



Support Learning Vovinam Exercises based on Computer Vision



Pham Son Tung, Thai Thanh Do, Pham Hong Giang

Instructor Phan Duy Hung

FPT University, Hanoi, Vietnam

TABLE OF CONTENTS

1. INTRODUCTION
2. METHODOLOGY
3. EXPERIMENTS RESULT AND CONCLUSION

INTRODUCTION

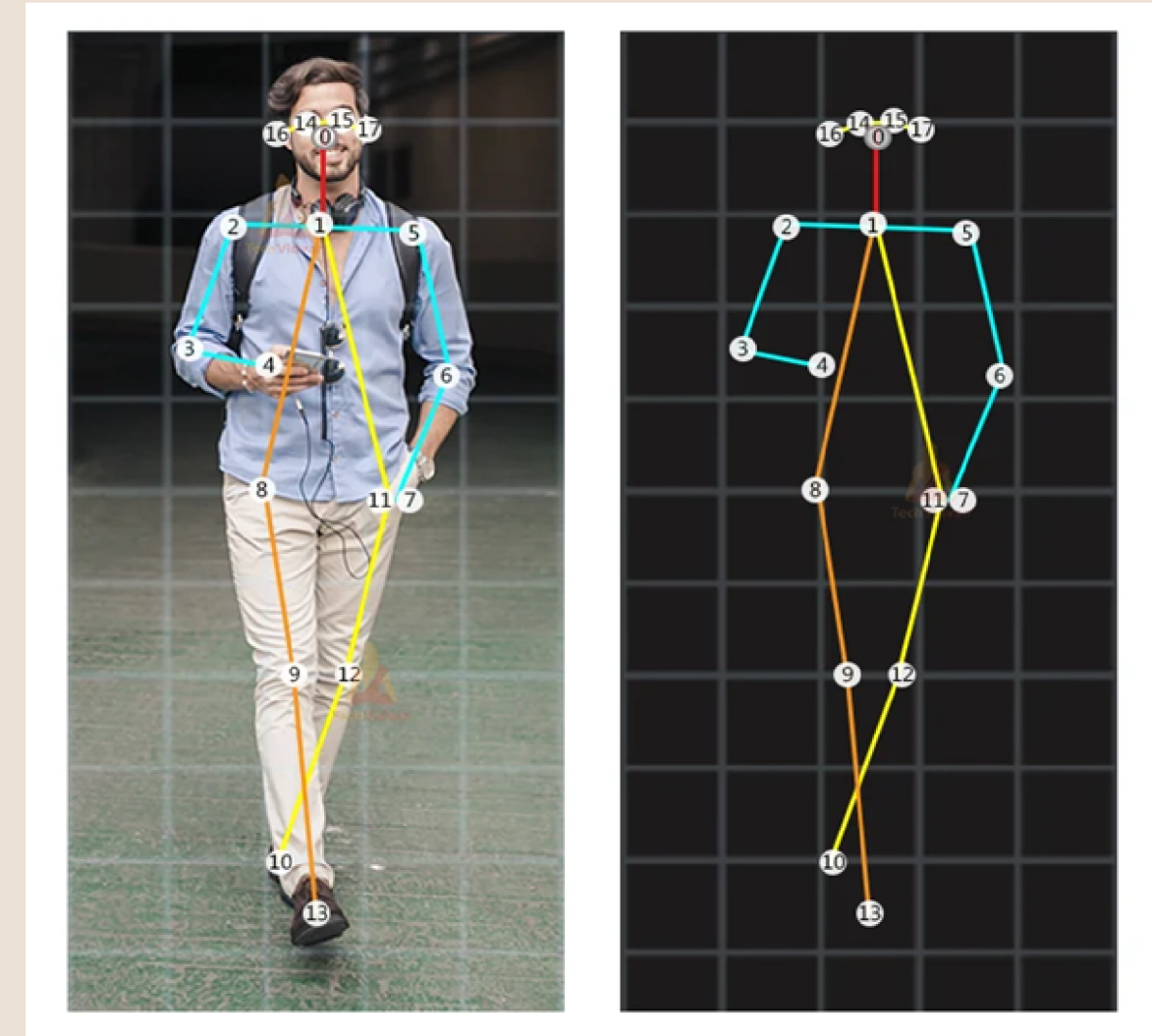
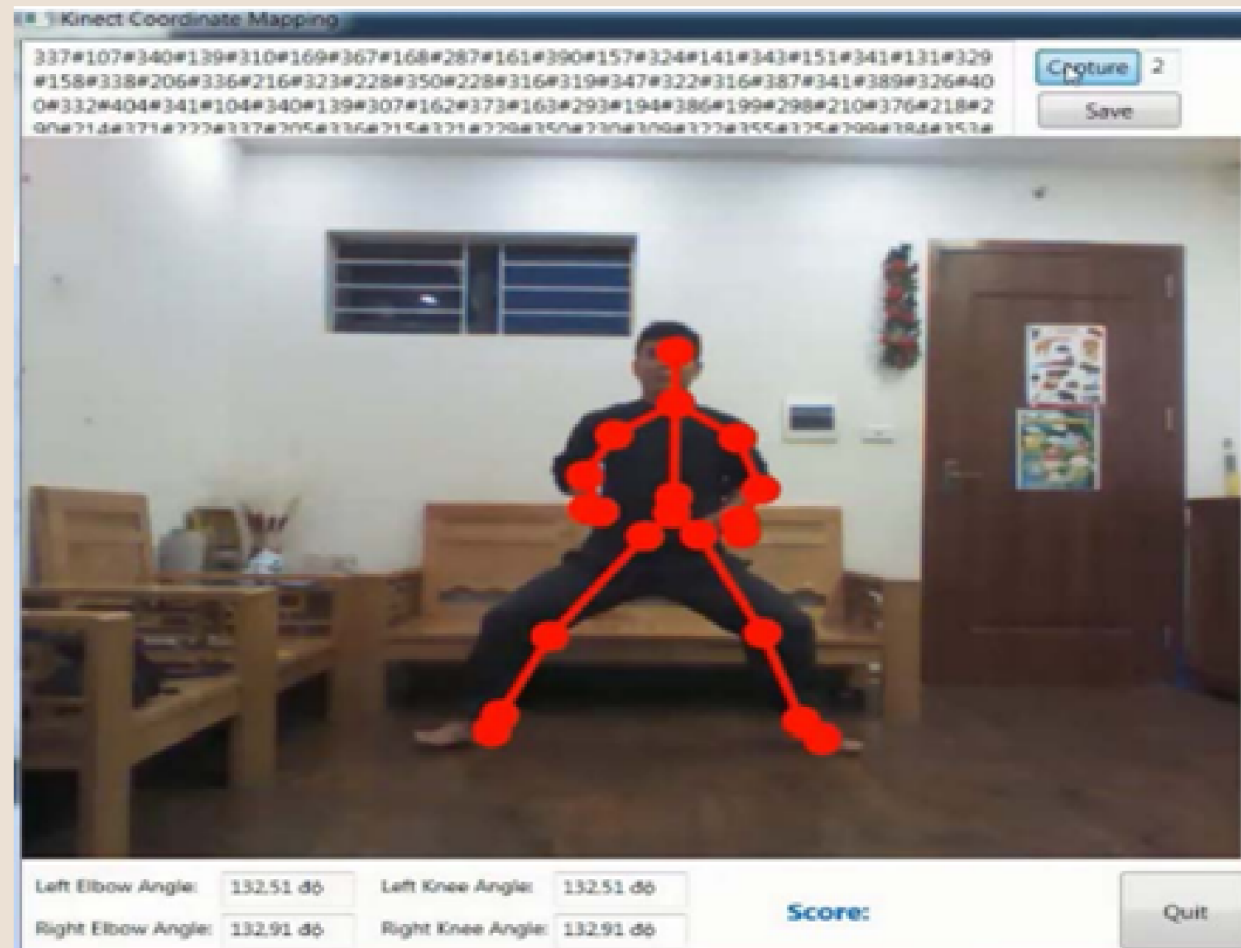
Overview

