Computer Vision for Crime Recognition Based on Skeleton Trajectories

AVERCHENKO ILLYA, University of Twente, The Netherlands

Fig. 1. CCTV Surveillance.

Abstract— Given the substantial amount of data generated daily by surveillance systems in urban areas, there is a growing necessity for automation in the crime detection process. Considering the limitations of the current approaches to detecting crime in surveillance videos, there is a need for a new approach that helps reduce human labor and its decision-making ability to ensure the safety of the public. The objective of this research to evaluate the accuracy of skeleton-based action recognition models within the crime domain and use the HR-Crime dataset as the reference point for comparison with other modalities.

Additional Key Words and Phrases: Deep-learning, PyTorch, Skeleton-based action recognition, Crime recognition, Surveillance videos, Human behavior analysis, PoseConv3D, ST-GCN++

1 INTRODUCTION

The increase in urban population has made it challenging to supervise and keep a check on high-risk crime zones, leading to more crime and insecurity in such areas [1]. Efficient and precise detection of criminal activity is crucial for ensuring the safety of residential areas. In general, any human activity can be recognized by means of appearance, depth, optical flows, and body skeletons.

Various approaches such as the spatial-temporal transformer-based model [2], encoder-decoder RNN model [3], CNN based methods [4] have been used to identify criminal or anomalous activity. These methods have several drawbacks: they are computationally intensive, they can not cover complete video understanding tasks, lack accuracy due to low quality of input data or lighting[2][8].

Despite the existence of AI models designed to identify anomalous activity in videos, they either suffer from inadequate accuracy or are irrelevant to the crime domain. Recently, new methods based on skeleton-based action recognition were utilized and presented competitive performance [5][6][7]. The essence of the skeleton-based method is to analyze human joints and classify their behavior by means of deep learning. Skeleton-type action recognition model has several advantages over other models, such as robustness, scalability, compactness, noise immunity, and lightweight modality[6].

In this research, the two new approaches for action recognition, namely PoseConv3D [6], and ST-GCN++ [7], will be evaluated and compared to the previous approaches in the crime detection domain. PoseConv3D and ST-GCN++ are part of the PYSKL - open source toolbox for skeleton-based action recognition based on PyTorch and MMAction2[13]. The two approaches will be evaluated on the HR-Crime dataset [12]. HR-Crime is a collection of YouTube videos, where each video is filmed in a different location. HR-Crime consists of normal videos and 13 categories of human-related crimes. The HR-Crime dataset is a subset of UCF-Crime. Matei et al. 2022 [11] mentioned that the current crime recognition approach could be improved by feeding the HR-Crime dataset with more data and labels that outline the anomalous activities more accurately can improve the model performance. However, taking into consideration annotation resources which can be costly, there is a need to investigate further various approaches that can outperform the current developments without requiring additional data annotation. Addi- tionally, Matei et al. 2022 [11] indicate that dataset imbalance has a significant influence on the complexity of the crime classification task.

This research paper investigates whether state-of-the-art models with novel algorithms in skeleton-based action recognition can mitigate the limitations of previous approaches and be applicable in real-life applications. This research aims to address the following questions:

1.1 Research Questions

This paper will answer the following main research question:

RQ1: How well do the state-of-the-art skeleton-based action recognition models, namely PoseConv3D and ST-GCN++, perform on the HR-Crime dataset?

TScIT 39, July 2, 2023, Enschede, The Netherlands

To support the main research question and explore potential areas for improvement, the following sub-questions will be answered:

RQ2: How does the balancing of the HR-Crime dataset impact the classification accuracy of the model?

RQ3: How the HR-Crime dataset can be further improved?

2 RELATED WORK

In order to gather related literature to the research domain ScienceDirect, arXiv, and IEEE were used. With search terms such as "skeleton-based action recognition", "skeleton trajectories", "crime recognition", "anomaly detection" and "pose estimation" several documents could be found that have done research in these fields.

In 2022, Talavera et al. [2] proposed a transformer-based model that relies on the spatial-temporal representation of extracted skeletal trajectories for fine-grained classification. The model was built on top of the PoseFromer - 3D human pose estimator. The model was evaluated on the HR-Crime dataset achieving a balanced accuracy of 49%. However, there was a lack of in-depth information with regard to the relation of action and classification of crime.

In 2019, Morais et al. [10] proposed the MPED-RNN architecture for anomaly detection in surveillance videos based on skeleton trajectories.MPEDRNN achieves competitive performance and is highly interpretable. However, MPEDRNN performed well on the HR-ShanghaiTech dataset that 1) contains video of the same location; 2) contained motions like jumping and running; and so the model detected less accurately on the HR-Crime dataset that has a different setup and type of actions.

In 2022, Matei et al. [11] proposed a trajectory-based crime classification framework based on MPED-RNN. The research studied whether specific human body movements correspond to a particular crime category. The research showed that human skeletal trajectory analysis is a feasible approach to crime-related anomaly classification. The research concluded that analyzing the trajectory of human skeletal movement can be a viable method for classifying anomalies related to criminal activity. Additionally, the impact of dataset imbalance of HR-Crime was examined and data augmentation techniques and extension of the HRCrime dataset were suggested for boosting the classification performance of the model.

3 METHODOLOGY

The research methodology used is experimental and focuses on observing the behavior of two PYSKL[14] models, namely PoseConv3D [6] and ST-GCN++[7], in various scenarios. The main objective is to assess the applicability of these approaches in the field of crime.

The study examines the impact of various data augmentation techniques on the performance of the models. Furthermore, it investigates how the models perform when trained solely on anomaly classes compared to being trained on both anomaly and normal classes.

The main steps of this paper include dataset preparation, skeleton extraction, model configuration, and evaluation.

3.1 Datasets

To assess the effectiveness of the two architectures, their performance is measured using a subset of the HR-Crime dataset, which comprises 13 anomaly classes that represent different criminal activities, along with a normal class where no criminal activity is present. The initial annotations of the HR-Crime dataset comprise a total of 1571 videos.

It is evident from these numbers that there is a significant class imbalance issue in the dataset, as the number of normal videos is more than ten times greater than the other classes. The other prevalent classes, besides the normal class, are Robbery, Stealing, and Burglary. Because of the limitations of time and graphical processing resources, this research utilizes only half of the HR-Crime dataset.

Three datasets are created in the following manner:

(1) Dataset 1 consists of 811 video annotations and includes randomly selected videos from the HR-Crime dataset (fig. 4 in Appendix).
(2) Dataset 2 contains 840 video annotations with balanced sampling, ensuring that each class is represented by 60 videos
(3) Dataset 3 consists of 780 video annotations exclusively from anomalous video classes, without any normal videos. This dataset is also balanced, with 60 videos per anomaly class.

The three datasets are utilized to compare the performance of the models: PoseConv3d, ST-GCN++, Spatial-Temporal Transformer [2], Encoded-based classifier [11], Deconded-based classifier [11], Encoder-decoder architecture [12]. The PYSKL models are trained and tested using an 80:20 ratio.

