# Human Action Recognition

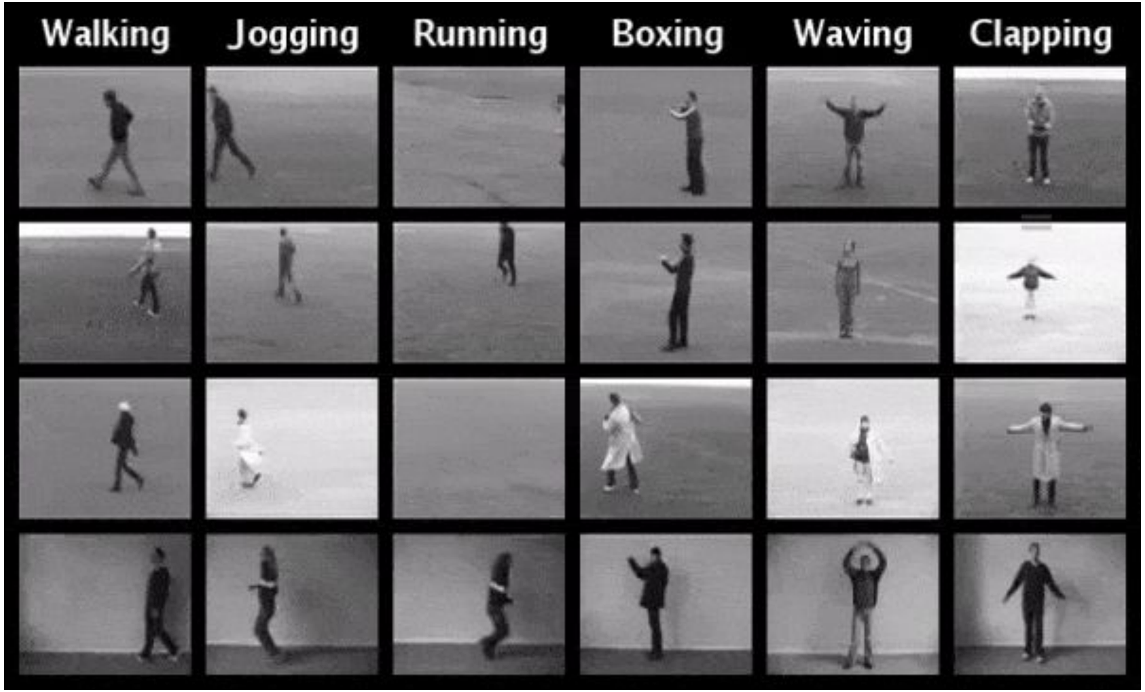
Ashok Prasad Neupane

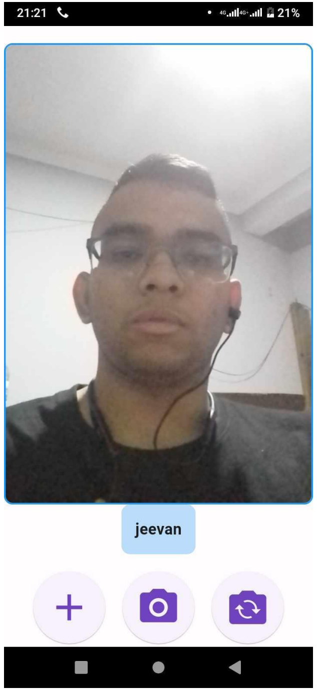
Anil Shrestha

Jeevan Neupane

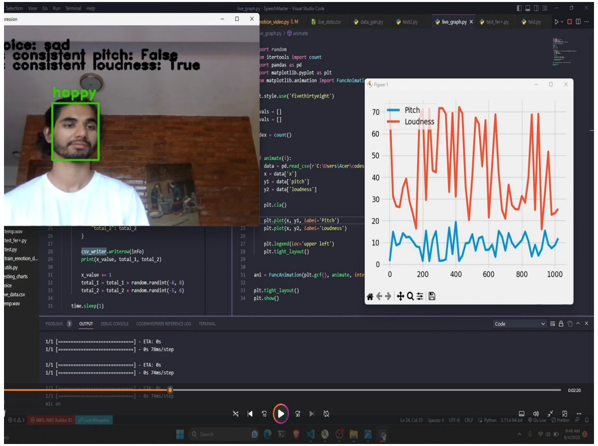
Contacts: neupane.ashok.9696@gmail.com

9818467416





Face Recognition using Siamese Network


Selection View Go Run Terminal Help
live_graph.py - SpeechMaster - Visual Studio Code
oice: sad
consistent pitch: False
consistent loudness: True
happy
Figure 1
Pitch
Loudness
70
60
50
40
30
20
10
0
0
200
400
600
800
1000
PROBLEMS
OUTPUT
DEBUG CONSOLE
CODEWHISPERER REFERENCE LOG
TERMINAL

```
        "total_2": total_2
    }

    csv_writer.writerow(info)
    print(x_value, total_1, total_2)

    x_value += 1
    total_1 = total_1 + random.randint(-6, 8)
    total_2 = total_2 + random.randint(-5, 6)

time.sleep(1)
```

```
port random
om itertools import count
