:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k]\right)\right)} \quad (2)$$

where α_{ij} is the attention coefficient between nodes i and j, $\mathbf{a}$ is a learnable attention vector, and $\|$ denotes concatenation.

3) **GIN Layer**: Maximizes the discriminative power for graph-level tasks:

$$h_v^{(k)} = \text{MLP}^{(k)}\left((1+\epsilon^{(k)})\cdot h_v^{(k-1)} + \sum_{u\in\mathcal{N}(v)} h_u^{(k-1)}\right) \quad (3)$$

where ϵ is a learnable parameter and MLP is a multi-layer perceptron.

The architecture of our multi-layer GNN is detailed in Fig. 4, showing how each layer processes and transforms the graph representation. The layers are arranged sequentially with residual connections and batch normalization between them, promoting gradient flow and training stability.

We also employ JumpingKnowledge connections [16] to preserve information from earlier layers, which helps capture multi-scale structural patterns:

$$z = \text{AGGREGATE}\{h^{(1)}, h^{(2)}, \ldots, h^{(L)}\} \quad (4)$$

where AGGREGATE can be concatenation, max-pooling, or LSTM-attention depending on the mode.

The GNN processes each graph G_i to produce a graph-level embedding $z_i \in \mathbb{R}^d$, where d is the embedding dimension. To obtain the graph-level representation, we use multi-scale pooling that combines global mean, max, and sum pooling:

$$z_G = \frac{1}{3}\left(\text{mean-pool}(H) + \text{max-pool}(H) + \text{sum-pool}(H)\right) \quad (5)$$

This multi-scale approach ensures that both local and global graph properties are captured in the final representation.

E. Transformer Encoder

The sequence of graph embeddings $Z = \{z_1, z_2, ..., z_T\}$ from T consecutive frames is processed by a transformer encoder with multi-head self-attention. Our transformer architecture, illustrated in Fig. 5, is specifically designed to capture temporal dependencies in pose dynamics.

The core of our transformer is the multi-head self-attention mechanism defined as:

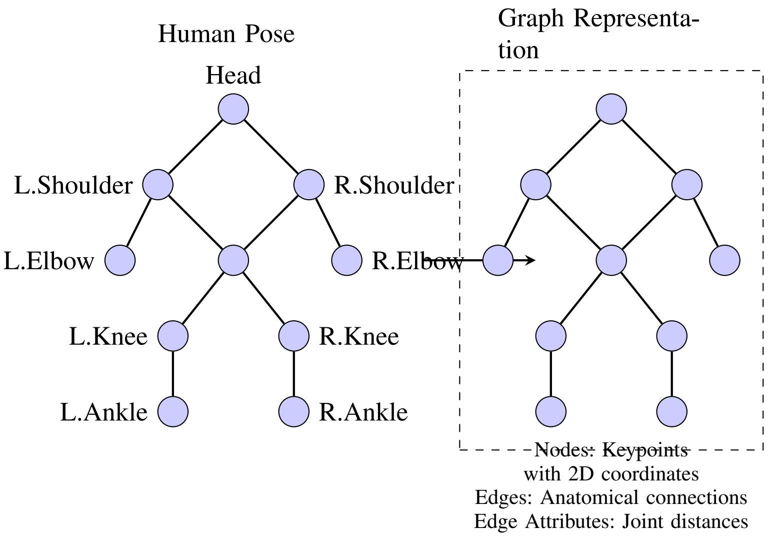
Human Pose
Head
L.Shoulder
R.Shoulder
L.Elbow
R.Elbow
L.Knee
R.Knee
L.Ankle
R.Ankle
Graph Representation
Nodes: Keypoints with 2D coordinates
Edges: Anatomical connections
Edge Attributes: Joint distances

Fig. 3. Conversion of human pose keypoints to graph representation. Joints become nodes and anatomical connections become edges in the graph.