3.2 Skeleton extraction with HRNet

Prior to training the models, preprocessing of the input data is essential, as both of the models require skeleton information - 2D poses. To accomplish this, HRNet2D was used to extract 2D joint coordinates, consisting of 17 keypoint pairs (x,y coordinates), and keypoint scores from the videos - confidence scores of the keypoints. HRNet 2D is the pose extractor pre-trained on the COCO key points. It takes video frames as input and outputs a sequence of coordinates that represent human positioning through the frames of a video. After, the skeleton annotations and the train/test splits are merged into a single pickle file which is used in model training and testing.

The example of the extracted poses is presented in figure 3.

3.3 PYSKL Models

The configuration settings for the models were derived from examples provided in the PYSKL library. These configurations were originally utilized to train the PoseConv3d and ST-GCN++ models on the UCF101 and NTURGB+D video datasets respectively. The adjustments made to these configurations included modifying hyperparameters such as the number of epochs, number of classes, learning rate, and clip length. The clip length determines the number of frames sampled in each clip used for training, validation, and testing. In this case, for the PoseConv3D and ST-GCN++ models, every 6th frame of the video was sampled, resulting in a loss of significant contextual information. With access to more processing units and memory, it would be possible to utilize all frames of the videos and capture more comprehensive contextual information.

3.3.1 *PoseConv3D.*

The PoseConv3D model is a 3D-CNN that utilizes 2D pose estimations from multiple views to reconstruct the 3D pose, shape, and motion of a human subject. It consists of several key components:

1. Volumetric Representation: PoseConv3D takes 2D pose estimations (joint coordinates) as input. The 2D pose estimations from different views are then used to create a volumetric representation of the human subject. This is achieved by projecting the 2D joint locations back into 3D space and encoding them as heatmaps or occupancy grids.
2. Pose Convolutional Networks: PoseConv3D operates on the volumetric representation. It uses 3D convolutional layers to capture spatial relationships and dependencies between joints across different views.
3. Pose Reconstruction: The pose convolutional networks process the volumetric representation to estimate the 3D pose, shape, and motion of the human subject. This is achieved by regressing the joint locations, body shape parameters, and temporal offsets from the volumetric features.

The overview of the PoseConv3D pipeline is presented in fig. 6. The backbone of the PoseConv3D is SlowOnlyR50 - a skeleton-based action recognition model. The state-of-the-art PoseConv3D achieves top accuracy of 0.86 on UCF101 dataset [19].

3.3.2 *STGCN++.*

ST-GCN++ builds upon the original ST-GCN model and incorporates improvements to enhance its performance in capturing spatial and temporal dependencies within human skeleton data. The main components of the ST-GCN++ are the following:

1. The human skeleton is represented as a graph, where each joint is treated as a node, and the edges represent the spatial connections between joints. The graph structure enables capturing spatial relationships between body parts.
2. ST-GCN++ utilizes spatial-temporal graph convolutional layers to learn and extract features from the skeleton data. These layers perform graph convolutions on the skeleton graph to capture spatial and temporal dependencies among the joints.
3. After feature extraction, ST-GCN++ employs a Softmax classifier, such as fully connected layers, to map the learned features to specific action classes for recognition.

The overview of the ST-GCN++ pipeline is presented in fig. 7. The state-of-the-art ST-GCN++ achieves top accuracy of 97.4% on the NTURGB+D dataset.

3.4 Evaluation

The models are assessed using metrics such as top1_accuracy, top5_accuracy, F1-score, AUROC (Area Under the Receiver Operating Characteristic curve), and confusion matrix. AUROC score - It tells how much the model is capable of distinguishing between classes.An AUROC score of 1 represents a perfect classifier, while a score of 0.5 indicates a random classifier. Top5_accuracy shows that the correct class gets to be in the top 5 probabilities for it to count as “correctly predicted”.

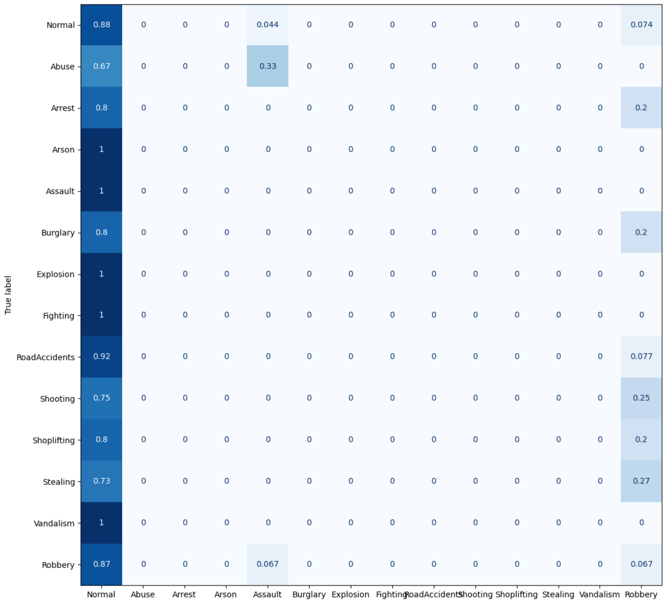


Fig. 2. PoseConv3D-1 confusion matrix

4 RESULTS

This section presents the results of the experiments conducted with the PoseConv3D and ST-GCN++ architectures introduced in Sections 2.3 and 2.4 respectively.

4.1 Dataset 1 - random sampling

4.1.1 *PoseConv3D with random sampling.*

The results of training the PoseConv3D model on Dataset 1 are presented in the confusion matrix (fig. 2). It is evident that the majority of the results were predicted as Normal_videos. Classes such as Arson, Assault, Explosion, Fighting, and Vandalism were classified as normal in 100% of the cases. The reason for these predictions is not entirely clear, but it could be attributed to the fact that the Normal_Videos class dominates the dataset, causing the model to focus on this class more frequently than the others. There are also other observations worth noting. For instance, Abuse was predicted as Assault with a probability of 33%. One possible explanation for this could be that these two types of crimes involve similar skeletal movements, which are more common and densely populated in the coordinate space.