port pandas as pd
port matplotlib.pyplot as plt
om matplotlib.animation import FuncAnimati

t.style.use('fivethirtyeight')

vals = []
vals = []

dex = count()


f animate(i):
    data = pd.read_csv(r'C:\Users\Acer\codes
    x = data['x']
    y1 = data['pitch']
    y2 = data['loudness']

    plt.cla()

    plt.plot(x, y1, label='Pitch')
    plt.plot(x, y2, label='Loudness')

    plt.legend(loc='upper left')
    plt.tight_layout()


ani = FuncAnimation(plt.gcf(), animate, inte

plt.tight_layout()
plt.show()
```

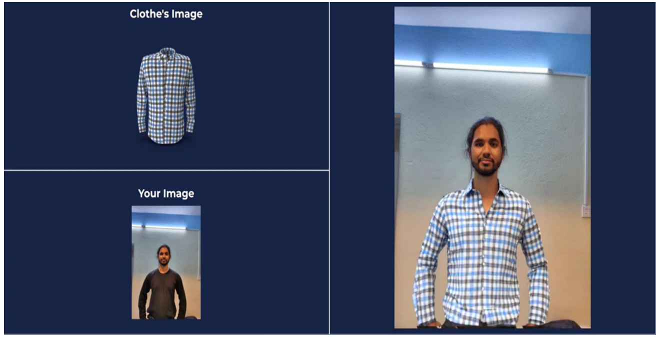
```
1/1 [==============================] - ETA: 0s
1/1 [==============================] - 0s 78ms/step

1/1 [==============================] - ETA: 0s
1/1 [==============================] - 0s 74ms/step
```

Emotion classification on audio and video

Studied about Full Self Driving Car(Tesla)