- In multiple forms of physical activities that people use today, martial arts is one of the most effective and popular ways. In Vietnam, the national traditional martial art form of the country.
- Along with a tremendous amount of time on self-training, the application of computer vision in supporting self-training will be very effective and convenient.



Related works

- The rise of new research on human body skeleton dynamics contributes an important part for recognizing human actions.

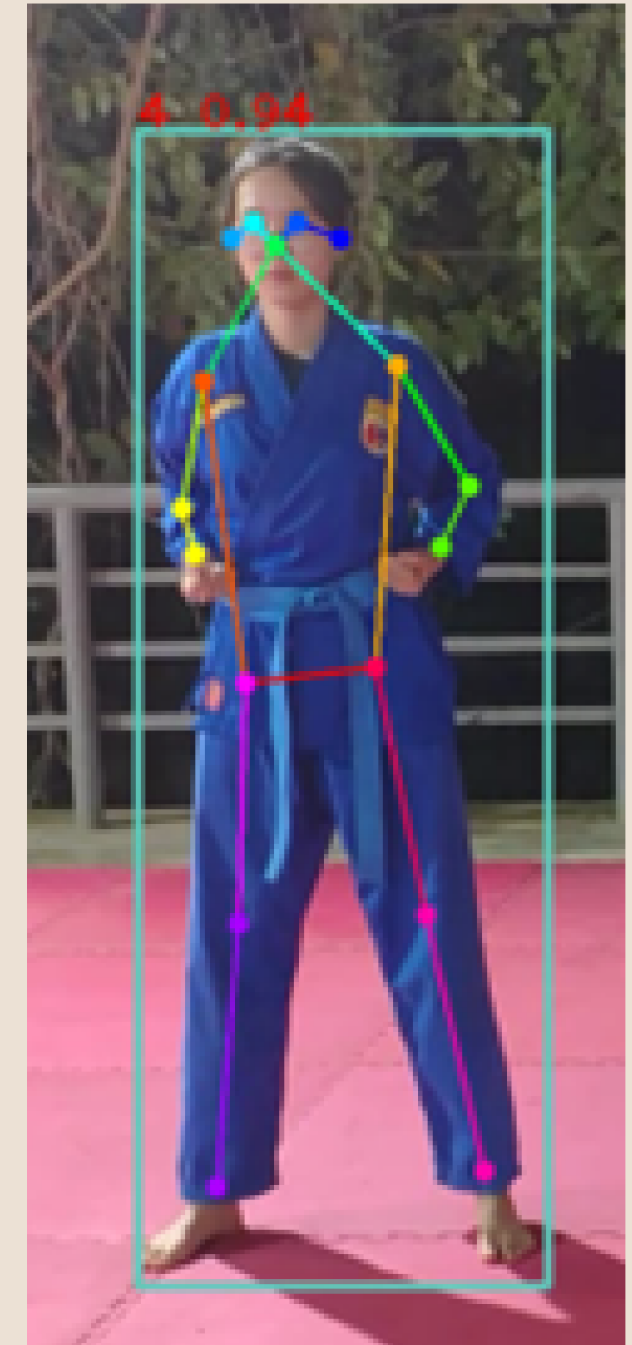


- For grading movements of Vietnamese martial arts, Tuong Thanh et al proposes the implementation of data analysis of the depth of scoring Kinect users' camera movements in grading Vietnamese traditional martial arts movements.