PoseConv3D-1 achieved an accuracy of 0.38, the weighted F1-score of 0.25, and an AUROC score of 0.51. The top5_accuracy is 0.68. The model has the highest accuracy among all six cases, primarily due to the high number of True Positives resulting from correctly predicting most of the Normal_Videos. However, the AUROC score suggests that the model struggles to differentiate between the different crime classes. Although the top1_accuracy initially appears competitive, considering the scores in the confusion matrix, there might be a reason for such a high score. Given that one-third of all

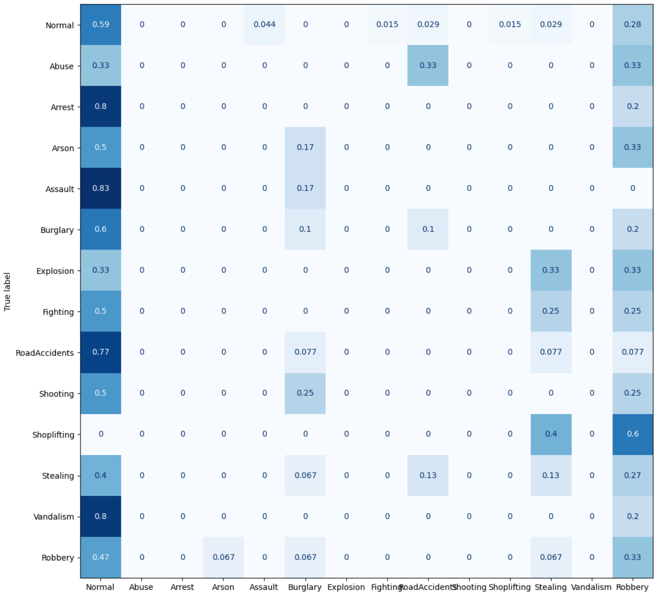


Fig. 3. ST-GCN++ -1 confusion matrix

videos in the dataset were normal videos, the accuracy merely indicates that the majority of normal videos were correctly predicted as normal (88%).

4.1.2 *ST-GCN++ with random sampling.*

The results of the ST-GCN++ model are presented in the confusion matrix (fig. 3). The image differs slightly from the PoseConv3D model, with fewer videos being predicted as normal. There are instances where the predicted labels intersect with the true labels. For example, Robbery was correctly predicted with a probability of 33% out of all predictions for that class. Similarly to the PoseConv3D confusion matrix, the classes Assault, Arrest, and Vandalism were also predicted as normal. Interestingly, the class Shoplifting was predicted as Stealing and Robbery, with percentages of 40% and 60% respectively, in terms of the total predictions for the class. This could be attributed to the similar nature of these crimes. The model's performance is somewhat worse than that of the PoseConv3D model: accuracy is 30%, f1-score is 0.25, and AUROC score is 0.50. The AUROC score indicates that the model distinguishes between the classes completely randomly, without any meaningful pattern.

Table 1. Results of the experiments

Model id	Dataset	top1_acc	top5_acc	F1-score	AUROC
PoseConv3D					
PoseConv3D-1	**random**	**0.3765**	**0.6790**	**0.2470**	0.5176
PoseConv3D-2	balanced	0.0238	0.32739	0.0144	0.4989
PoseConv3D-3	balanced, only anomaly	0.0764	0.3949	0.0357	**0.5262**
ST-GCN++					
STGCN++ -1	**random**	**0.2962**	**0.5740**	**0.2495**	**0.5033**
STGCN++ -2	balanced	0.0714	0.3273	0.0673	0.4796
STGCN++ -3	balanced, only anomaly	0.0445	0.3375	0.0327	0.4571
Previous approaches					
ST-Tran	HR-Crime	**0.4900**	-	**0.6300**	-
MPED-C5.2	HR-Crime	0.3040	0.8160	0.3730	-
MPED-M1.2	HR-Crime	0.3820	**0.8460**	0.4280	-
MPED-RNN	HR-Crime	-	-	-	**0.7346**

4.2 Dataset 2 - balanced sampling

4.2.1 *PoseConv3D with balanced sampling.*

The situation with balanced sampling presents a completely different picture compared to the previous results. Based on the confusion matrix (fig. 8), most of the classes were identified as Road Accidents, with 50% or more cases per class. However, Road Accidents were not correctly identified with their true label in 83% of the cases. The dominant prediction of the Road Accidents class could be attributed to a difference in the number of skeleton annotations. When a person is sitting in a car, they become invisible to the pose estimator, resulting in fewer tracked skeleton trajectories and a lower total number of skeleton key points compared to the annotations of other classes.

The accuracy of the model is only 2.3%. The remaining metrics can be found in Table 1.

4.2.2 *STCGN++ with balanced sampling.*

Similarly to the PoseConv3D model, the ST-GCN++ model with balanced sampling did not perform well either. The accuracy is only 7%. However, the scores in the confusion matrix (fig. 9) are distributed in a different and more chaotic manner, indicating that the model does not distinguish between the classes at all.

4.3 Dataset 3 - balanced sampling only anomaly

4.3.1 *PoseConv3D with balanced sampling and only anomaly videos.*

The results of the model trained on a dataset without normal videos demonstrate similarities to the results discussed in section 3.2. The corresponding confusion matrix is provided in Figure 10. Once again, the majority of the predictions are assigned to the Road Accidents class, particularly for normal videos. However, in this case, the Road Accidents class is correctly classified with a probability of 42%. The AUROC score, which is 0.53, stands as the highest among all the other cases, although it still falls short of the desired performance.

4.3.2 *ST-GCN++ with balanced sampling and only anomaly videos.*

The confusion matrix (fig. 11 in Appendix) shows scattered predictions with a low number of True Positives. The accuracy score of 4,5% is the lowest among the three configurations.

4.4 Comparison with state of the art

The results of the experiments conducted on the two PYSKL models with the three datasets are presented in Table 1. Additionally, the results of the previous approaches are summarised for comparison.
* ST-Tran - Spatipatial-temporal transformer [2].
* MPED-C - Encoded-based classifier (MPED-C5.2) [11].
* M1.2 - Deconded-based classifier (M1.2) [11].
* MPED-RNN - encoder-decoder architecture for anomaly detection in surveillance videos based on skeleton trajectories described by local and global body movement [12].

5 DISCUSSION AND FUTURE WORK

This research paper evaluated the performance of two skeleton-based action recognition models on the HR-Crime dataset. It aimed to determine the most suitable data sampling techniques for these models. The results demonstrate whether the two models can be applied in the crime domain and what could be the possible limitations of the dataset regarding the data quality.

5.1 PoseConv3D on the HR-Crime

PoseConv3D-1 achieves the highest accuracy of 37.6%. The recall of PoseConv3D-1 is 0.376, while the precision is 0.184. The low recall indicates that the model fails to identify most of the true positives and has a higher number of false negatives. In the crime domain, minimizing false negatives is crucial. It is acceptable if a normal video is flagged as anomalous, meaning it contains a crime scene, but it is essential to avoid labeling crime videos as non-crime.

5.2 ST-GCN++ on the HR-Crime

ST-GCN++ -1 achieves the second-highest accuracy rate of 29.6% among the six different model configurations. The recall rate is 0.296 and the precision rate is 0.223. Even though the accuracy score is lower than that of PoseConv3d-1, the confusion matrix reveals that the model's predictions exhibit more diversity across different classes compared to Normal_Videos. For instance, it shows a stronger correlation between the Shoplifting class and Robbery and Stealing, indicating that the model recognizes patterns among these three types of crimes.