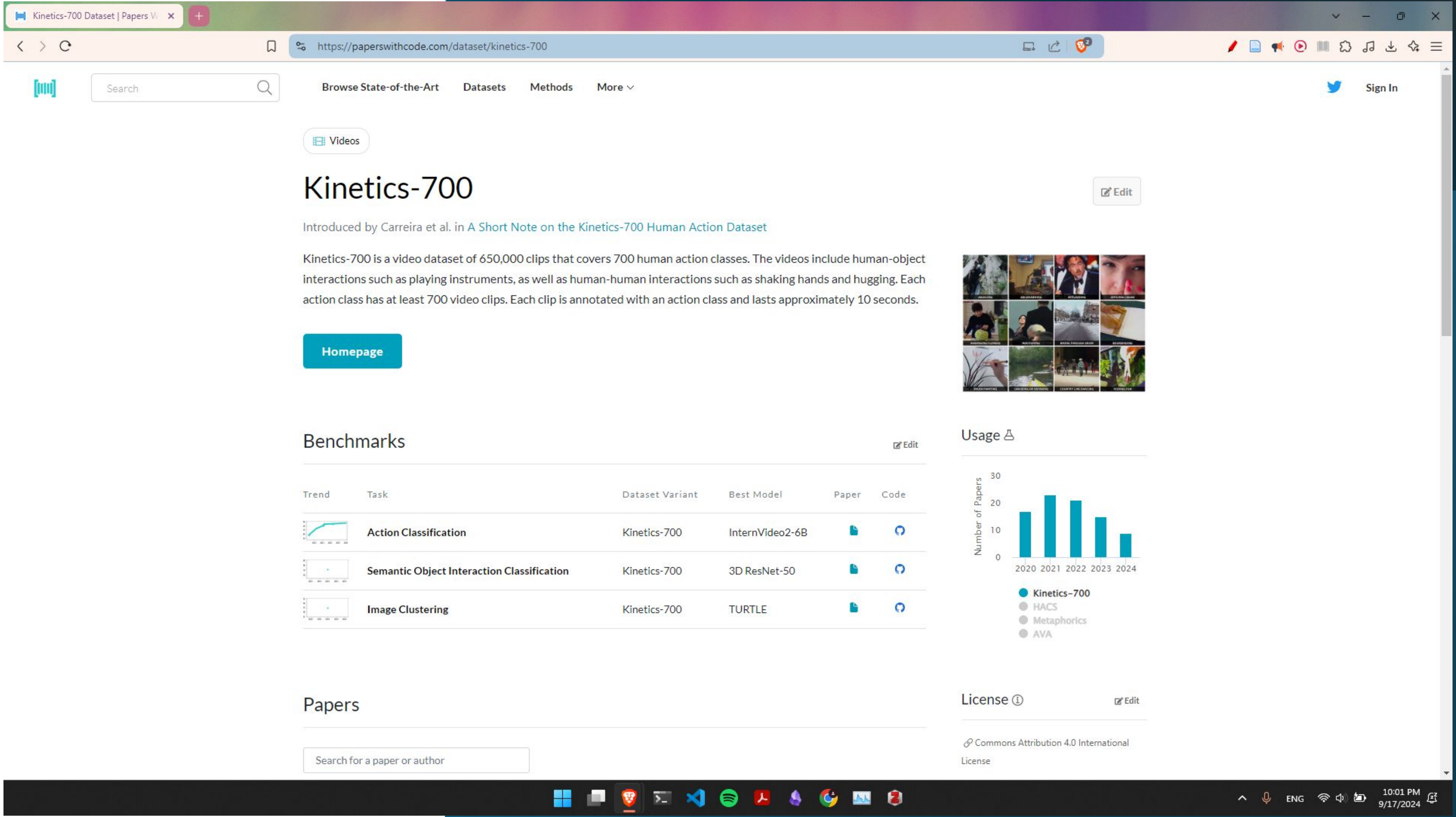
Smoking detection using fine tuned YOLO v8



Viton(Virtual Try On)

# Research question

How can deep learning models be optimized to accurately and efficiently recognize human actions in real-time from RGB video sequences, considering challenges like occlusion, poor lighting, and motion variability?


Kinetics-700 Dataset | Papers W
https://paperswithcode.com/dataset/kinetics-700
Search
Browse State-of-the-Art
Datasets
Methods
More
Sign In
Videos
Edit

# Kinetics-700

Introduced by Carreira et al. in A Short Note on the Kinetics-700 Human Action Dataset

Kinetics-700 is a video dataset of 650,000 clips that covers 700 human action classes. The videos include human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands and hugging. Each action class has at least 700 video clips. Each clip is annotated with an action class and lasts approximately 10 seconds.