Contribution

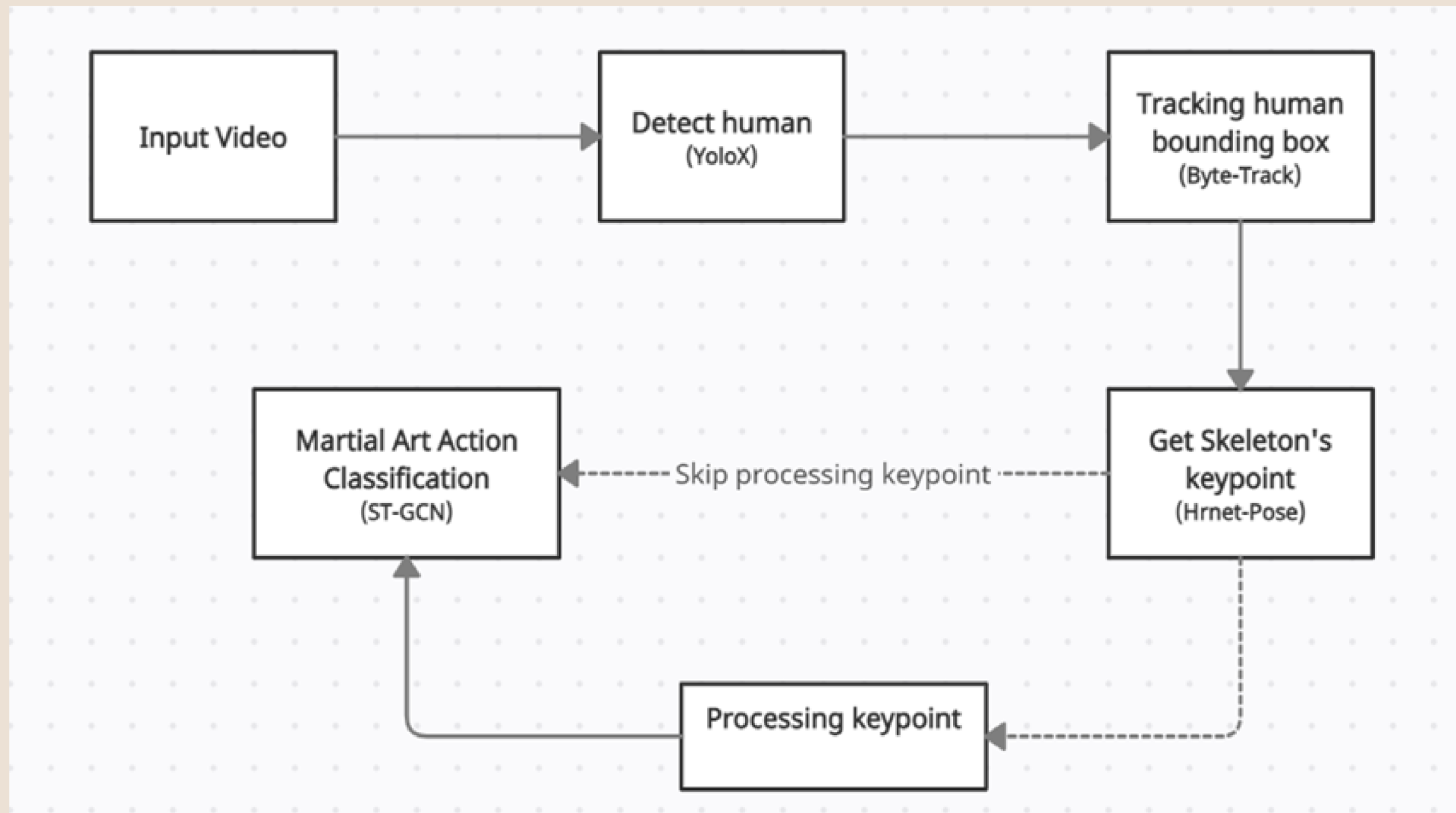


- Our goal in this thesis is to develop an end-to-end action recognition model that leverages ST-GCN kernels with several modifications and some recent computer vision techniques to deal with the complex moves of vovinam martial arts.
- This study suggests a pipeline where we have performed auto-generator data from video to keypoints sequence and processed them appropriately for martial arts movements. Later on, the data would be put into the ST-GCN model to deal with the problem of classification of martial arts poses.



METHODOLOGY

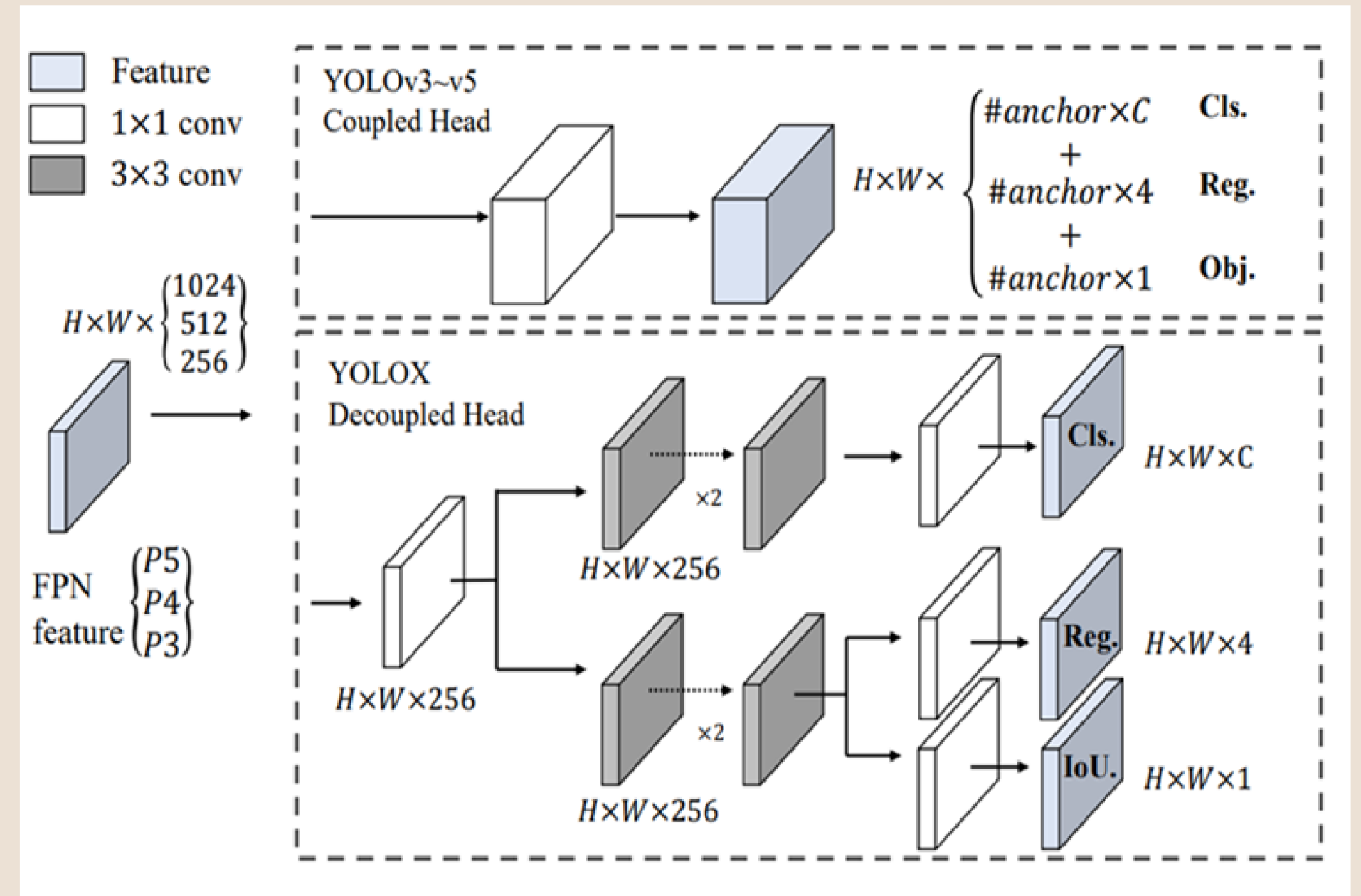
Overview pipeline



Block diagram for classifying martial art action

YoloX

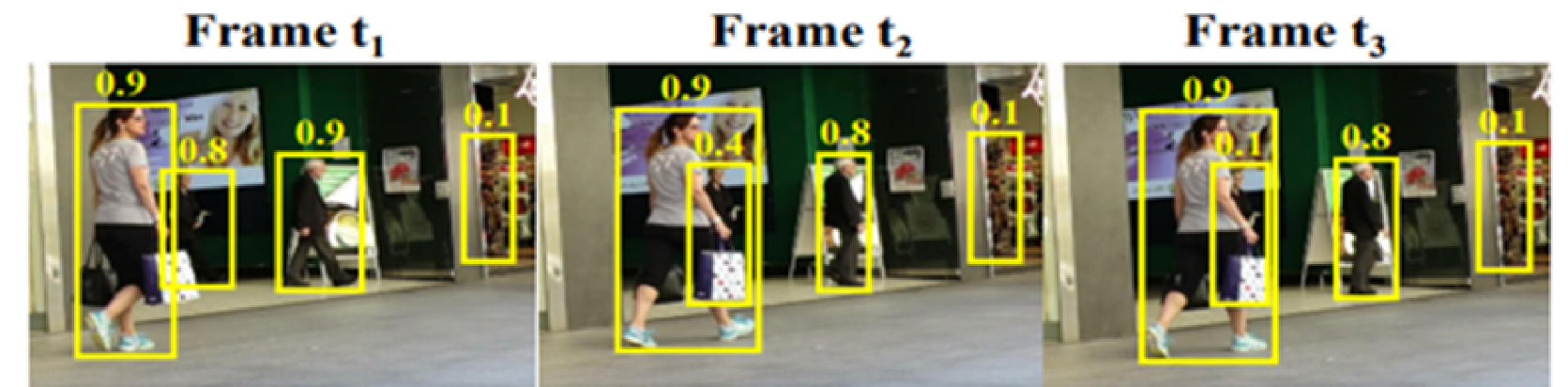
- We chose YoloX for this model because it improved object detection by emphasizing anchor-free detectors and had better detector performance than the Yolo-series.
- And It is also recommended that YoloX be used with the byte-track method for better performance.



