# DLCV Final Project-Talking to Me

Po-Wen Hsueh (薛博文), Ping-Mao Huang (黃秉茂), Ya-Ching Hsu (許雅晴), Shang-Yen Lee (李尚宴)

## Introduction

The TTM Challenge is a task that focuses on improving multimodal perception in egocentric video using the Ego4D dataset. Through this task, AI can learn from daily life experiences around the world by observing what we do.

## Pre-processing

First, the human face is cropped from each frame. Next, a selection of frames is chosen and data augmentation techniques such as RandomCrop, and ColorJitter are applied. To maintain the temporal relationship, we combined the selected frames into a series and input them into the vision encoder. This process transforms the temporal information into the spatial domain, allowing the model to learn to recognize patterns within the data.

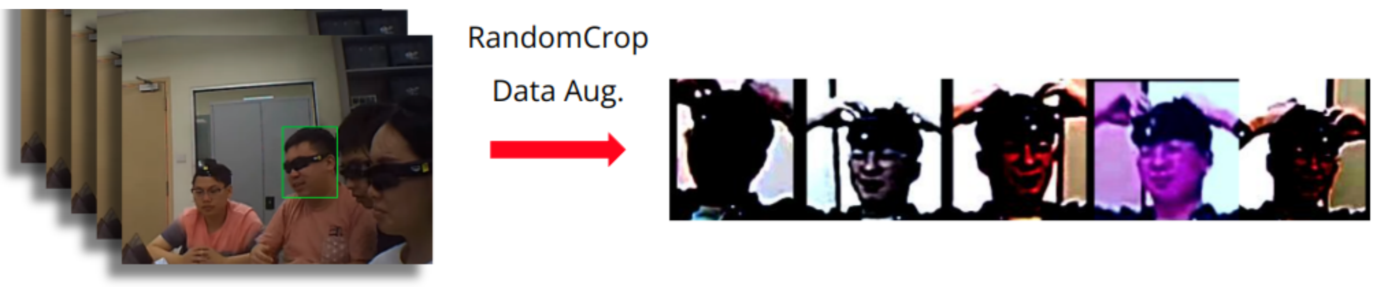

Figure 1: A sample of frames after data augmentation

## Model

We partitioned the model into two components: one that processes the visual information from the images. And another that handles the audio data extracted from the videos. We try to use different backbones as vision encoders, such as ResNet50, Swin Transformers. For the audio encoder, we try to use wav2vec, and some BERT-Style model.