## Benchmarks

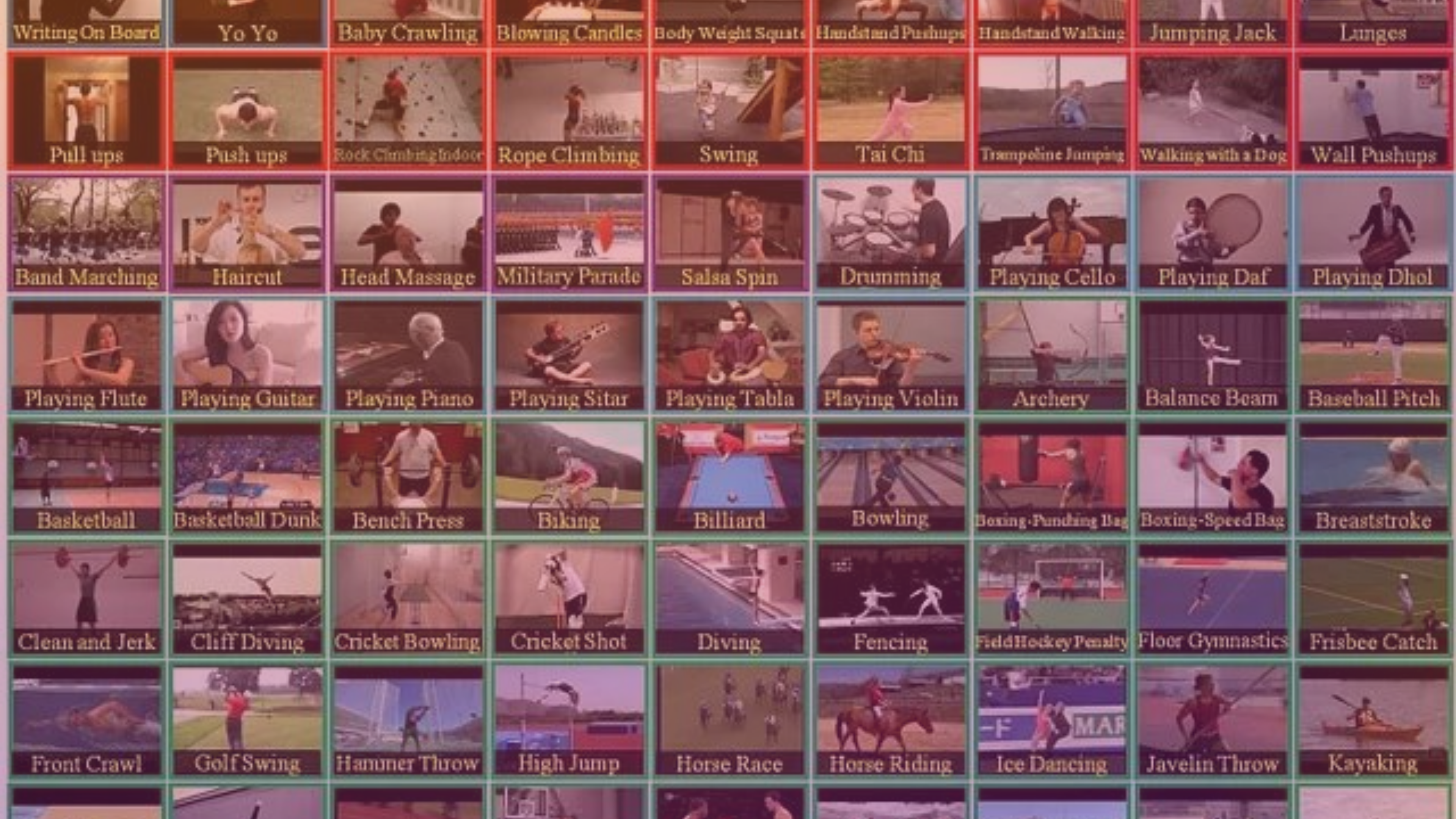
| Trend | Task | Dataset Variant | Best Model | Paper | Code |
|---|---|---|---|---|---|
| | Action Classification | Kinetics-700 | InternVideo2-6B | | |
| | Semantic Object Interaction Classification | Kinetics-700 | 3D ResNet-50 | | |
| | Image Clustering | Kinetics-700 | TURTLE | | |