YOLOv3 head and YoloX decoupled head

Byte-Track

- Most of previous approach identify identities by identifying association detection boxes with scores over a threshold, then discarding the rest of the objects with low detection scores, resulting in significant actual object loss and fractured trajectories.
- Byte-track tracks nearly every detection box, not just the highest-scoring ones.



(a) detection boxes



(b) tracklets by associating high score detection boxes

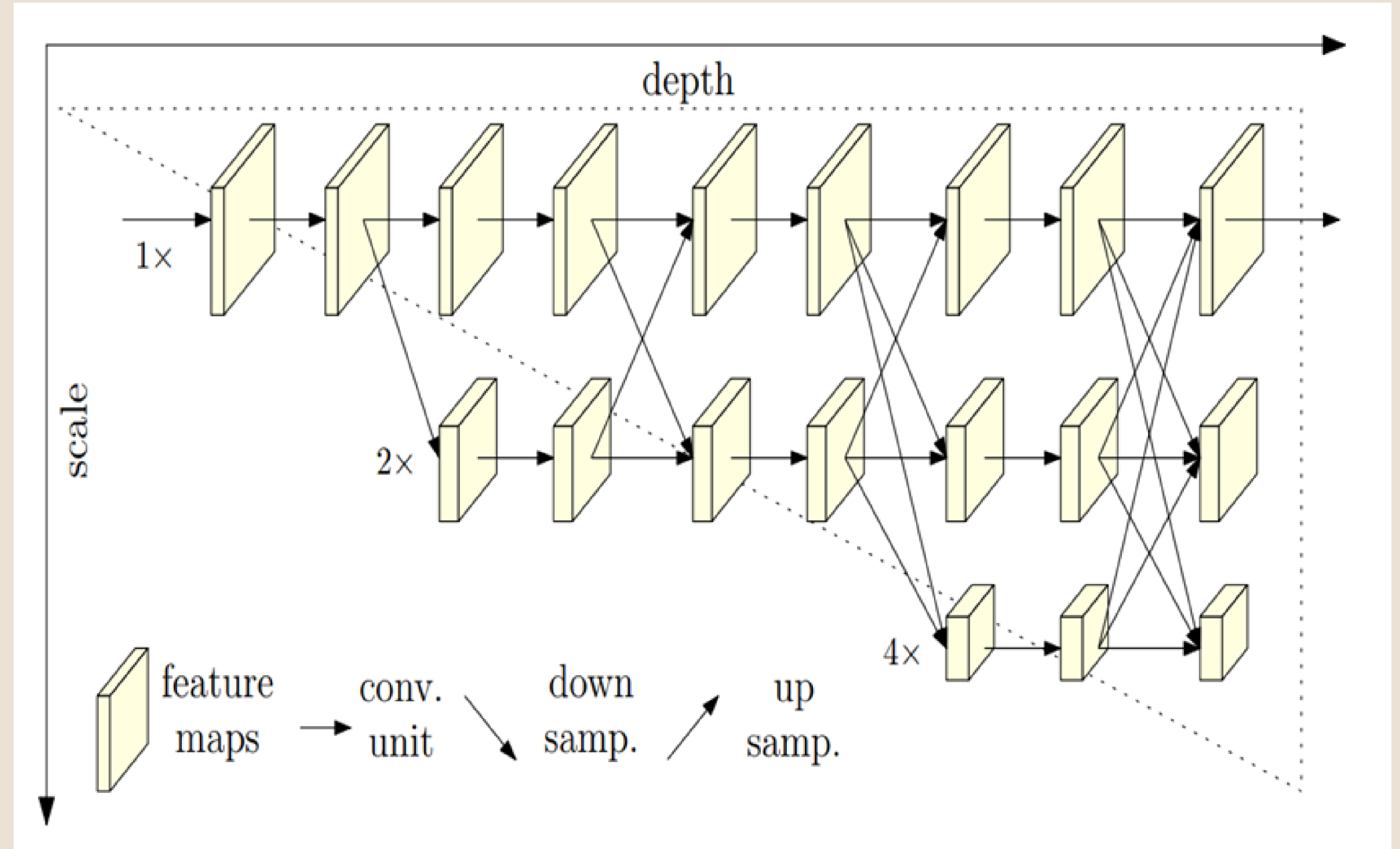


(c) tracklets by associating every detection box

Byte-track associates every detection box

HRNet-Pose

- In contrast with most previous systems that were linked in a series, high-to-low resolution subnetworks are linked in parallel.
- In this method use recurrent multiscale fusions, high-resolution and low-resolution will be augmented representations . As a result, the anticipated heatmap may be more accurate.



Illustrating the architecture of the proposed HRNet

Processing Keypoints

- Technique 1: Interpolation missing frames.



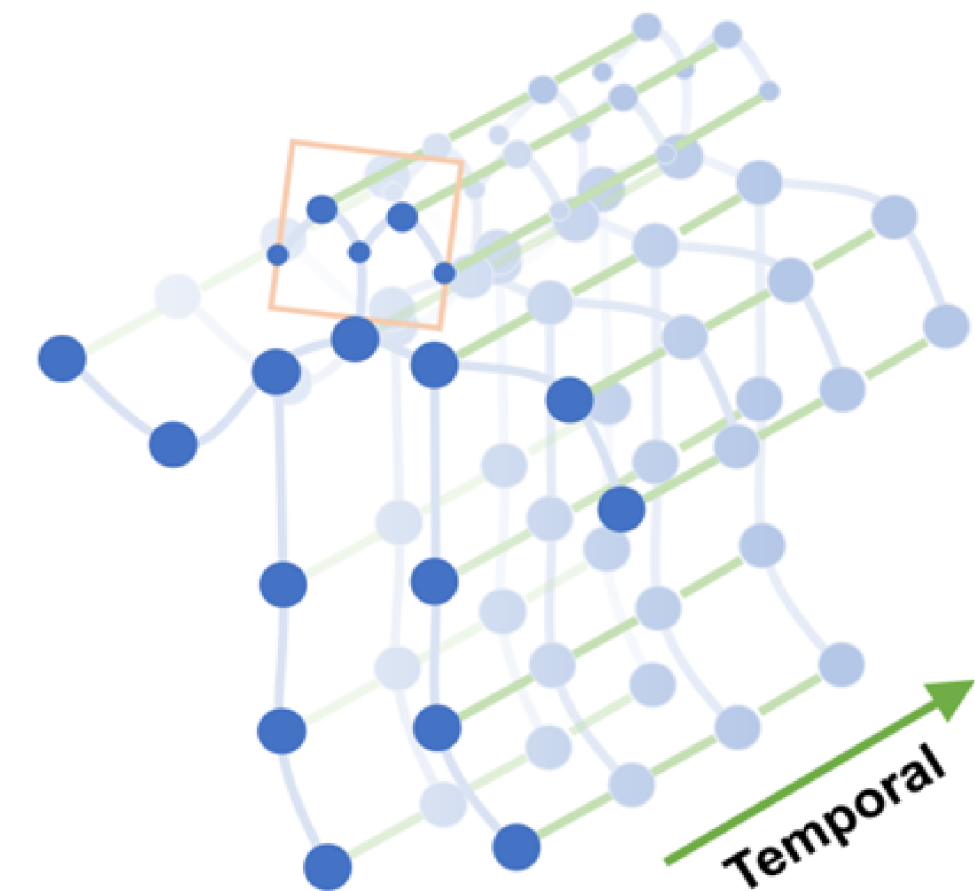
- Technique 2: data augmentation techniques .

