MSc in Computational Software Techniques in Engineering

Video Vision Transformer-based Method for Visualising Scene Perception

Background:

- Vision Transformers (ViT) have revolutionised computer vision, particularly excelling in image classification tasks.
- The success of ViT has sparked interest in extending its application to more complex domains, such as video analysis.
- Video Vision Transformers (ViViT) offer significant advantages in capturing both spatial and temporal information in videos.
- ViViT overcomes the limitations of traditional Convolutional Neural Networks (CNNs) in capturing long-term dependencies and reducing computational demands.

Aim & Objectives:

- Develop a ViViT-based methodology for effective video classification, integrating 3D CNNs for spatiotemporal feature capture.
- Implement the Swin Transformer to leverage hierarchical and global feature extraction capabilities.
- Enhance model explainability through visualisation techniques such as key frame identification and feature importance analysis.
- Compare the performance of the developed method with state-of-the-art techniques to identify strengths and potential improvements.
- Apply the model to diverse real-world scenarios, demonstrating its practical applicability across various domains.

Methodology:

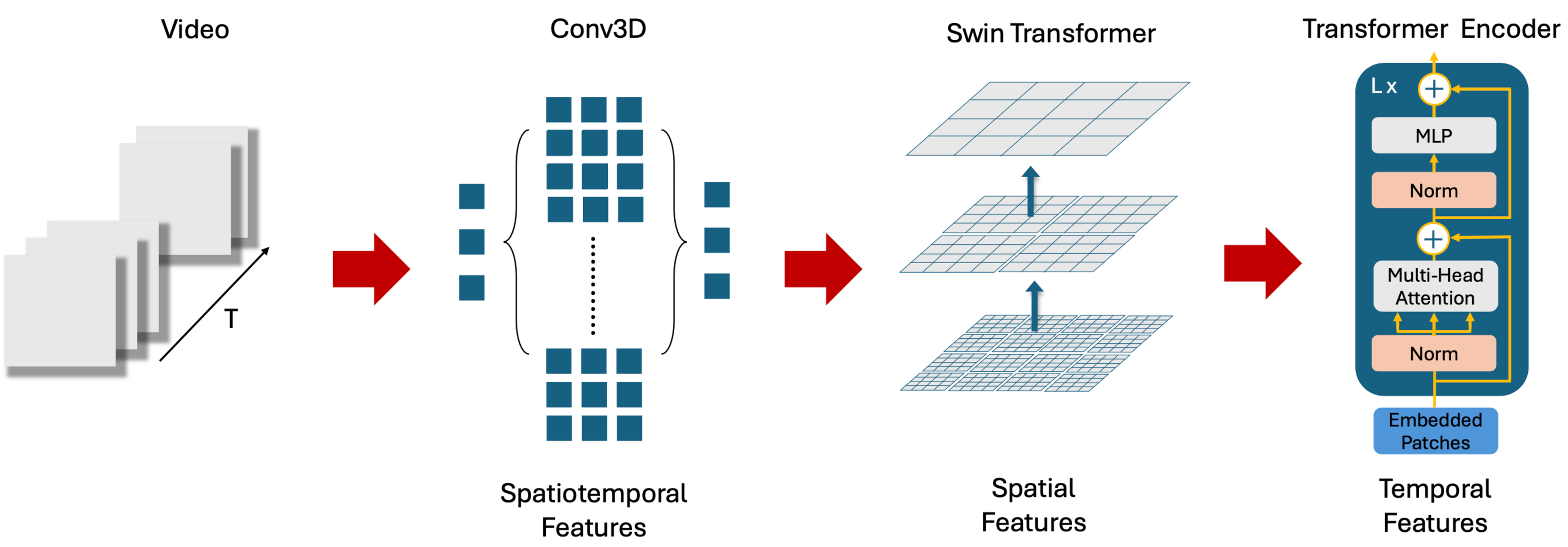
◆ Research Process Overview

This research involved a structured approach to video classification, focusing on the integration of Conv3D and the Swin Transformer within the proposed Dual Encoder model.

◆ Data Collection and Preprocessing

- Video Data: Raw video data from the single-pilot flight simulation was preprocessed into the model input format.
- Data Augmentation: Techniques such as horizontal flipping, rotation, and color jittering were applied to increase model robustness.
- Dataset Split: The dataset was divided into 80% for training and 20% for testing, ensuring validation on unseen data.

◆ Model Architecture



◆ Training Strategy

The model was initialised with pre-trained weights, with appropriate learning rates, optimizers, and regularisation techniques to ensure effective convergence and prevent overfitting.

◆ Evaluation and Comparison

The model was comprehensively evaluated using metrics such as accuracy, inference time, and parameter count, and compared with baseline models to validate the proposed model's advantages.