To compare the PYSKL models with the previous models (Table 2), accuracy scores, F1-scores, and AUROC scores are considered. ST-Tran [2] demonstrates the highest performance among all the models, with an accuracy of 49% and an F1-score of 0.63. Thus, the PoseConv3D model, which relies on 3D heatmaps as the base representation of human skeletons, is not yet applicable in the crime domain. Nevertheless, both of the PYSKL models demonstrate potential even with limited training data. By employing better augmentation techniques and video pre-processing, the models can achieve improved results. The discussion section outlines future steps for further adaptation of the models to the crime domain.

5.3 Sampling Techniques

Upon analyzing the accuracy scores, it becomes apparent that the models trained on randomly sampled datasets perform better compared to the balanced versions of HR-Crime. The initial expectation was that balancing the dataset would remove bias among classes and result in higher accuracy. However, this approach led to the dispersion of predictions across all classes, thus eliminating the accurate guesses achieved when the Normal_Videos class was dominant. Nevertheless, there might be another reason influencing this behavior, and balanced sampling could still be a valuable sampling strategy once this reason is fixed.

5.4 Dataset quality

During the data preprocessing stage, an overview of the dataset's overall quality was conducted. Some cases were identified that might provide additional insights into the poor performance of the model in specific video classes within the HR-Crime dataset. For instance, in videos, like Abuse038, the focus is on crimes against animals rather than humans. In this video, a dog was the main object of interest, while humans were present but not involved in the crime. Consequently, the model considered it a normal video because it couldn't detect any anomalous human activity. Another observation pertains to the camera's focus and perspective in videos such as Robbery113. The model failed to record accurate skeleton trajectories for individuals located far from the CCTV camera's focus. Additionally, in videos like Robbery113 and Stealing106, poor lighting conditions and shadows prevented the capture of human skeleton movements. Furthermore, in these videos, the individual gets into a car, rendering them invisible for skeleton motion capture by HRNet. These observations lead to two conclusions: first, the HR-Crime dataset needs to undergo revision and filtering; second, considering that some videos reflect real-world conditions such as low-quality CCTV footage, unfavorable perspectives, and inadequate lighting, there is a need to explore alternative approaches for crime recognition. This may involve considering objects with which humans interact, animals, and specific locations or adopting a multimodal approach to improve accuracy and robustness.

5.5 Future work

Model training proved to be challenging due to the need to process a substantial amount of data frames. This limitation arises from both the quantity of videos used for model training and the number of frames utilized in the model. On average, the videos in the HR-Crime dataset have a duration of approximately 3.3 minutes. In the majority of crime videos, a significant portion of the footage is normal human behavior, with no occurrence of anomalous activities.The findings from the models using random sampling indicate a tendency to classify videos as normal, which is consistent with the previously mentioned observation. By reducing the duration of videos and focusing on key stages of the events, the average number of frames can be minimized, allowing for better utilization of frames during model training. Prior to classification, incorporating a detection stage would be necessary to identify and truncate frames that do not align with the annotation label of a video.

The PYSKL models demonstrated good performance on the UCF101 dataset with 87% accuracy, using uniform sampling with 10 clips and other parameters. Since most UCF101 videos are shorter (around 15 seconds) and focused on actions, it is worth reevaluating the annotation strategy to trim the video set to a shorter length, including only the frames capturing the action, and subsequently reevaluating the PYSKL models.

6 CONCLUSION

In this work, the two novel skeleton-based action recognition models, namely PoseConv3D and ST-GCN++ were discussed and evaluated on the HR-Crime subset. To answer the main research question the experiments on three dataset configurations were conducted. PoseConv3D achieved the highest accuracy of 0.37 with the random sampling configuration for the 13 crimes and 1 normal classification.

The obtained accuracy of 0.29 shows that ST-GCN++ indicates that it also holds promise as a competitive model. The two PYSKL models were also compared to the previous implementations of crime recognition. Spatipatial-temporal transformer[2] remains the benchmark model with accuracy of 0.49.

The sub-questions "How does the balancing of the HR-Crime dataset impact the classification accuracy of the model?" and "How the HR-Crime dataset can be further improved?" gave a reason to further analyze the HR-Crime dataset on data quality and propose certain adjustments in terms of the duration of the videos.

Due to the similarity between the content of crime videos and that of normal videos, as well as other crime videos, balancing the dataset did not result in any significant improvements in model performance compared to the random sampling approach.

Regarding dataset improvement, although it was not pursued due to the significant annotation effort and resource requirements involved, the findings from previous approaches, along with the novel ones, suggest the need for adjustments in the HR-Crime dataset. One possible step towards improvement is to trim the videos to focus on key actions and filter out misclassified videos. This initial refinement can serve as the first step towards enhancing the dataset, leading to subsequent evaluations of the novel algorithms.

In conclusion, it was demonstrated that skeleton-based action recognition is a feasible approach to crime classification. Potential areas for future work include modifying the HR-Crime dataset to include shorter videos that specifically capture key activities.

7 REFERENCES

[1] S. Tobon, D. Mejia, and S. Gómez, 'The Deterrent Effect of Surveillance Cameras on Crime'. Rochester, NY, Mar. 24, 2020. doi: 10.2139/ssrn.3560356.

[2] K. Boekhoudt and E. Talavera, 'Spatial-Temporal Transformer for Crime Recognition in Surveillance Videos', in 2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Nov. 2022, pp. 1–8. doi: AVSS56176.2022.9959414.

[3] R. Morais, L. Vong, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, 'Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos', pp. 11988–11996, Jun. 2019, doi: 10.1109/CVPR.2019.01227.

[4] K. Simonyan and A. Zisserman, 'Two-Stream Convolutional Networks for Action Recognition in Videos', arXiv.org, Jun. 09, 2014. https://arxiv-org.ezproxy2.utwente.nl/abs/1406.2199v2

[5] W. Xin, R. Liu, Y. Liu, Y. Chen, W. Yu, and Q. Miao, 'Transformer for Skeleton-based action recognition: A review of recent advances', Neurocomputing, vol. 537, pp. 164–186, Jun. 2023, doi: 10.1016/j.neucom.2023.03.001.

[6] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, 'Revisiting Skeleton-based Action Recognition'. arXiv, Apr. 02, 2022. doi: arXiv.2104.13586.

[7] H. Duan, J. Wang, K. Chen, and D. Lin, 'PYSKL: Towards Good Practices for Skeleton Action Recognition'. arXiv, May 19, 2022. doi: 10.48550/arXiv.2205.09443.

[8] Z. Du, G. Zhang, W. Lu, T. Zhao, and P. Wu, 'Spatio-Temporal Transformer for Online Video Understanding', J. Phys.: Conf. Ser., vol. 2171, no. 1, p. 012020, Jan. 2022, doi: 10.1088/1742-6596/2171.

[9] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. 'Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition'. Proceedings of the AAAI Conference on Artificial Intelligence 32, no. 1 (27 April 2018). https://doi.org/10.1609/aaai.v32i1.12328.