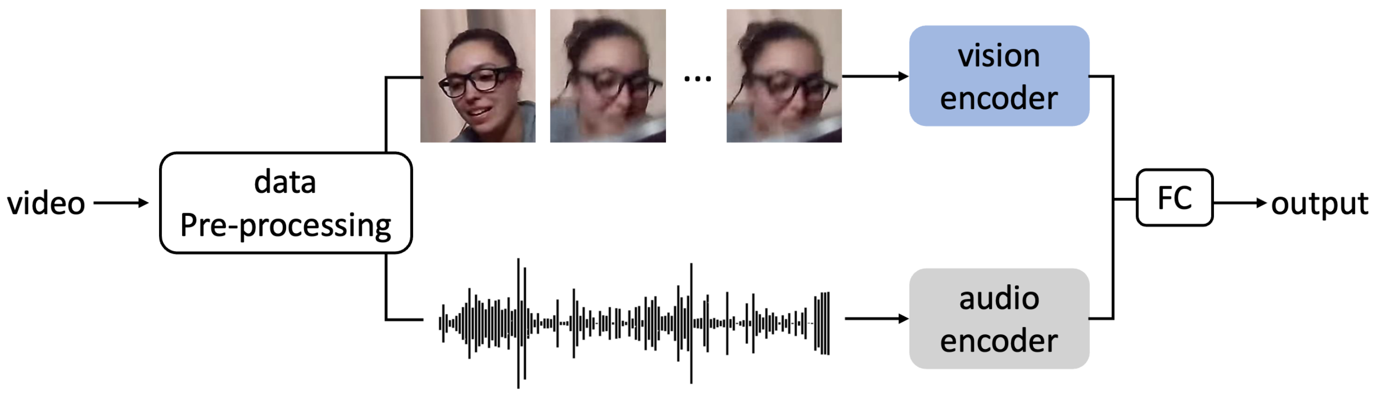

Figure 2: model structure

## Vision

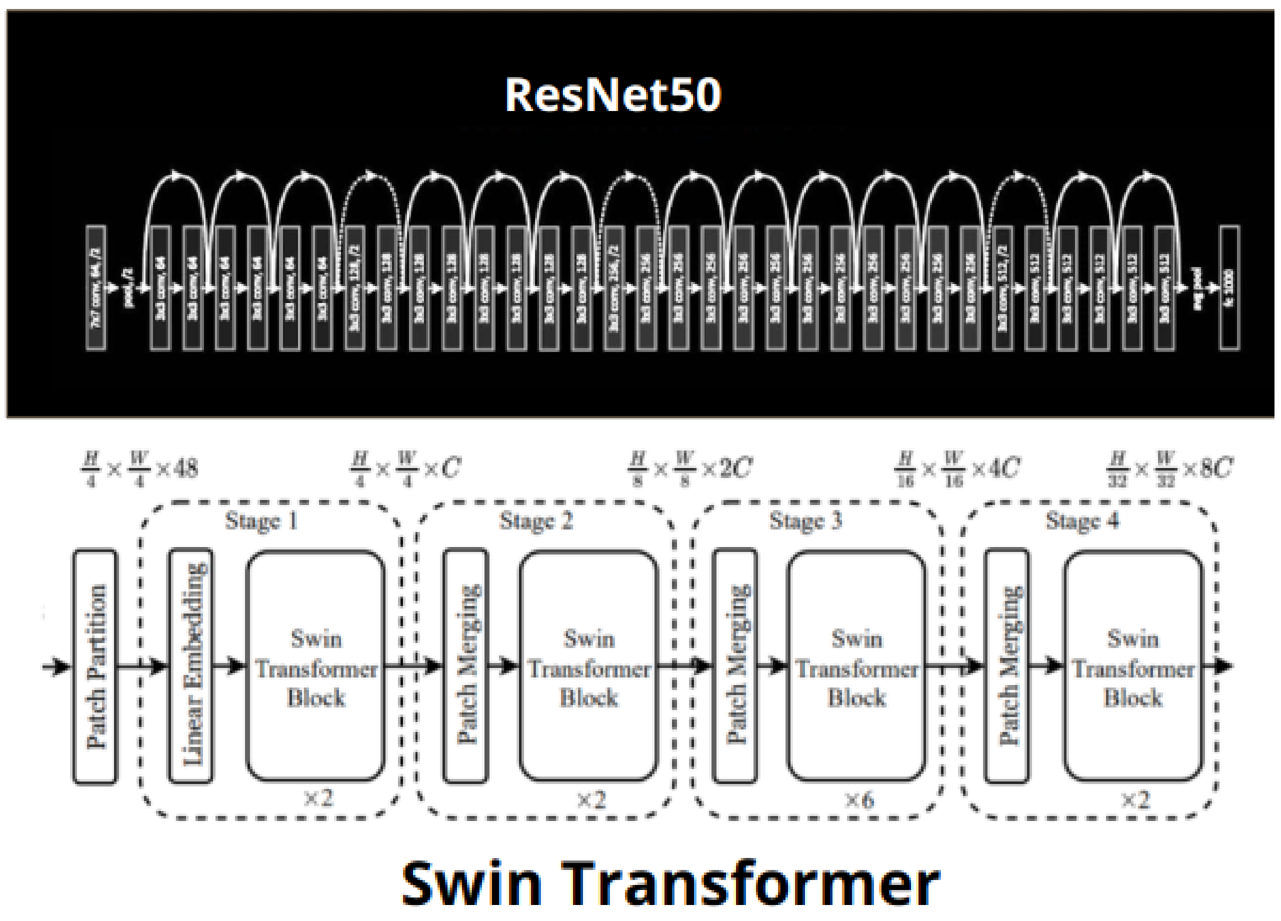
ResNet50
$\frac{H}{4} \times \frac{W}{4} \times 48$
$\frac{H}{4} \times \frac{W}{4} \times C$
$\frac{H}{8} \times \frac{W}{8} \times 2C$
$\frac{H}{16} \times \frac{W}{16} \times 4C$
$\frac{H}{32} \times \frac{W}{32} \times 8C$
Stage 1
Stage 2
Stage 3
Stage 4
Patch Partition
Linear Embedding
Swin Transformer Block
Patch Merging
×2
×2
×6
×2
Swin Transformer