- After processing the above technique, we process the keypoints sequences of each processed frame into the ST-GCN model to train and predict the martial art movements.



ST-GCN

- Relatively new approach to automatically capturing the patterns stored in the joint's spatial configuration and also their temporal dynamics.
- The use of GCNs to represent dynamic graphs spanning large-scale datasets, like as human skeletal sequences, has not yet been examined. By extending GCNs to a spatial-temporal graph model known as ST-GCN for Skeleton-Based Action Recognition.

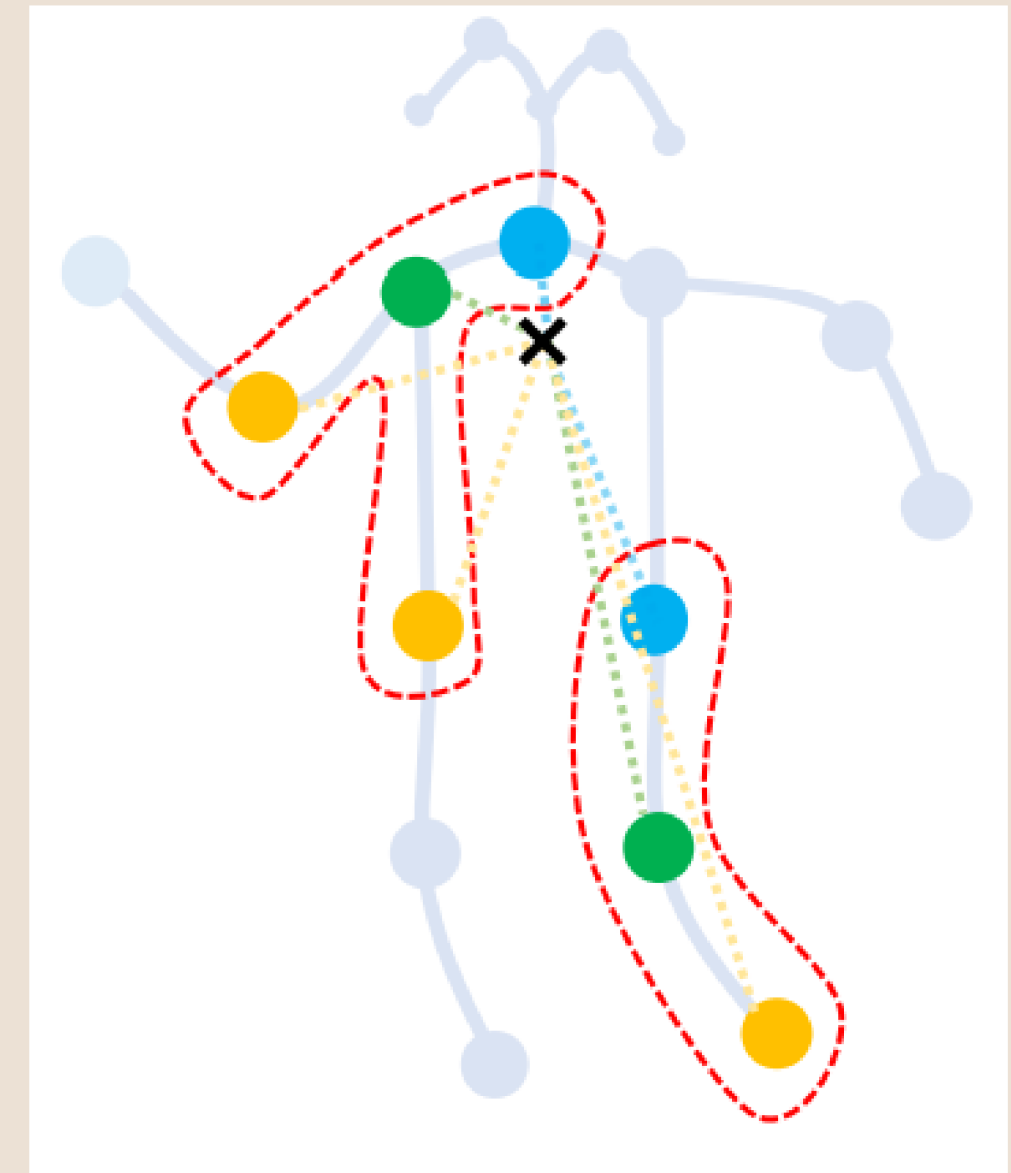


A skeleton sequence's spatial-temporal graph

$$l_{ti}(v_{tj}) = \begin{cases} 0 & \text{if } r_j = r_i \\ 1 & \text{if } r_j < r_i \\ 2 & \text{if } r_j > r_i \end{cases}$$

1. The root node itself.
2. Centripetal group: the neighboring nodes that are closer to the gravity center of the skeleton than the root node.
3. Otherwise, the centrifugal group.

» $l_{ST}(v_{qj}) = l_{ti}(v_{tj}) + (q - t + \lfloor \Gamma/2 \rfloor) \times K$



Spatial configuration partitioning

Implementing ST-GCN

GCN

$$\mathbf{f}_{out} = \mathbf{\Lambda}^{-\frac{1}{2}} (\mathbf{A} + \mathbf{I}) \mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{f}_{in} \mathbf{W}$$

ST-GCN

$$\mathbf{f}_{out} = \sum_j \mathbf{\Lambda}_j^{-\frac{1}{2}} \mathbf{A}_j \mathbf{\Lambda}_j^{-\frac{1}{2}} \mathbf{f}_{in} \mathbf{W}_j$$

The graph convolution is implemented by performing a $1 \times \Gamma$ standard 2D convolution and multiplies the resulting tensor with the normalized adjacency matrix $\mathbf{\Lambda}^{-\frac{1}{2}} (\mathbf{A} + \mathbf{I}) \mathbf{\Lambda}^{-\frac{1}{2}}$ on the second dimension.

Network architecture and training

EXPERIMENTS RESULT AND CONCLUSION

Data Collection

Table 1. Data statistics

Type	Value
Number people	10
Number class	9
Total number action	778



Examples of Vovinam movement recognition

Experiments

Environment

- GPU-3090, Cuda 11.3, Ubuntu 22.04, Python 3.7

Setup virtual environment

- Step 1 : Install Anaconda.
- Step 2: Setup Environment and Install Package.

```
conda create --name capstone python=3.7
```

```
conda activate capstone
```

```
conda install pytorch pytorch-cuda=11.3 -c pytorch -c nvidia
```

```
pip install -r requirements.txt
```

- Step 3: Import video data into the pipeline's model with the configuration parameters.



