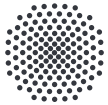


University of Stuttgart
Germany

Gaze-based Transformer for Improving Action Recognition in Egocentric Videos

Master Thesis

Mengze Lu

July 14, 2024

Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

Table of Contents

Introduction

Related Works

Video Swin Transformer

State-of-the-Art Transformer EgoViT

Gaze Enhanced EgoViT

The Structure of Gaze Enhanced EgoViT

Key Components

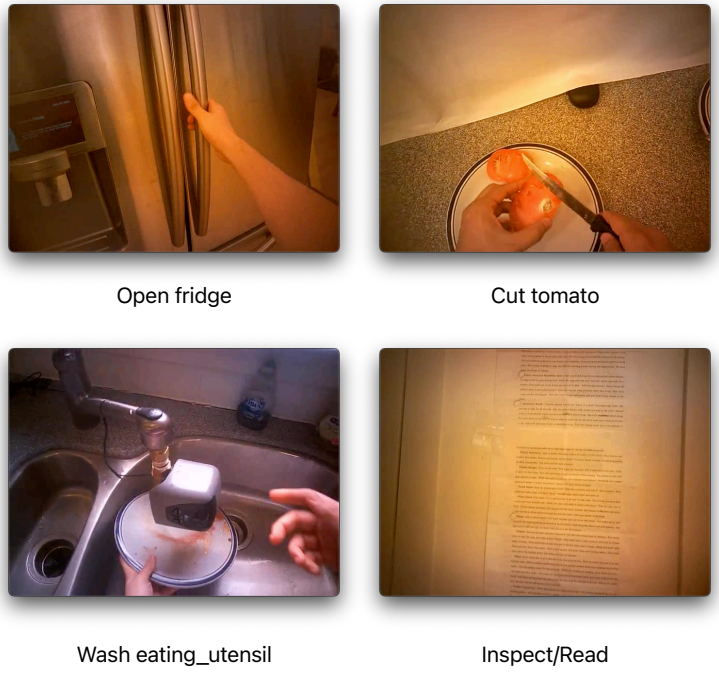
Experiments and Results

Conclusion and Future Works

Action Recognition

Task:

- Identify and categorize human actions into predefined classes.



Source: Videos with action label

Egocentric Video

Source: Li et al. [2020] Eye tracking glasses used in video recording.

Key Features:

- First-person View (FPV)
- Hand and object interactions
- Large scene changes and diverse backgrounds

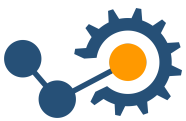
Significance of Gaze in Action Recognition

Gaze information:

- Crucial for understanding human intention
- Closely linked to object-oriented actions

Potential of gaze information:

- Enhances the precision of action recognition



The Goal of this Thesis

Objective: Improve the accuracy of transformers for action recognition by integrating gaze data.