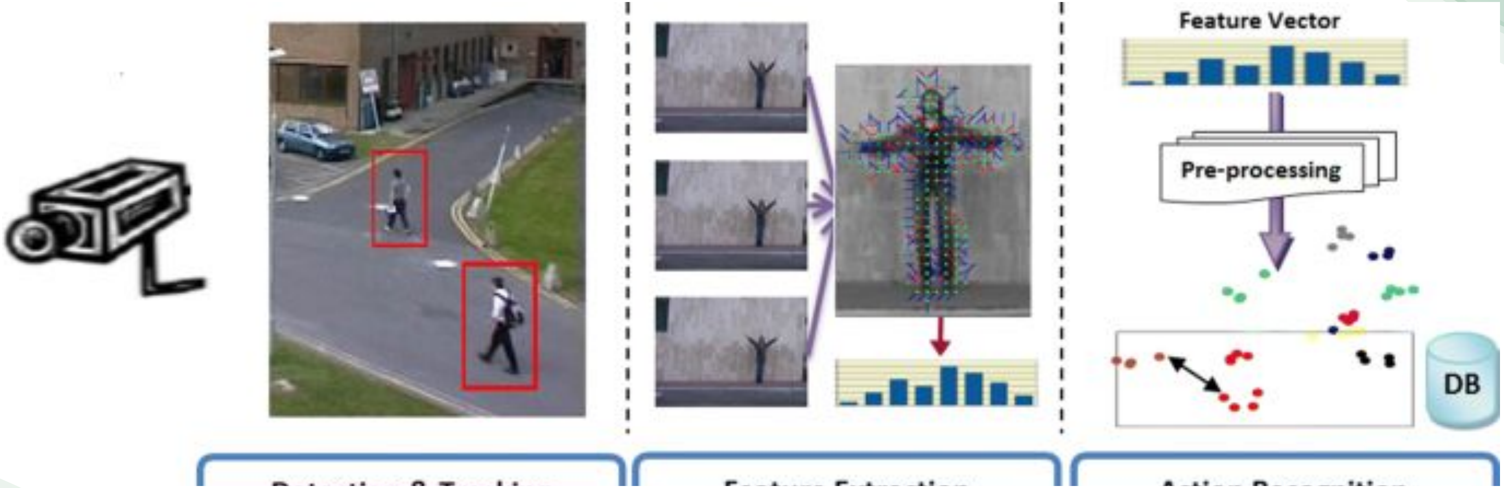
## Usage

## Papers

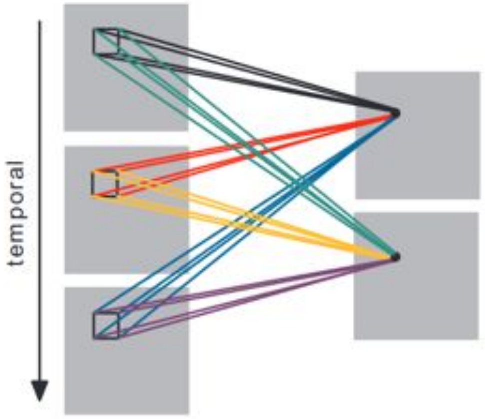
## License

Commons Attribution 4.0 International License

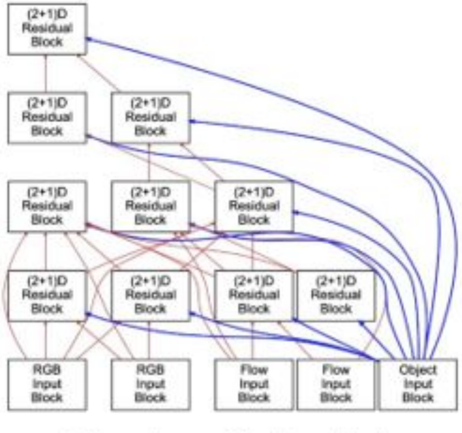

Writing On Board
Yo Yo
Baby Crawling
Blowing Candles
Body Weight Squat
Handstand Pushups
Handstand Walking
Jumping Jack
Lunges
Pull ups
Push ups
Rock Climbing Indoor
Rope Climbing
Swing
Tai Chi
Trampoline Jumping
Walking with a Dog
Wall Pushups
Band Marching
Haircut
Head Massage
Military Parade
Salsa Spin
Drumming
Playing Cello
Playing Daf
Playing Dhol
Playing Flute
Playing Guitar
Playing Piano
Playing Sitar
Playing Tabla
Playing Violin
Archery
Balance Beam
Baseball Pitch
Basketball
Basketball Dunk
Bench Press
Biking
Billiard
Bowling
Boxing-Punching Bag
Boxing-Speed Bag
Breaststroke
Clean and Jerk
Cliff Diving
Cricket Bowling
Cricket Shot
Diving
Fencing
Field Hockey Penalty
Floor Gymnastics
Frisbee Catch
Front Crawl
Golf Swing
Hammer Throw
High Jump
Horse Race
Horse Riding
Ice Dancing
Javelin Throw
Kayaking