[10] Morais, Romero, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. 'Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos'. arXiv, 17 April 2019. https://doi.org/10.48550/arXiv.1903.03295.

[11] Matei, Alina-Daniela, Estefania Talavera, and Maya Aghaei. 'Crime Scene Classification from Skeletal Trajectory Analysis in Surveillance Settings'. arXiv, 4 July 2022. http://arxiv.org/abs/2207.01687.

[12] Boekhoudt, Kayleigh, Alina Matei, Maya Aghaei, and Estefanía Talavera. 'HR-Crime: Human-Related Anomaly Detection in Surveillance Videos'. CAIP 2021. https://doi.org/10.1007/978-3-030-89131-2_15.

[13] MMAction2 is an open-source toolbox for video understanding based on PyTorch. https://github.com/open-mmlab/mmaction2

[14] PYSKL - toolbox focusing on action recognition based on Skeleton data with PYTorch. https://github.com/kennymckormick/pyskl

8 APPENDIX

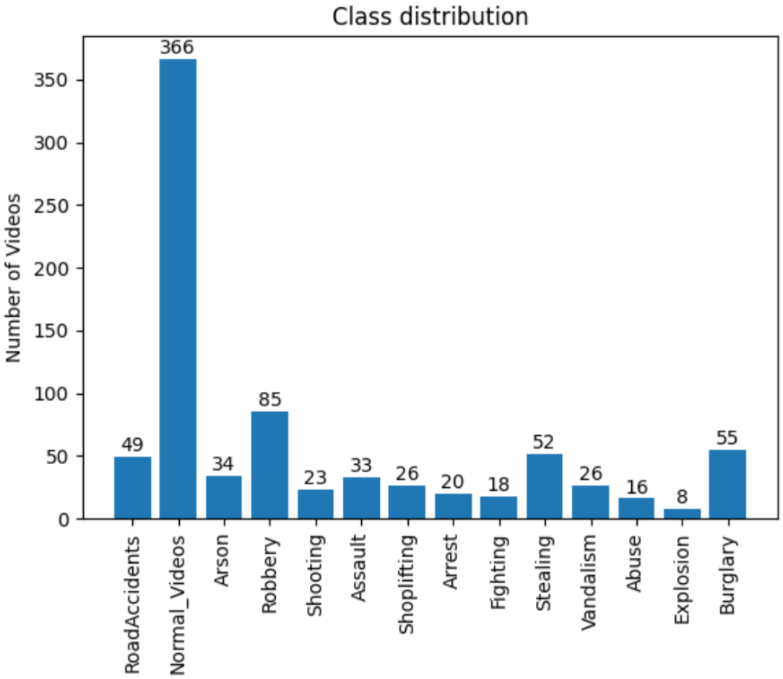


Fig. 4. Class distribution of the dataset 1

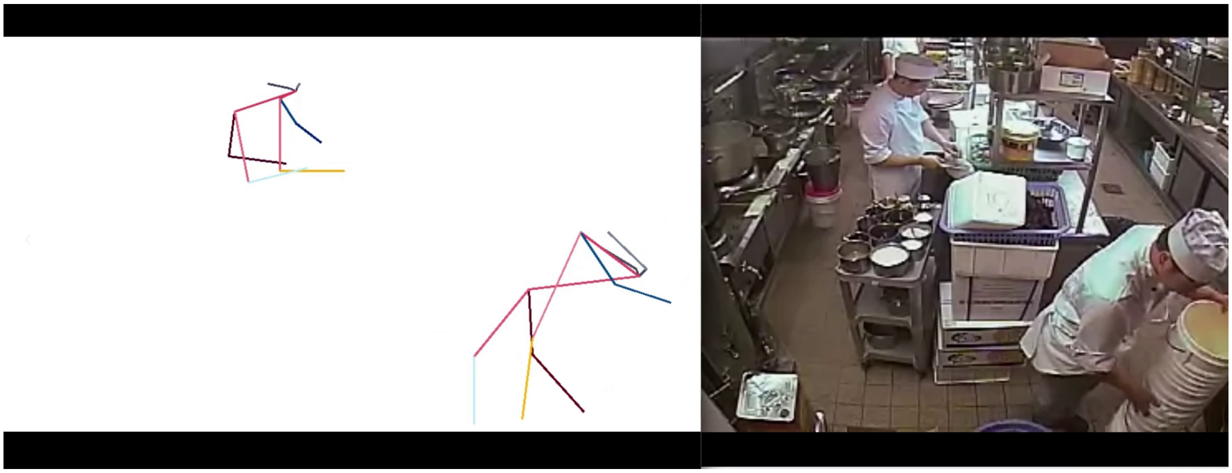


Fig. 5. Example of the skeleton extraction by HRNet