YoloX

Parameter	Value
Input Shape	640,640
Score Threshold	0.1
NMS Threshold	0.7

ByteTrack

Parameter	Value
Track threshold	0.7
Track buffer	30
Match threshold	0.8

ST-GCN

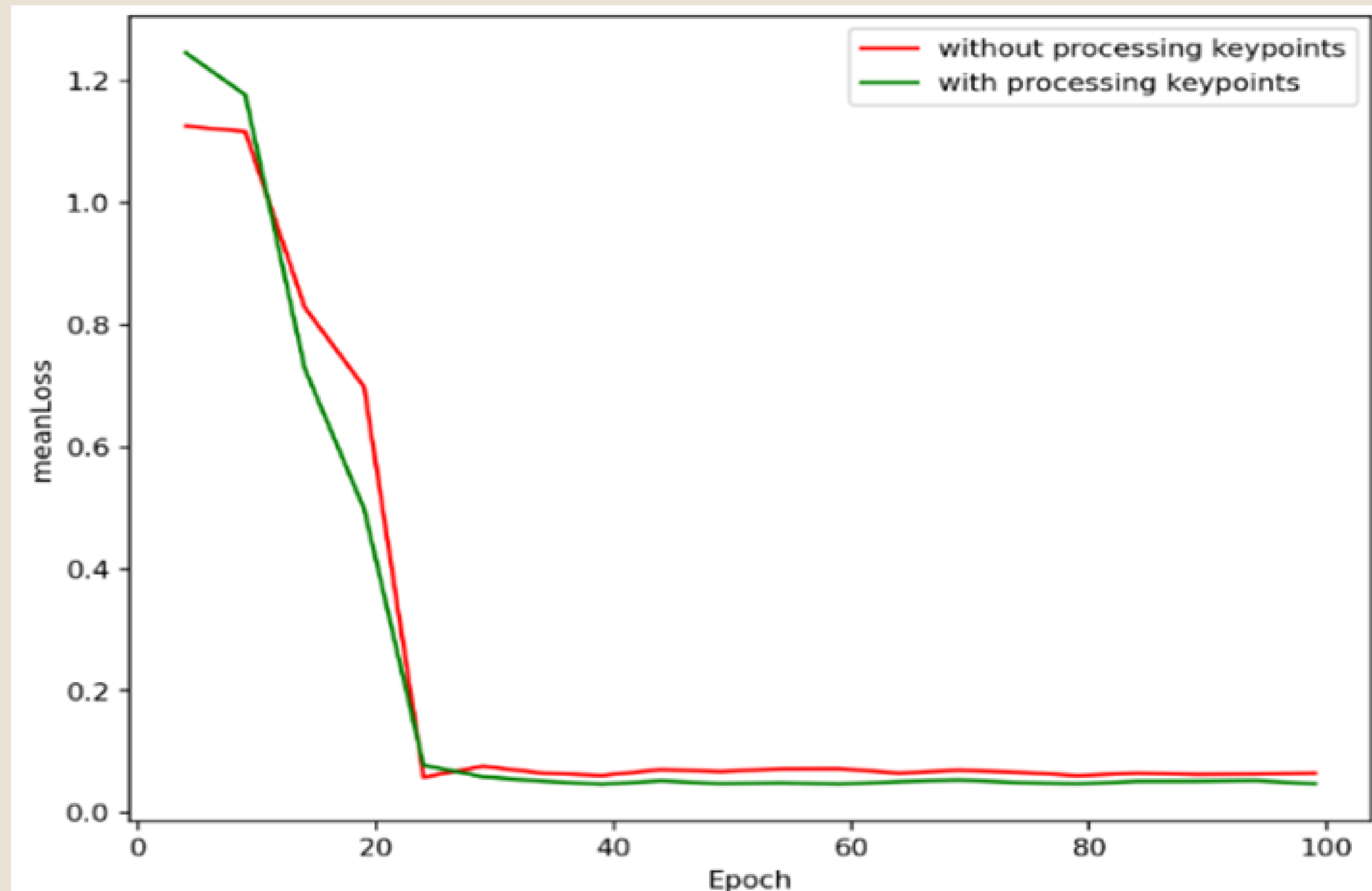
Parameter	Value
base lr	0.1
batch size	32
num epoch	100
optimizer	SGD
weight decay	0.0001