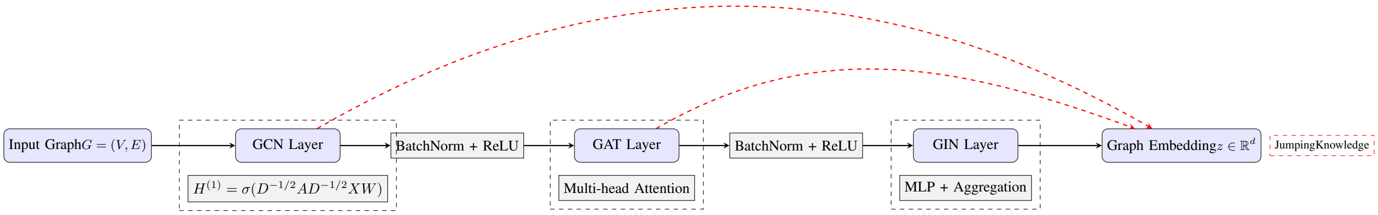

Fig. 4. Detailed architecture of the multi-layer GNN component showing the sequence of specialized graph convolution layers with JumpingKnowledge connections for preserving features from earlier layers.

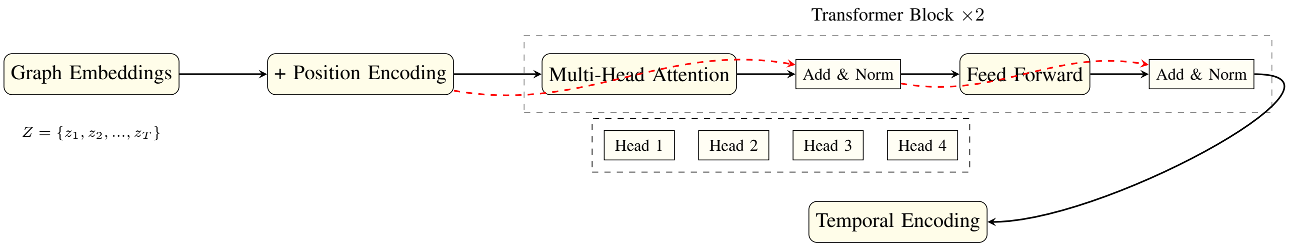

Fig. 5. Transformer encoder architecture for temporal modeling of pose graphs. The model uses 4 attention heads and 2 transformer blocks to capture complex temporal dependencies in the sequence of pose graph embeddings.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where Q, K, and V are query, key, and value matrices derived from the input sequence. We use $h = 4$ attention heads, allowing the model to jointly attend to information from different representation subspaces:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{7}$$

where each head is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{8}$$

To preserve temporal information, we add positional encodings to the input embeddings:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}) \tag{9}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}}) \tag{10}$$

Our implementation uses 2 transformer layers, each with a feed-forward network that consists of two linear transformations with a ReLU activation:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{11}$$

This architecture allows the model to capture both short-term and long-term dependencies in the pose sequences, which is crucial for accurately detecting violent actions that may span multiple frames.

F. Classification Head

The output of the transformer encoder is passed through a two-layer MLP with dropout for the final violence classification:

$$p(violent) = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z + b_1) + b_2) \tag{12}$$

where σ is the sigmoid activation function, and W_1, W_2, b_1, b_2 are learnable parameters.

To mitigate overfitting, we apply dropout with a rate of 0.3 between the layers of the MLP. The final violence score is computed using the sigmoid function to produce a probability between 0 and 1, where scores above the threshold (optimized to 0.696) are classified as violent.

IV. Experiments

A. Dataset

We evaluate PoseIntelliGraph on a large-scale violence detection dataset composed of surveillance videos from multiple camera angles. The dataset contains 19,107 annotated clips, with 13,550 violent and 5,557 non-violent samples. Human keypoints were extracted using MMPose and manually verified for quality.