# Feature extraction techniques like optical flow plus SVM classifier

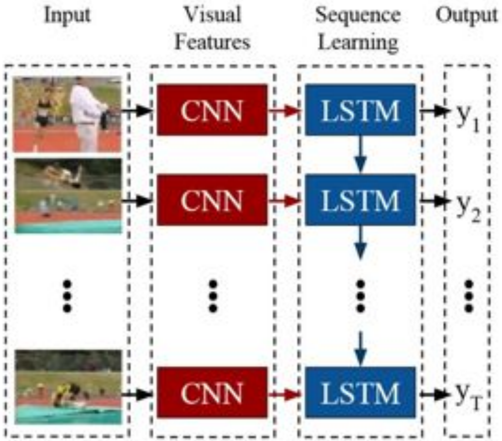


3D Convolution



(a) Learned connectivity between blocks

Multi-stream net



Long term recurrent convolution net

# Deep Learning Based Approach

# Video Foundation models

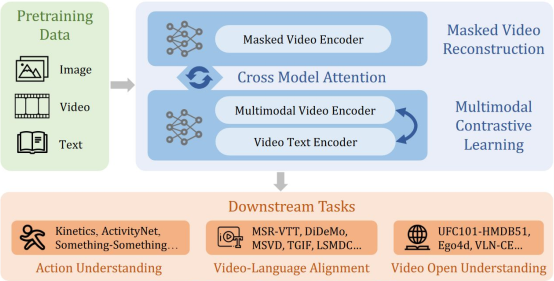
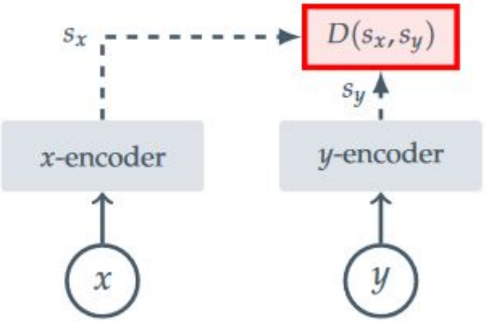


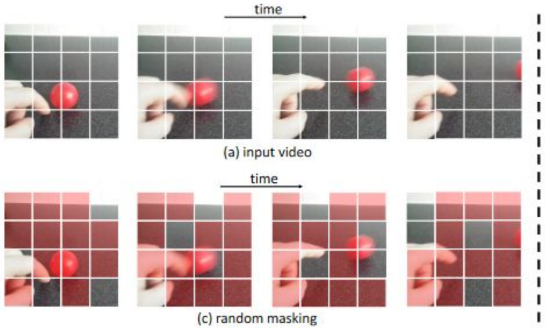
Figure 2: The overall framework of InternVideo.