Figure 3: vision models

## Audio

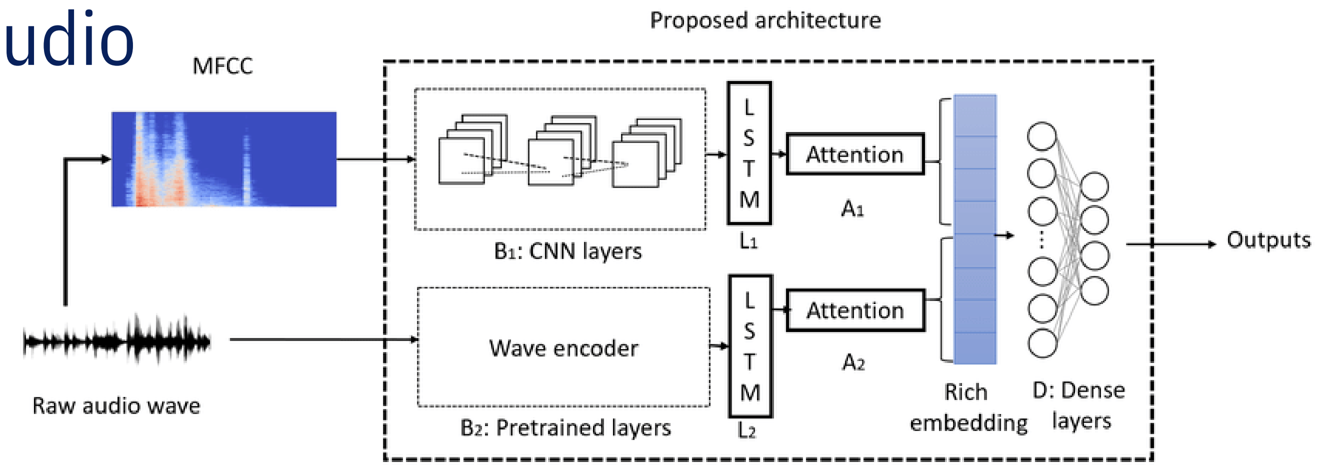

Figure 4: speaker identification system with CNN

(ref: https://www.sciencedirect.com/science/article/abs/pii/S0045790622001719?via%3Dihub)

## Fusion

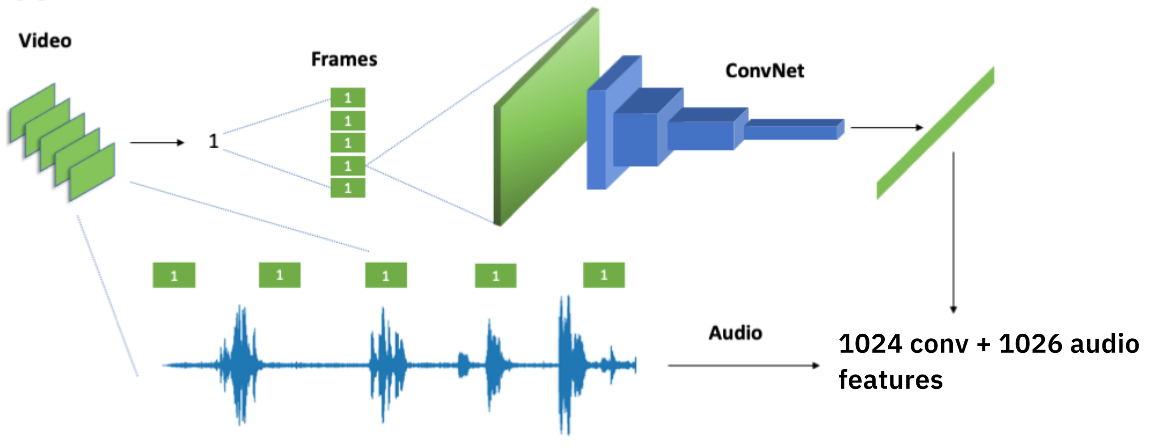

Figure 5: combining Image with Audio features

(ref: https://francescopochetti.com/video-classification-experiments-combining-image-with-audio-features/)