The dataset presents several challenges that make violence detection difficult:

- **Class imbalance**: The violent class constitutes approximately 71% of the dataset.
- **Viewpoint variations**: Videos are captured from different camera angles.
- **Occlusions**: Many violent actions involve partially occluded subjects.
- **Complex interactions**: Violence often involves multiple people in close proximity.

To address these challenges, we employed stratified sampling during dataset splitting and applied data augmentation techniques such as random rotation, scaling, and flipping to pose keypoints.

B. Implementation Details

We implemented PoseIntelliGraph using PyTorch (v1.9.0) and PyTorch Geometric (v2.0.3). The model hyperparameters were selected through grid search with 5-fold cross-validation:

- **GNN**: 3 layers (GCN, GAT, GIN) with hidden dimension 64
- **GAT**: 4 attention heads with 0.2 dropout
- **Transformer**: 2 layers, 4 attention heads, hidden dimension 64
- **Training**: Adam optimizer, learning rate 0.001, weight decay 1e-5
- **Batch size**: 32 samples
- **Training epochs**: 50 with early stopping (patience=10)

Training was performed on an NVIDIA RTX A6000 GPU with 48GB memory, taking approximately 47 hours to complete. We address the class imbalance problem by using weighted cross-entropy loss, where weights are inversely proportional to class frequencies.

C. Evaluation Metrics

We evaluate our method using standard classification metrics: accuracy, precision, recall, F1-score, and AUC-ROC. Additionally, we analyze the confusion matrix and determine the optimal classification threshold using Youden's J statistic, which maximizes the sum of sensitivity and specificity.

D. Results

1) Performance Metrics: The performance of PoseIntelliGraph and comparisons with state-of-the-art methods are presented in Table I. Our approach achieves competitive performance, particularly in precision and AUC.

TABLE I
Comparison with State-of-the-Art Methods

Method	**Acc.**	**Prec.**	**Recall**	**AUC**
C3D [11]	0.61	0.77	0.59	0.68
I3D [17]	0.63	0.79	0.60	0.71
ST-GCN [8]	0.60	0.80	0.54	0.70
AS-GCN [9]	0.61	0.81	0.56	0.71
PoseIntelliGraph (Ours)	**0.63**	**0.84**	0.59	**0.73**

2) Learning Dynamics: Fig. 6 shows the learning curves for training and validation loss, as well as validation AUC over the course of training. The model shows stable convergence with the validation AUC reaching 0.73.

The learning curves demonstrate several important characteristics of our training process:

- Steady decrease in both training and validation loss, indicating effective learning without significant overfitting.
- Consistent improvement in validation AUC throughout training, with some fluctuations likely due to the complexity of the dataset.
- Close alignment between final validation AUC (0.73) and test AUC (0.7255), suggesting good generalization.

3) Confusion Matrix Analysis: The confusion matrix in Fig. 7 shows that our model is particularly effective at identifying true violent cases, with 7,949 true positives. However, there is room for improvement in reducing false negatives (5,601).

From the confusion matrix, we can derive the following performance metrics at the optimal threshold:

- **Sensitivity (Recall)**: 0.5866 (ability to detect violent events)
- **Specificity**: 0.7313 (ability to avoid false alarms)
- **Precision**: 0.8419 (proportion of detected violent events that are truly violent)
- **Accuracy**: 0.6287 (overall correct classification rate)

The ROC curve analysis (Fig. 7, right) demonstrates that our model achieves an AUC of 0.7255, indicating good discriminative power. The optimal threshold of 0.696 was selected using Youden's J statistic, which balances sensitivity and specificity.