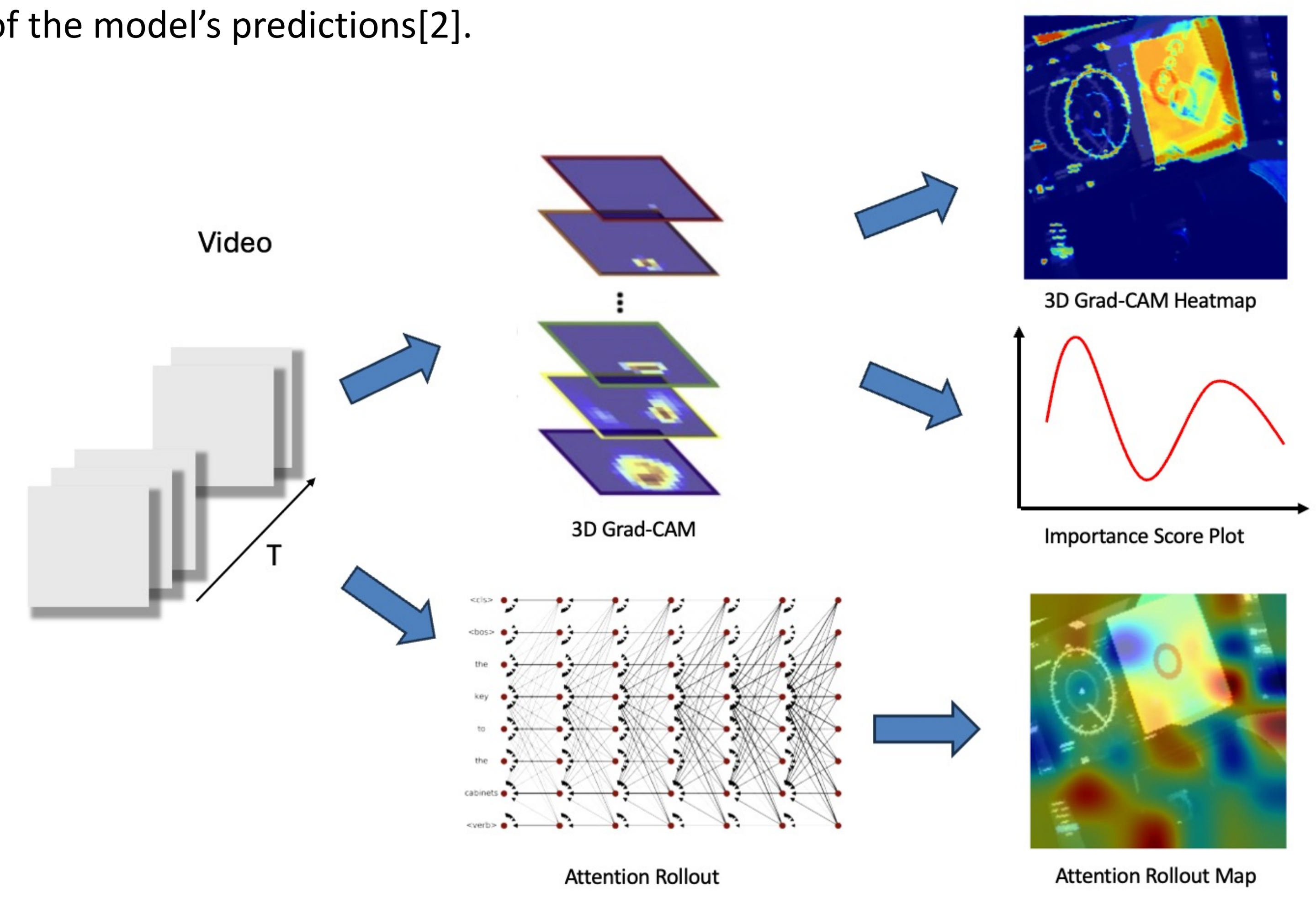
◆ Visualisation Techniques

1. Attention Rollout:

Aimed at visualising how attention is distributed across layers in the model, helping to understand the internal decision-making process[1].

2. 3D Grad-CAM:

Identify key spatial-temporal regions in the video, further validating the reasonableness of the model's predictions[2].



Result:

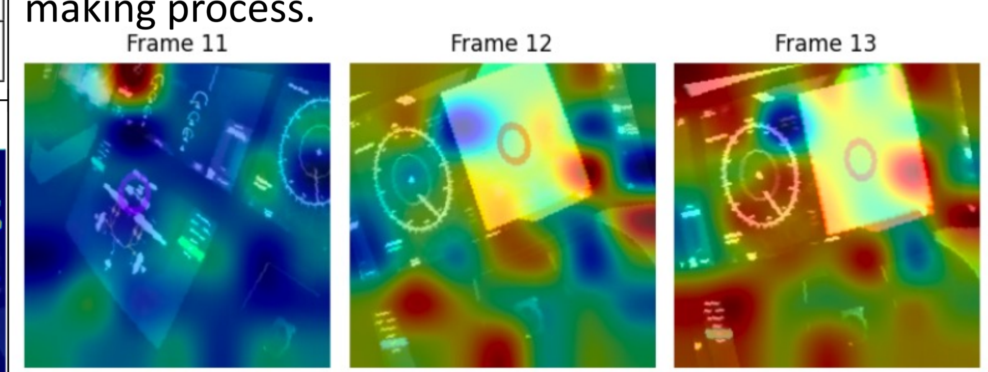
Model Comparison

The Dual Encoder model outperformed the Single Encoder and ResNet3D18 models, achieving higher accuracy in video classification.