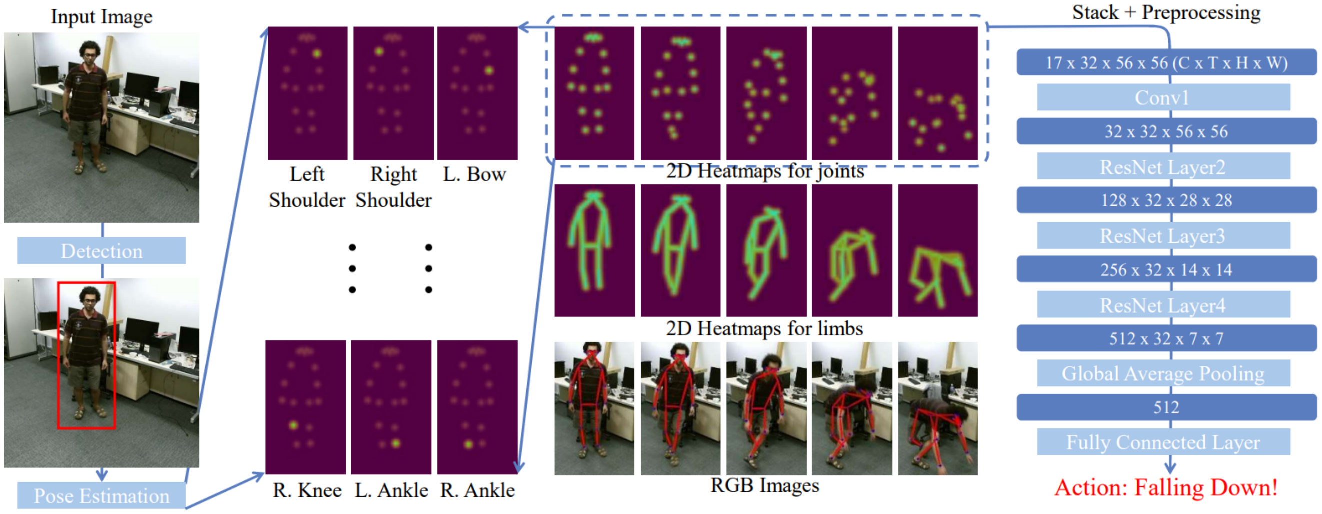


Fig. 6. The pipeline of PoseConv3D

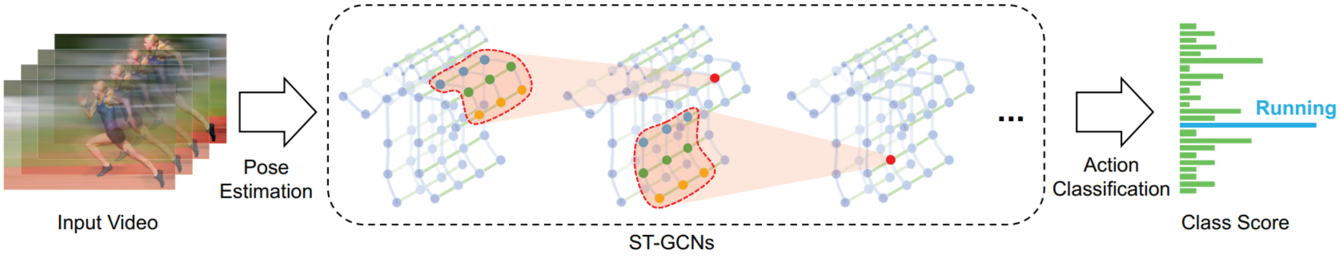
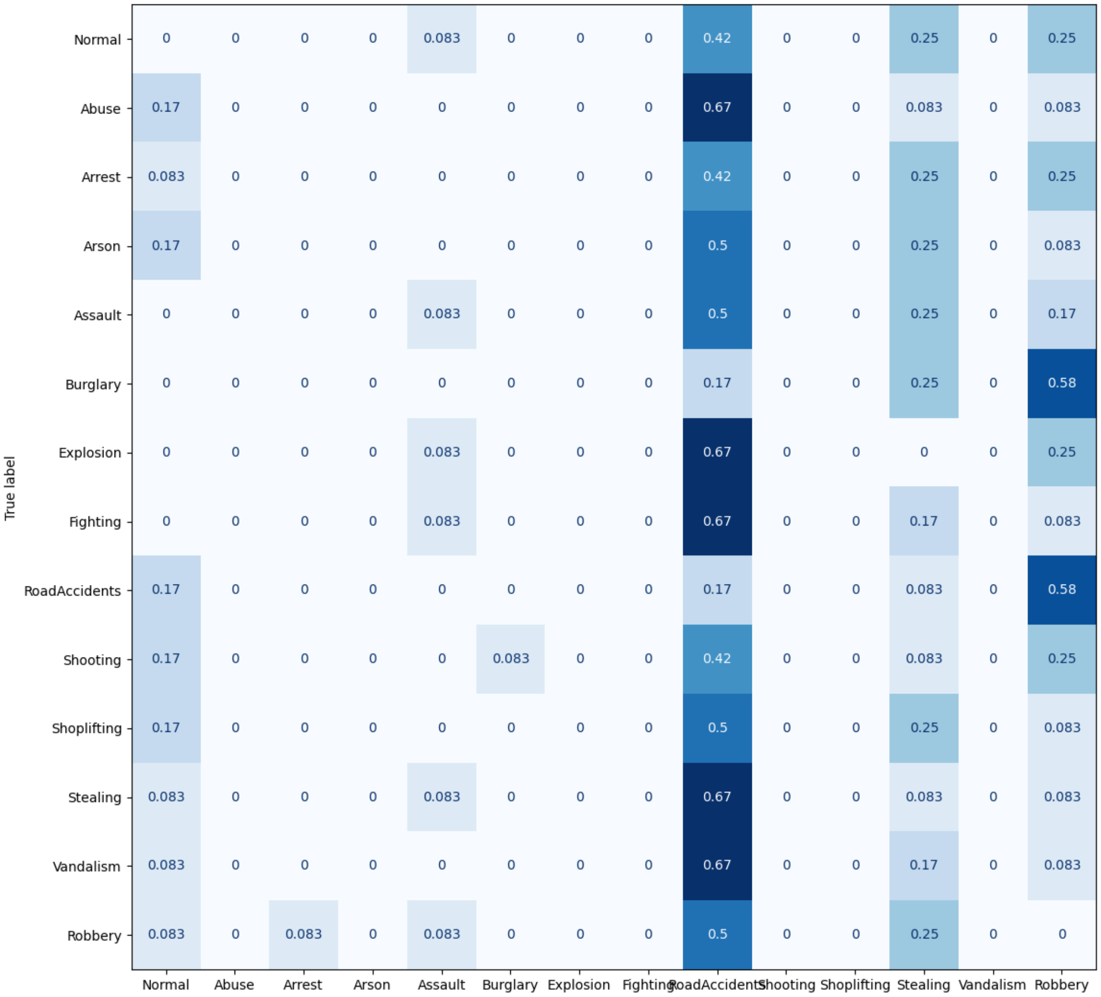


Fig. 7. The pipeline of ST-GCN++



	Normal	Abuse	Arrest	Arson	Assault	Burglary	Explosion	Fighting	RoadAccidents	Shooting	Shoplifting	Stealing	Vandalism	Robbery
Normal	0	0	0	0	0.083	0	0	0	0.42	0	0	0.25	0	0.25
Abuse	0.17	0	0	0	0	0	0	0	0.67	0	0	0.083	0	0.083
Arrest	0.083	0	0	0	0	0	0	0	0.42	0	0	0.25	0	0.25
Arson	0.17	0	0	0	0	0	0	0	0.5	0	0	0.25	0	0.083
Assault	0	0	0	0	0.083	0	0	0	0.5	0	0	0.25	0	0.17
Burglary	0	0	0	0	0	0	0	0	0.17	0	0	0.25	0	0.58
Explosion	0	0	0	0	0.083	0	0	0	0.67	0	0	0	0	0.25
Fighting	0	0	0	0	0.083	0	0	0	0.67	0	0	0.17	0	0.083
RoadAccidents	0.17	0	0	0	0	0	0	0	0.17	0	0	0.083	0	0.58
Shooting	0.17	0	0	0	0	0.083	0	0	0.42	0	0	0.083	0	0.25
Shoplifting	0.17	0	0	0	0	0	0	0	0.5	0	0	0.25	0	0.083
Stealing	0.083	0	0	0	0.083	0	0	0	0.67	0	0	0.083	0	0.083
Vandalism	0.083	0	0	0	0	0	0	0	0.67	0	0	0.17	0	0.083
Robbery	0.083	0	0.083	0	0.083	0	0	0	0.5	0	0	0.25	0	0