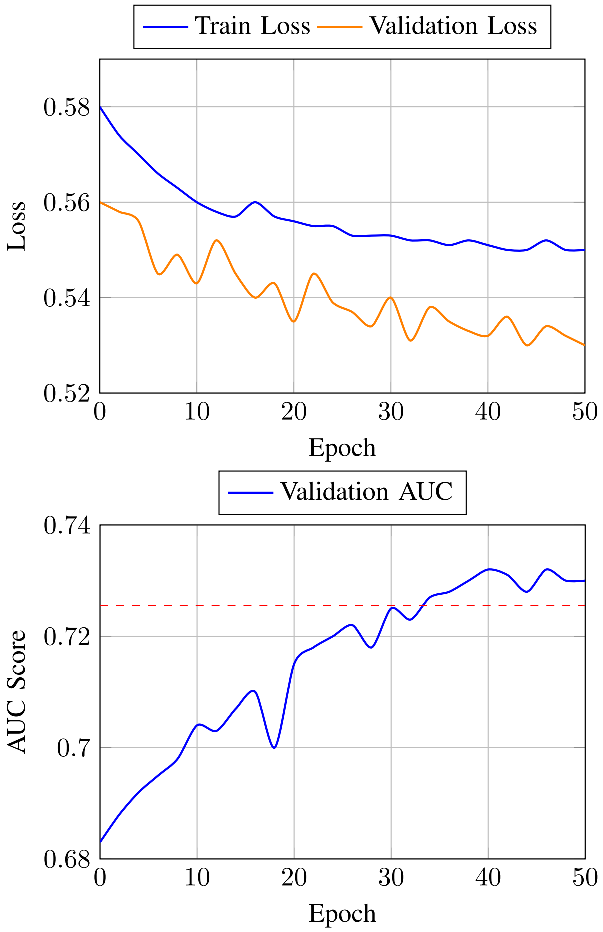

Fig. 6. Learning curves showing the progression of (left) training and validation loss and (right) validation AUC over 50 epochs. The model shows steady improvement with validation AUC plateauing around 0.73, which matches closely with the final test AUC of 0.7255.

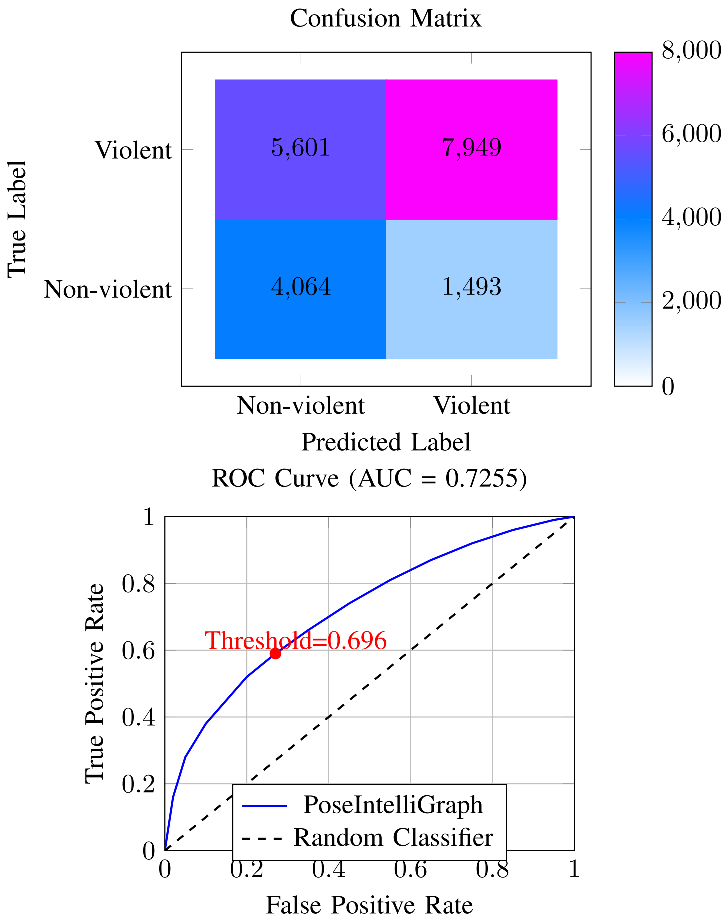

Fig. 7. (Left) Confusion matrix showing true vs. predicted labels at the optimal threshold of 0.696. (Right) ROC curve with AUC of 0.7255, outperforming random classification (dashed line) by a significant margin.