Intern video



(a) Joint-Embedding Architecture

V JEPA

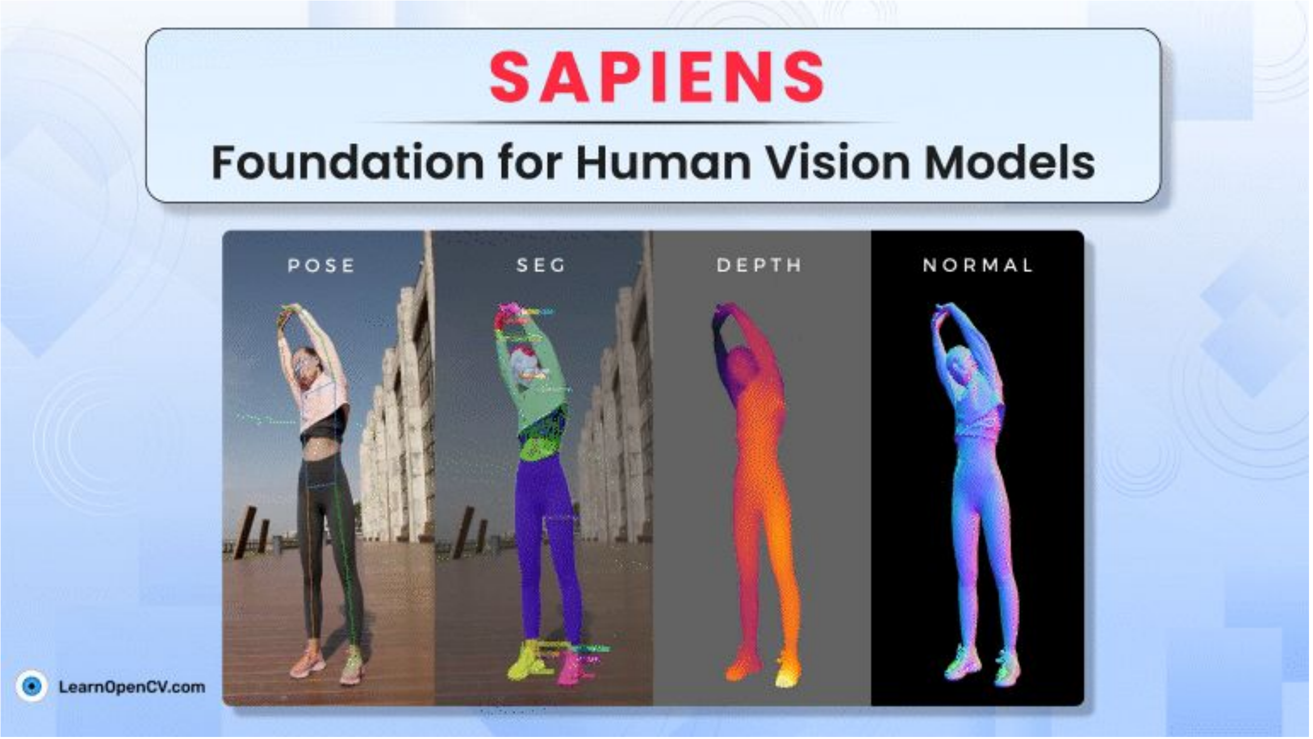


(a) input video

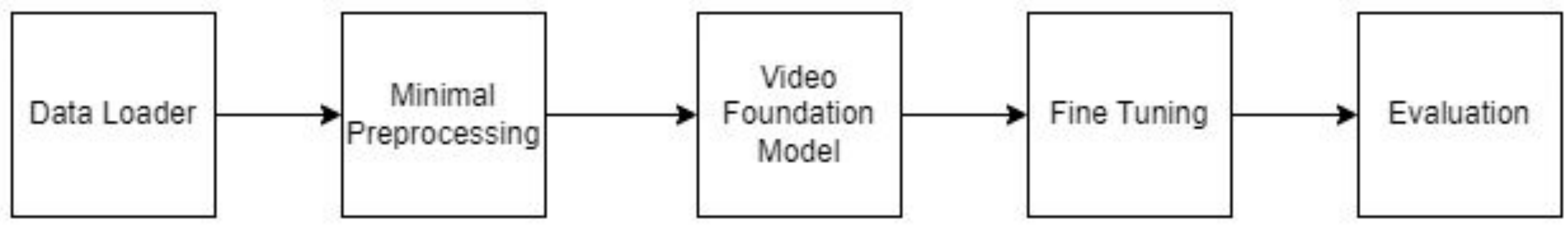
(c) random masking

Video MAE

# Opportunities for us

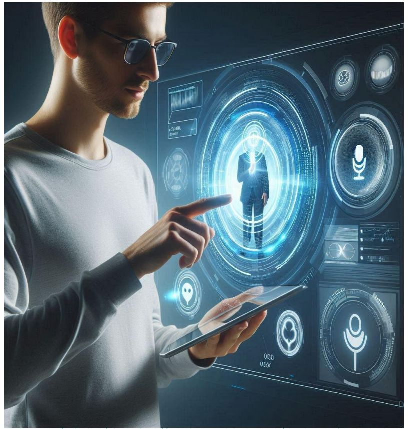


# Block diagram



# Applications of HAR:

Fraud Detection



Human-Computer Interaction

Theft Detections in Supermarkets

Health Care Surveillance

Crime Surveillance

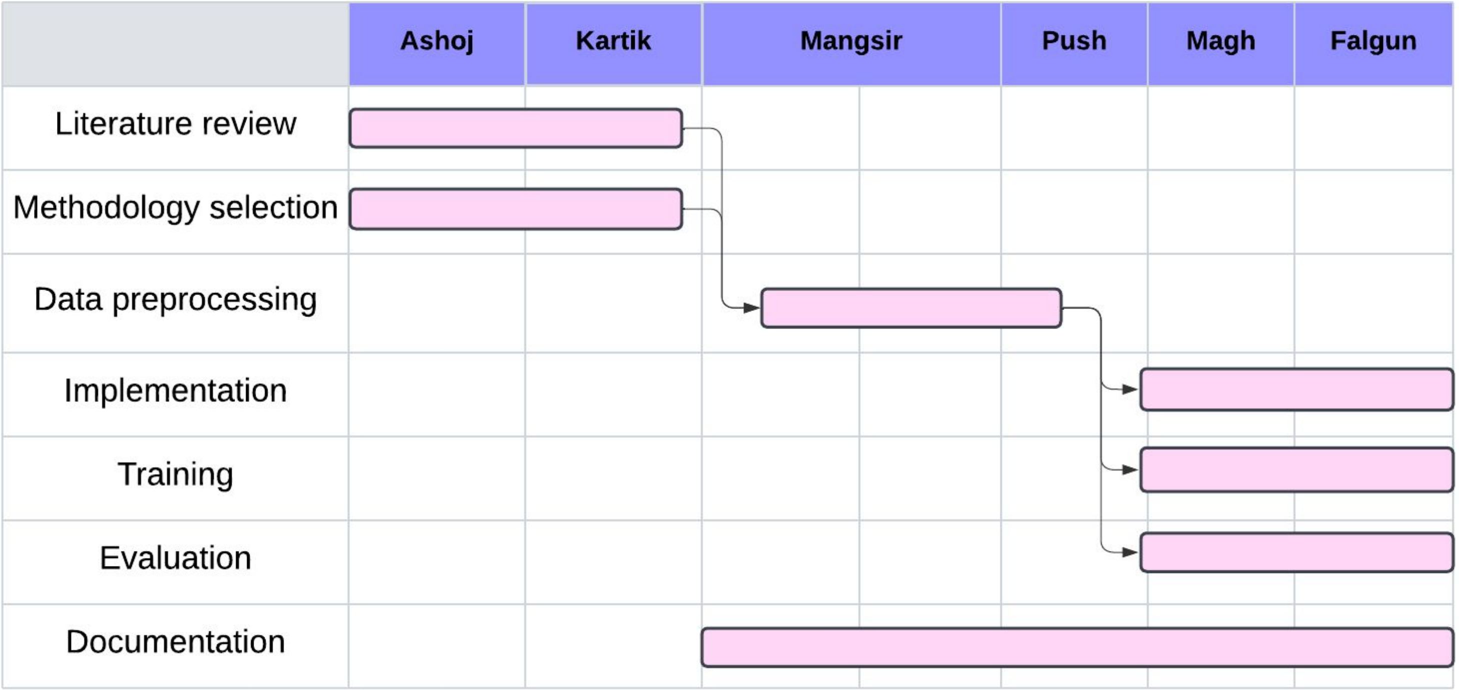
Sports Analysis

Accident Detection

Video Processing

# Timeline

## Human Action Recognition



Ashoj

Kartik

Mangsir

Push

Magh

Falgun

Literature review

Methodology selection

Data preprocessing

Implementation

Training

Evaluation

Documentation

# References

- **Human Action Recognition and Prediction: A Survey**: https://arxiv.org/abs/1806.11230
- **Revisiting 3D ResNets for Video Recognition**: https://arxiv.org/pdf/2109.01696v1
- **AssembleNet++: Assembling Modality Representations via Attention Connections**: https://arxiv.org/pdf/2008.08072v1
- **InternVideo**: http://arxiv.org/abs/2212.03191
- **VideoMAE**: http://arxiv.org/abs/2203.12602
- **Long Term Recurrent Convolutional Neural Network**: http://arxiv.org/abs/1411.4389
- **V-JEPA**: http://arxiv.org/abs/2301.08243
- **Dataset**: https://paperswithcode.com/dataset/kinetics
- **Shop Lifting dataset**: https://www.kaggle.com/datasets/mateohervas/dcsass-dataset
- **Sapiens**: https://about.meta.com/realitylabs/codecavatars/sapiens/

# Thank you