Fig. 8. PoseConv3D-2 confusion matrix

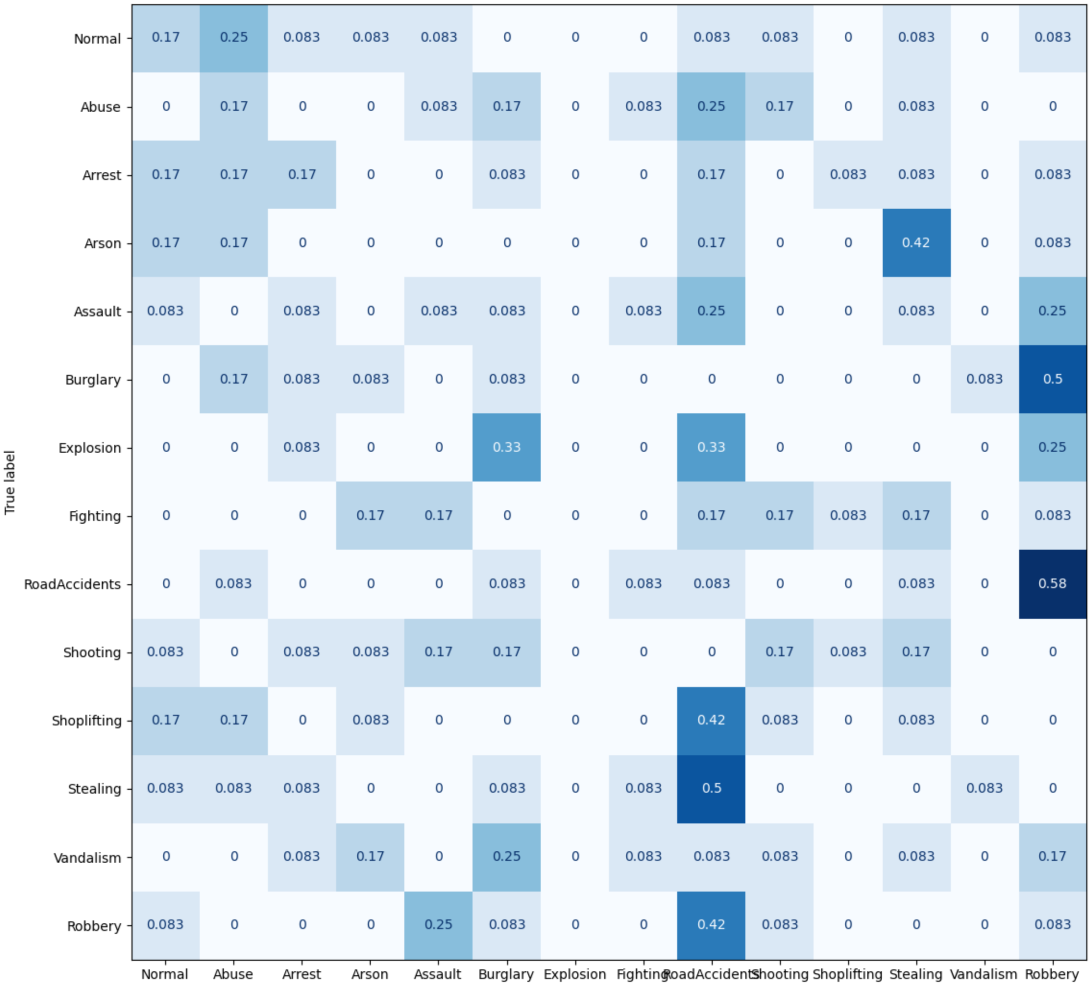
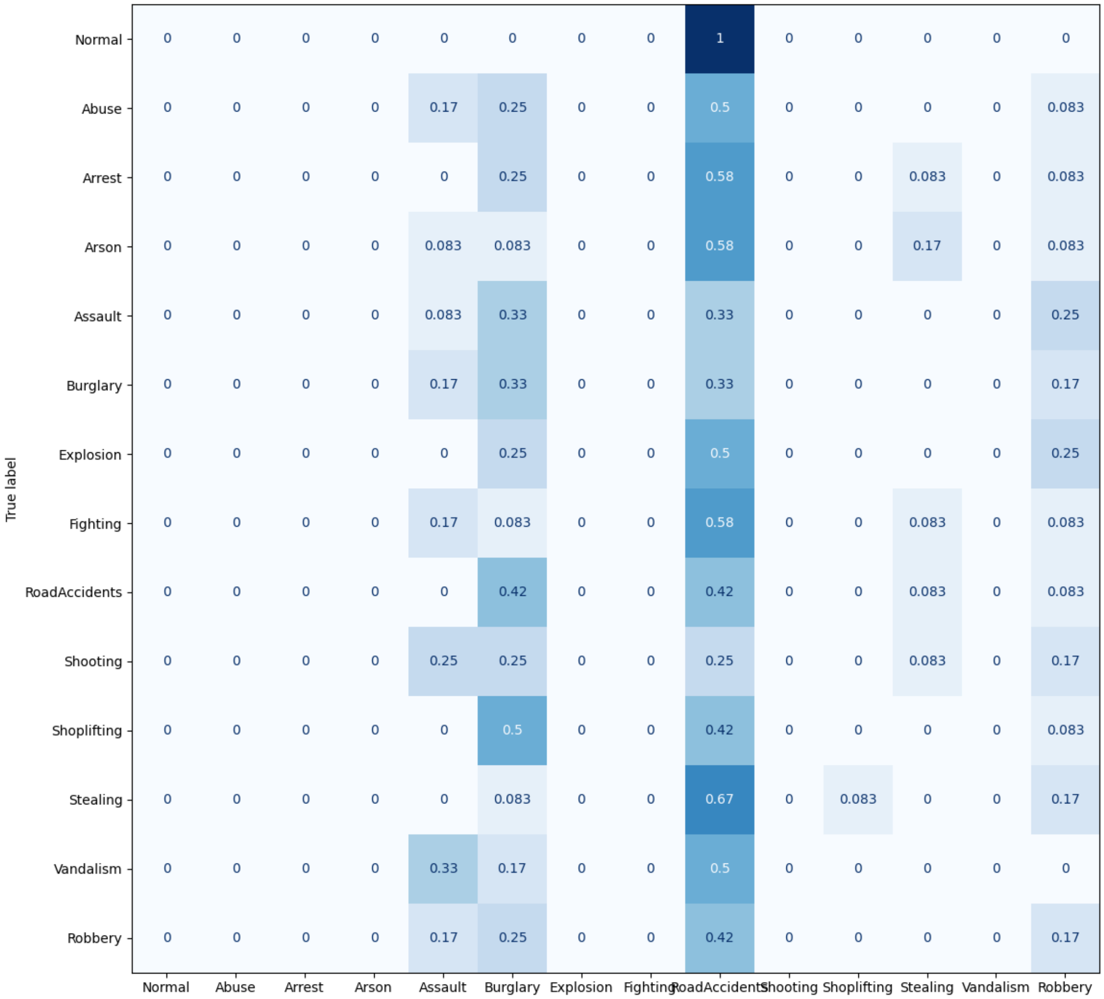
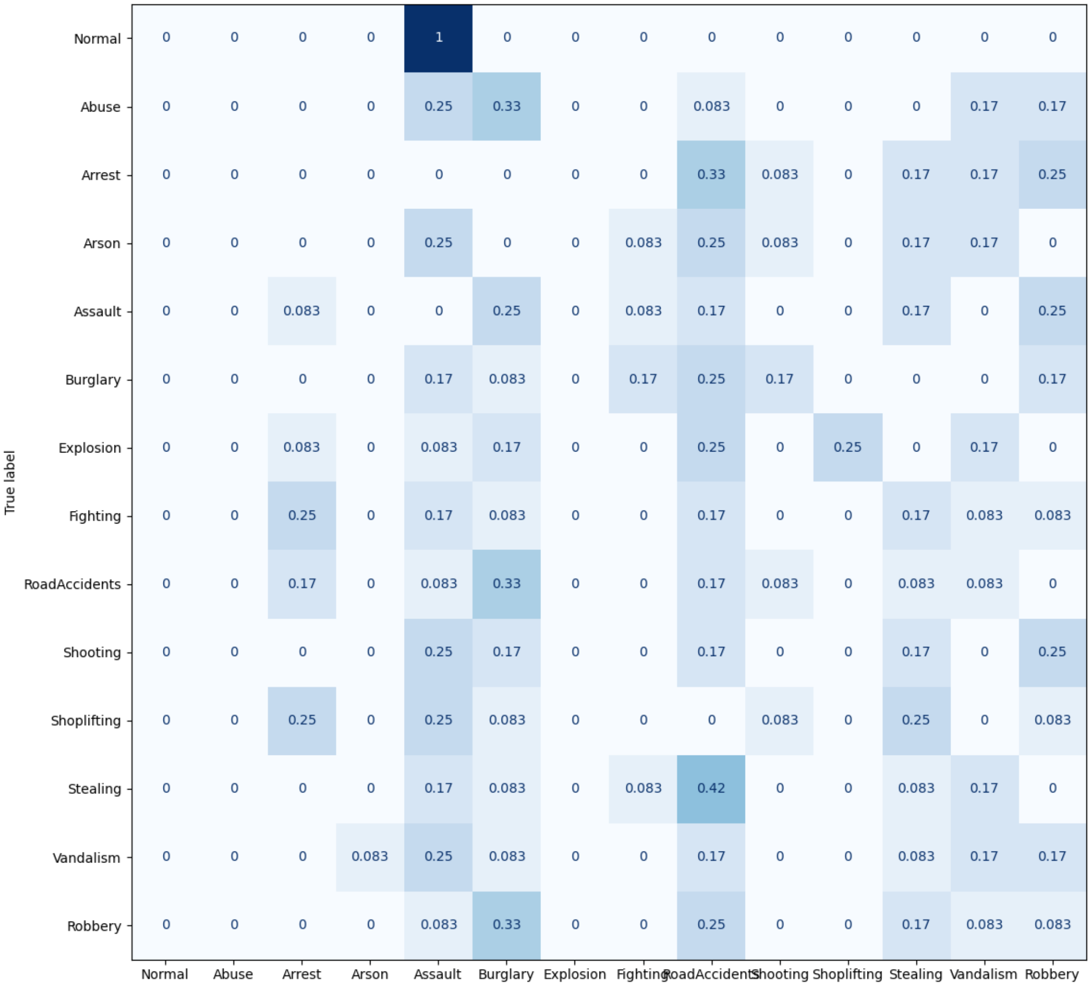