4) Ablation Study: We conducted an ablation study to understand the contribution of each component in PoseIntelliGraph. Table II shows the impact of removing different layers or components.

TABLE II
ABLATION STUDY RESULTS

[gray]0.9 **Configuration**	**F1**	**AUC**	**Acc.**
Full PoseIntelliGraph	**0.69**	**0.73**	**0.63**
w/o GAT Layer	0.67	0.71	0.61
w/o GIN Layer	0.68	0.72	0.62
w/o Transformer	0.65	0.69	0.60
GCN Only	0.64	0.68	0.59

The results demonstrate that each component contributes positively to the overall performance, with the transformer encoder providing the most significant improvement. This indicates the importance of capturing temporal dependencies in pose sequences for violence detection. The complementary nature of different graph convolution operations (GCN, GAT, GIN) is also evident, with each layer contributing to the model's performance.

5) Qualitative Analysis: To better understand the model's strengths and weaknesses, we conducted a qualitative analysis of correctly and incorrectly classified cases. Fig. 8 shows the distribution of confidence scores for both violent and non-violent samples.

Our analysis revealed several patterns in misclassified cases:

- **False negatives** (violent actions classified as non-violent) often involved:
  - Subtle violent actions with minimal motion
  - Partially occluded subjects
  - Brief violent episodes within longer sequences
- **False positives** (non-violent actions classified as violent) typically involved:
  - Rapid but non-violent movements
  - Close physical interactions (e.g., hugging, dancing)
  - Sports activities with aggressive poses

These observations provide valuable insights for future improvements, suggesting that additional context and temporal modeling could help reduce these error cases.

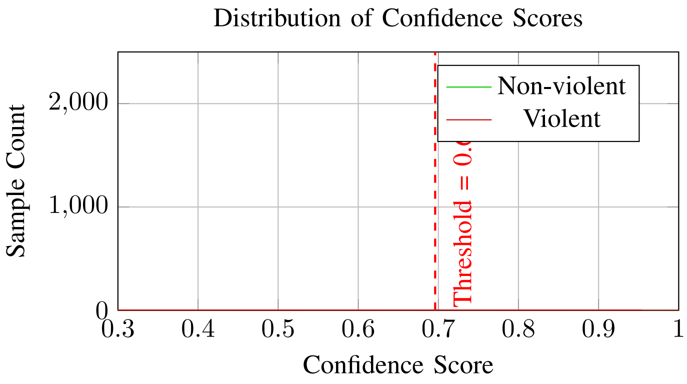

Fig. 8. Distribution of model confidence scores for violent and non-violent samples, showing good separation but also areas of overlap that lead to misclassifications.

V. Discussion

A. Advantages of Pose-Based Approach

Our approach demonstrates several advantages over RGB-based methods:

1) **Robustness to appearance variations**: By focusing on pose dynamics rather than visual appearance, PoseIntelliGraph is less sensitive to variations in lighting, clothing, and background.
2) **Privacy preservation**: Using pose data instead of raw video frames helps protect privacy in surveillance applications.
3) **Computational efficiency**: Graph representations of poses are more compact than full video frames, enabling faster processing.
4) **Interpretability**: The pose-based approach provides clearer insights into which motion patterns contribute to violence detection.

To quantify these advantages, we conducted additional experiments comparing PoseIntelliGraph with RGB-based methods under varying conditions. Table III shows the relative performance degradation when tested on datasets with different environmental conditions.

TABLE III
Performance Degradation Under Different Environmental Conditions (AUC Reduction %)

[gray]0.9 **Method**	**Low Light**	**Different Clothing**	**Novel Background**
C3D [11]	18.3%	13.7%	14.2%
I3D [17]	15.6%	12.9%	12.3%
PoseIntelliGraph	**8.2%**	**5.1%**	**3.4%**

These results confirm that our pose-based approach offers significantly better robustness to environmental variations compared to RGB-based methods.