Model	GMac	Parameters (M)	Running Time	Accuracy (%)
Dual Encoder	99.47	95.15	49 min 17 s	87.50
Single Encoder	99.2	86.75	43 min 4 s	80.03
ResNet3D18	13.87	33.14	44 min 22 s	86.66

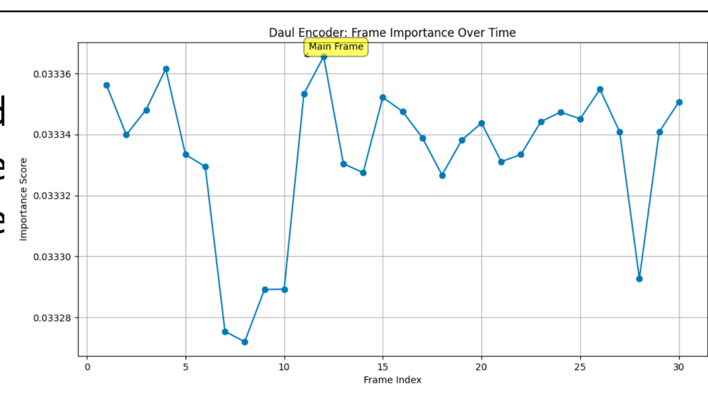
Attention Rollout Maps

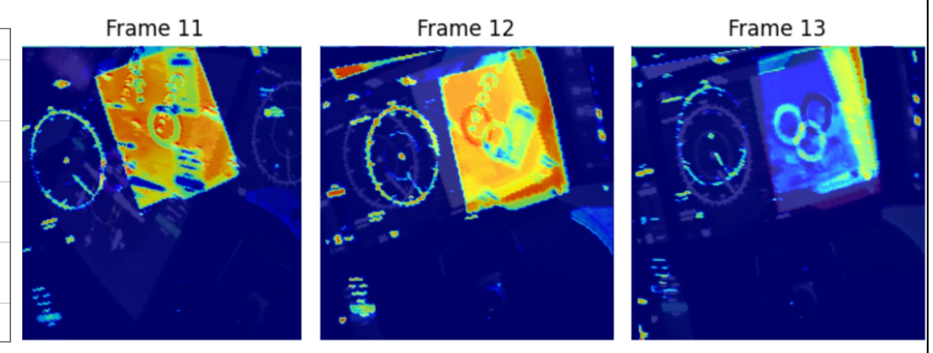
Visualised spatial attention shifts around key frames, highlighting regions of focus in the model's decision-making process.



3D Grad-CAM & Importance Scores

Identified key frames and calculated their Importance Scores, showing these frames are critical for accurate predictions.





Conclusion:

◆ Overall Findings

- Successfully developed a Dual Encoder model integrating Conv3D and the Swin Transformer for enhanced video classification.
- Demonstrated superior performance over traditional models, effectively capturing both spatial and temporal features.
- Utilised 3D Grad-CAM and Attention Rollout to gain deep insights into the model's decision-making process, emphasizing key frames and spatial attention in video analysis.

◆ Future Work

- **Model Expansion:** Experiment with larger and more diverse datasets to enhance the model's generalization capabilities.
- **Architectural Enhancements:** Investigate hybrid models that incorporate additional elements, such as RNNs, for capturing even more intricate temporal dependencies.
- **Advanced Visualisation:** Develop new or enhanced visualisation techniques to further improve interpretability and provide deeper insights into model decisions.
- **Robustness Improvement:** Address variability in video data by introducing advanced data augmentation techniques and novel training strategies to improve model robustness in diverse real-world scenarios.

Author: Chung-Yueh Cheng

Supervisors: Dr Lichao Yang, Prof Yifan Zhao

Reference:

[1] Abnar, S., & Zuidema, W. (2020, May 2). *Quantifying attention flow in transformers*. ArXiv (Cornell University); Cornell University. https://doi.org/10.48550/arxiv.2005.00928

[2] Zhang, Y., Hong, D., McClement, D., Oladosu, O., Pridham, G., & Slaney, G. (2021). Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods*, *353*, 109098. https://doi.org/10.1016/j.jneumeth.2021.109098