Configuration parameter

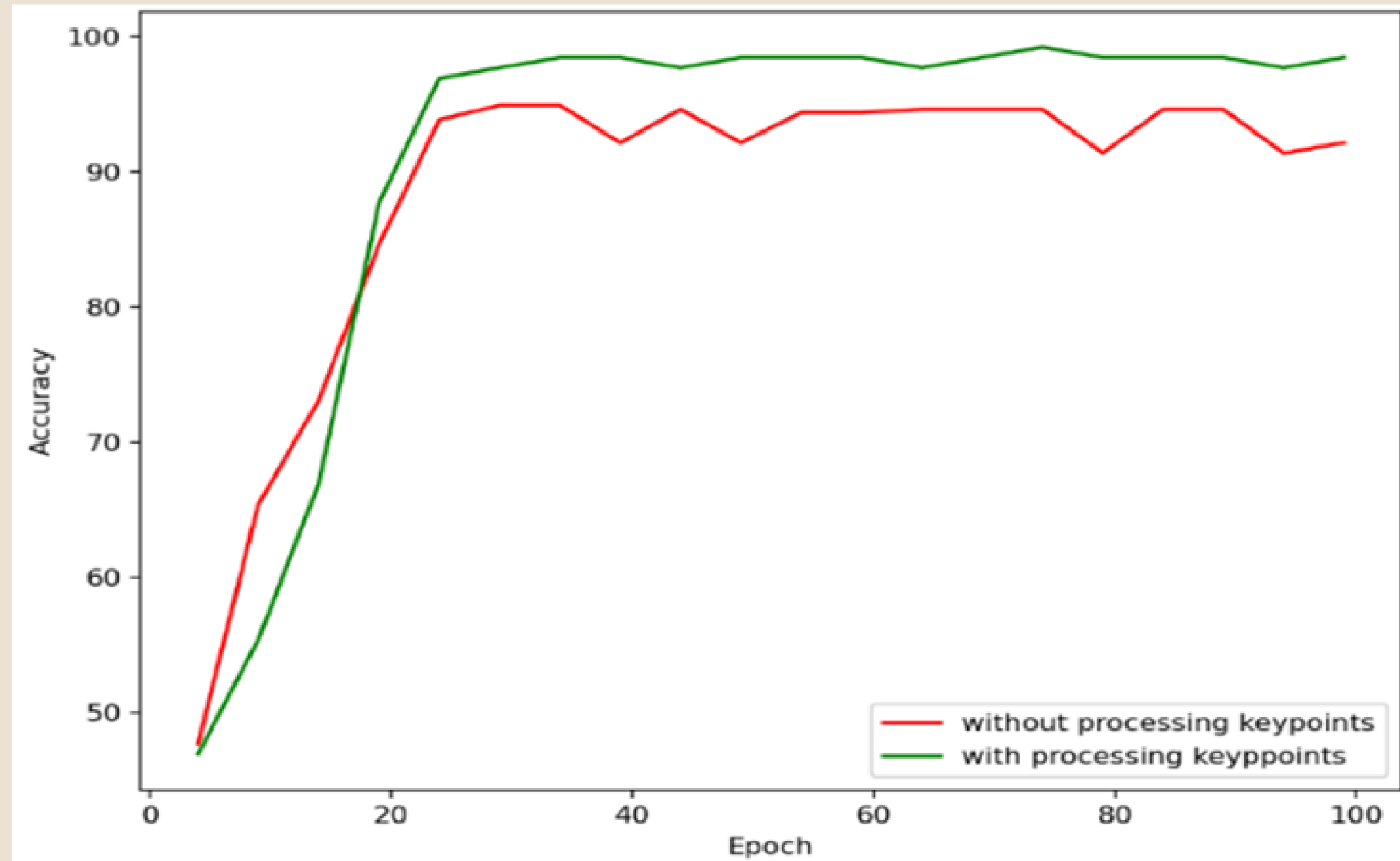
Result and analysis

Model	Top1-Accuracy	Mean loss
Without processing keypoints	94.62%	0.065
With processing keypoints	99.23%	0.047

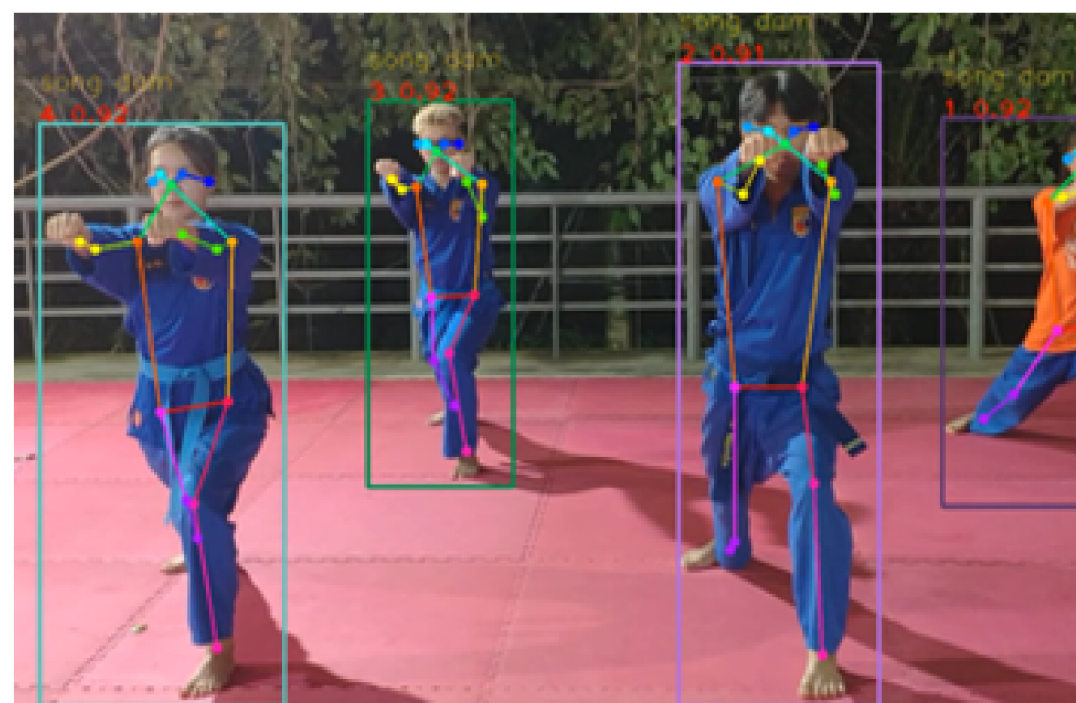
Results of two cases with and without processing keypoints