## Results and discussion

Vision and audio models can significantly improve the accuracy of a task. In this case, the use of DL models increased accuracy by over 14% compared to the baseline.

In our experiments with vision models, we found that using a cosine scheduler leads to better performance. While increasing the image size to 224 or the learning rate to 1e-4 may also be helpful.

| model | image size | sample | batch size | scheduler | accuracy (val) |
|---|---|---|---|---|---|
| swin t | 128 | 5 | 64 | exp | 0.7000 |
| resnet18 | 128 | 5 | 64 | exp | 0.6892 |
| resnet50 | 128 | 5 | 64 | exp | 0.6941 |
| swin t | 224 | 5 | 32 | exp | 0.6740 |
| resnet18 | 224 | 5 | 32 | exp | 0.6944 |
| resnet50 | 224 | 5 | 32 | exp | 0.6903 |
| swin t | 128 | 10 | 32 | exp | 0.6907 |
| resnet50 | 128 | 10 | 64 | exp | 0.7015 |
| swin t | 128 | 5 | 64 | cos | 0.6996 |
| resnet50 | 128 | 5 | 64 | cos | 0.7100 |
| swin t | 128 | 5 | 64 | exp | 0.5465 |
| resnet18 | 128 | 5 | 64 | exp | 0.6788 |
| resnet50 | 128 | 5 | 64 | exp | 0.6881 |

Table1 : performance comparison among configuration

| | vision (swin t, resnet) | audio (wave encoder) | fusion |
|---|---|---|---|
| accuracy (val) | 0.710 | 0.673 | 0.722 |

Table2 : model ablation study

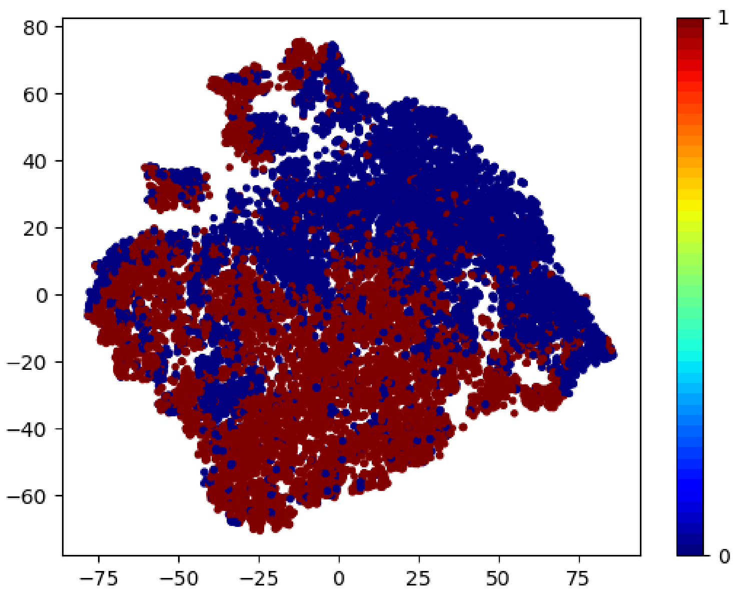

Figure 6: T-SNE of fusion features

## Conclusions

- We applied a multimodal neural network that concatenates the visual representation and the audio representation for final classification, resulting in improved performance.
- Our proposed approach involves preprocessing the video data and transforming the temporal information into the spatial domain, potentially leading to outstanding performance with a smaller number of sampled data.

## References

Kristen Grauman et al. "Ego4D: Around the World in 3, 000 Hours of Egocentric Video". In: CoRR abs/2110.07058 (2021). arXiv: 2110.07058. URL: https://arxiv.org/abs/2110.07058.