Fig. 9. ST-GCN++ -2 confusion matrix



	Normal	Abuse	Arrest	Arson	Assault	Burglary	Explosion	Fighting	RoadAccidents	Shooting	Shoplifting	Stealing	Vandalism	Robbery
Normal	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Abuse	0	0	0	0	0.17	0.25	0	0	0.5	0	0	0	0	0.083
Arrest	0	0	0	0	0	0.25	0	0	0.58	0	0	0.083	0	0.083
Arson	0	0	0	0	0.083	0.083	0	0	0.58	0	0	0.17	0	0.083
Assault	0	0	0	0	0.083	0.33	0	0	0.33	0	0	0	0	0.25
Burglary	0	0	0	0	0.17	0.33	0	0	0.33	0	0	0	0	0.17
Explosion	0	0	0	0	0	0.25	0	0	0.5	0	0	0	0	0.25
Fighting	0	0	0	0	0.17	0.083	0	0	0.58	0	0	0.083	0	0.083
RoadAccidents	0	0	0	0	0	0.42	0	0	0.42	0	0	0.083	0	0.083
Shooting	0	0	0	0	0.25	0.25	0	0	0.25	0	0	0.083	0	0.17
Shoplifting	0	0	0	0	0	0.5	0	0	0.42	0	0	0	0	0.083
Stealing	0	0	0	0	0	0.083	0	0	0.67	0	0.083	0	0	0.17
Vandalism	0	0	0	0	0.33	0.17	0	0	0.5	0	0	0	0	0
Robbery	0	0	0	0	0.17	0.25	0	0	0.42	0	0	0	0	0.17

Fig. 10. PoseConv3D-3 confusion matrix



	Normal	Abuse	Arrest	Arson	Assault	Burglary	Explosion	Fighting	RoadAccidents	Shooting	Shoplifting	Stealing	Vandalism	Robbery
Normal	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Abuse	0	0	0	0	0.25	0.33	0	0	0.083	0	0	0	0.17	0.17
Arrest	0	0	0	0	0	0	0	0	0.33	0.083	0	0.17	0.17	0.25
Arson	0	0	0	0	0.25	0	0	0.083	0.25	0.083	0	0.17	0.17	0
Assault	0	0	0.083	0	0	0.25	0	0.083	0.17	0	0	0.17	0	0.25
Burglary	0	0	0	0	0.17	0.083	0	0.17	0.25	0.17	0	0	0	0.17
Explosion	0	0	0.083	0	0.083	0.17	0	0	0.25	0	0.25	0	0.17	0
Fighting	0	0	0.25	0	0.17	0.083	0	0	0.17	0	0	0.17	0.083	0.083
RoadAccidents	0	0	0.17	0	0.083	0.33	0	0	0.17	0.083	0	0.083	0.083	0
Shooting	0	0	0	0	0.25	0.17	0	0	0.17	0	0	0.17	0	0.25
Shoplifting	0	0	0.25	0	0.25	0.083	0	0	0	0.083	0	0.25	0	0.083
Stealing	0	0	0	0	0.17	0.083	0	0.083	0.42	0	0	0.083	0.17	0
Vandalism	0	0	0	0.083	0.25	0.083	0	0	0.17	0	0	0.083	0.17	0.17
Robbery	0	0	0	0	0.083	0.33	0	0	0.25	0	0	0.17	0.083	0.083

Fig. 11. ST-GCN++ -3 confusion matrix