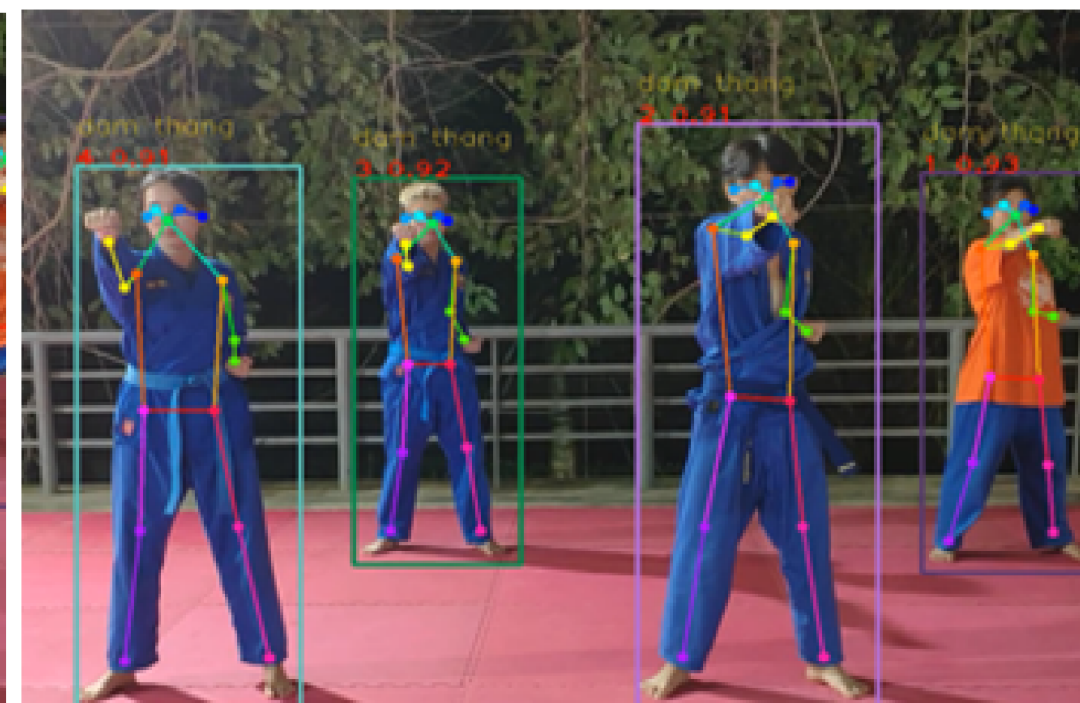
The meanLoss



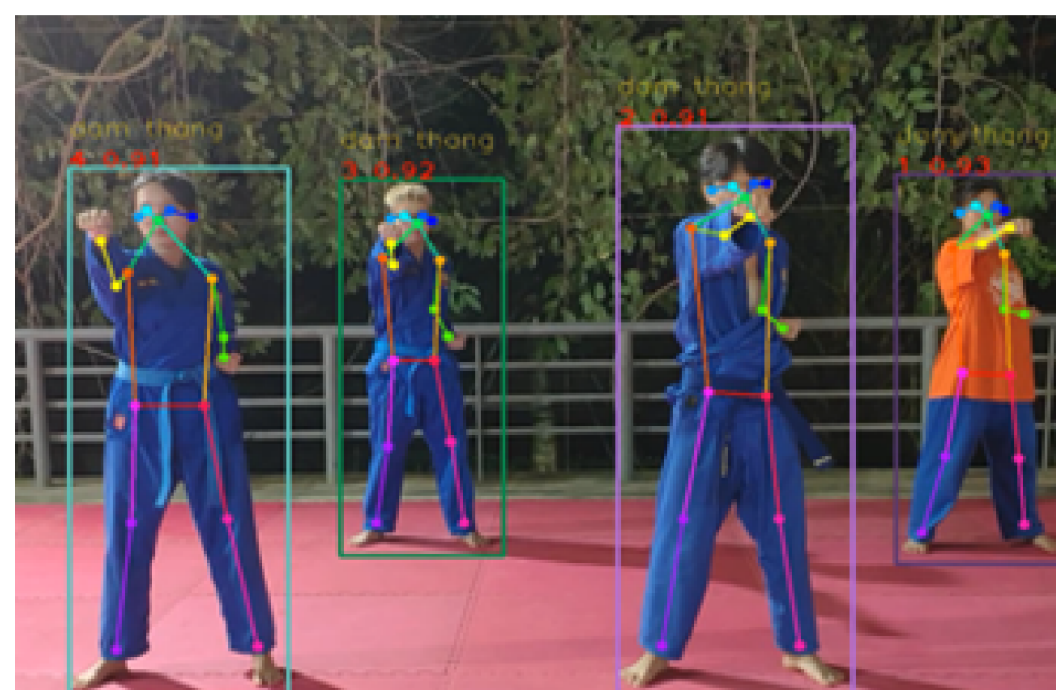
The accuracy



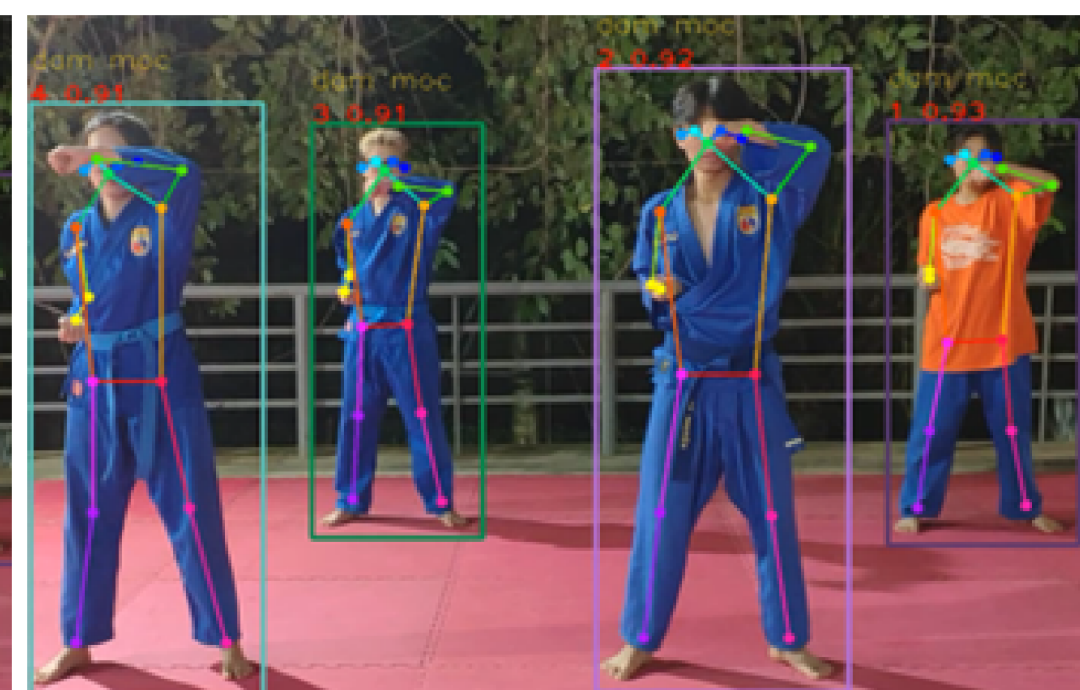
Double punch movement (song đấm)



Straight punch movement (đấm thẳng)



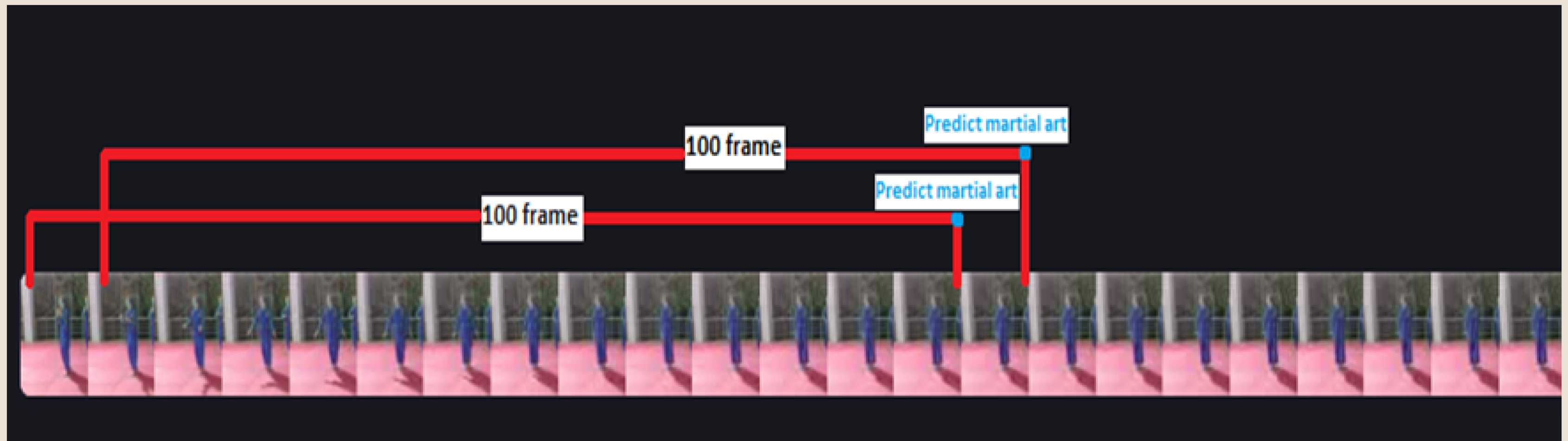
Straight punch movement (đấm thẳng)



Hook punch movement (đấm móc)


Examples of Vovinam movement recognition


Inference



Predictive process of the model on long video

Conclusion & Future Works

- 
- This study proposed a pipeline for identifying martial arts movements in Vovinam, a Vietnamese martial art. Each phase's appropriate approaches are thoroughly considered and selected from the most recent methods. The data was gathered and labeled, including nine classes separated into three categories: standing still, defense, and basic martial arts movement. A new processing phase for keypoints is added to enhance input for the ST-GCN model.

- 
- The model can be further developed on a complete database of all lessons, movements. The development of the application into a mobile application product in order to provide the easiest support for learners is also worth paying attention to.