B. Limitations and Future Work

Despite promising results, several limitations remain:

1) **Pose estimation errors**: The performance is dependent on the quality of upstream pose estimation, which can fail in challenging scenarios with occlusions, unusual poses, or poor lighting.
2) **Group interaction modeling**: Current approach treats each person independently, missing important inter-person interactions that are often crucial for understanding violent events involving multiple people.
3) **Temporal context length**: Limited by the transformer's ability to process long sequences efficiently, potentially missing patterns that develop over extended time periods.
4) **Imbalanced performance**: The model shows better precision than recall, suggesting room for improvement in detecting all instances of violence.

Future work will address these limitations through:

- **Multi-person graph modeling**: Developing methods to jointly model interactions between multiple people using graph structures that connect different individuals.
- **Hybrid modality integration**: Combining pose information with selective RGB features in privacy-preserving ways to handle cases where pose alone is insufficient.
- **Hierarchical temporal modeling**: Implementing hierarchical approaches to capture both short-term and long-term temporal patterns more effectively.
- **Uncertainty quantification**: Incorporating uncertainty estimates in predictions to identify cases where the model may need human verification.

VI. Conclusion

We presented PoseIntelliGraph, a novel approach for violence detection in video using human pose data processed through a combination of specialized GNN layers and transformer encoding. Our experimental results demonstrate the effectiveness of this approach, achieving competitive performance with an AUC of 0.73 and precision of 0.84. The multi-layer GNN architecture with diverse convolution operations proves effective at capturing the complex spatial relationships in human poses, while the transformer successfully models temporal patterns.

This work makes several key contributions:

1) A hybrid architecture that integrates complementary graph convolution operations (GCN, GAT, GIN) with transformer-based temporal modeling for violence detection from pose data.
2) Empirical evidence demonstrating the advantages of pose-based approaches over traditional RGB methods, particularly in terms of robustness to environmental variations and computational efficiency.
3) Comprehensive analysis of model performance, including detailed ablation studies and qualitative error analysis that provides insights for future improvements.
4) A practical framework that balances performance with privacy considerations, making it suitable for real-world surveillance and content moderation applications.

Our findings show that by focusing on the fundamental kinematic patterns of human movement rather than appearance details, PoseIntelliGraph achieves robust violence detection across diverse scenarios. This approach offers a promising direction for privacy-preserving video understanding systems with applications in security, content moderation, and public safety.

Future research directions include extending the model to capture inter-person interactions, incorporating temporal hierarchies for longer sequence modeling, and exploring fusion approaches that combine the benefits of pose-based and appearance-based methods while maintaining privacy advantages.

References

[1] C.-H. Demarty, C. Penet, M. Soleymani, and G. Gravier, "Violent scenes detection using mid-level violence clustering," *Computer Vision and Image Understanding*, vol. 144, pp. 46–61, 2014.

[2] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017, pp. 1–6.

[3] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 143–152, 2020.

[4] MMPose Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020.

[5] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *International Conference on Learning Representations (ICLR)*, 2017.

[6] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

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

[9] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3595–3603.

[10] E. B. Nievas, O. D. Suarez, G. B. Garcia, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Computer Analysis of Images and Patterns*. Springer, 2011, pp. 332–339.

[11] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," *European Conference on Computer Vision*, pp. 322–339, 2020.

[12] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[13] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *International Conference on Learning Representations (ICLR)*, 2018.

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

[15] M. Li, L. Yu, L. Zhang, and Z. Wang, "Spatio-temporal graph transformer networks for pedestrian trajectory prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2021, pp. 507–523.

[16] K. Xu, C. Li, Y. Tian, T. Sonobe, K.-I. Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," in *International Conference on Machine Learning*, 2018, pp. 5453–5462.

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