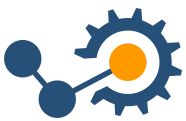


Table of Contents

Introduction

Related Works

Video Swin Transformer

State-of-the-Art Transformer EgoViT

Gaze Enhanced EgoViT

The Structure of Gaze Enhanced EgoViT

Key Components

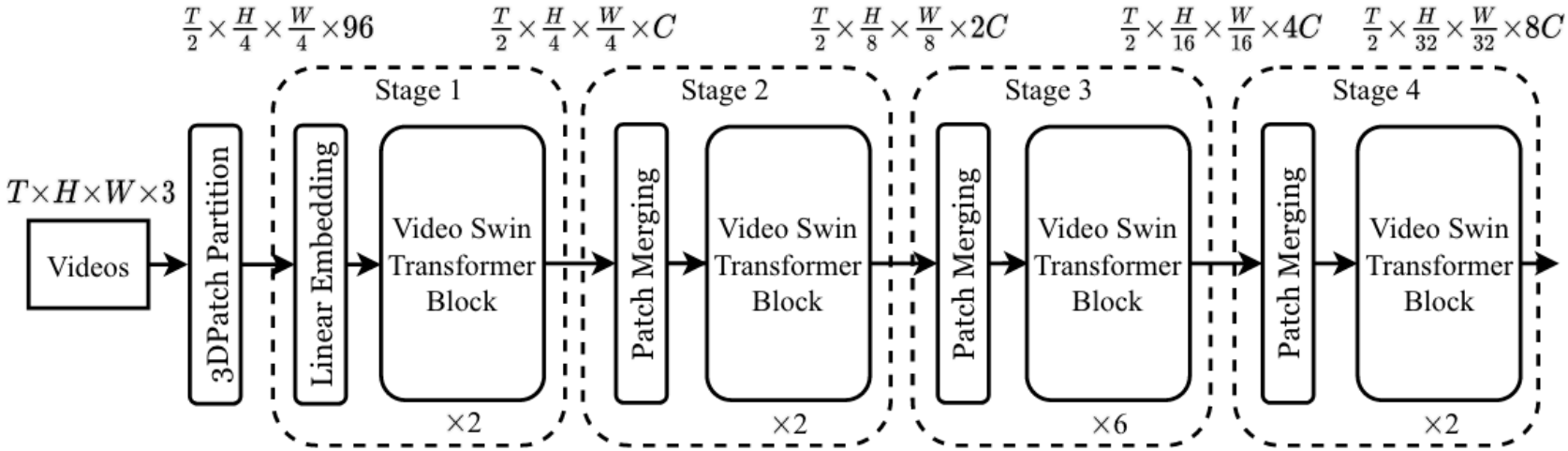
Experiments and Results

Conclusion and Future Works

Related Works

Video Swin Transformers

Video Swin Transformer



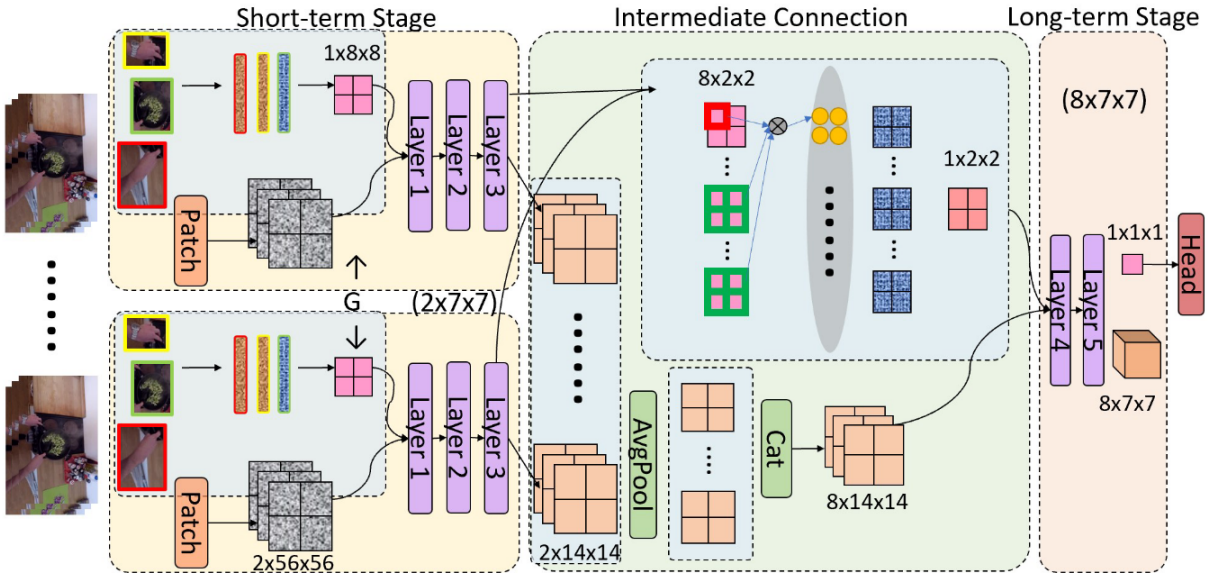
Source: Overview of the Video Swin Transformer Architecture. Liu et al. [2021]

- Processes inputs through a series of Swin Transformer blocks
- 3D Shifted Window Self-Attention
- Used as the backbone in the proposed model.

Related Works

State-of-the-Art Vision Transformer EgoViT

State-of-the-Art Vision Transformer EgoViT



Source: Pan [2023] The structure of EgoViT.

Features of the Structure:

- Designed for egocentric video
- Seamlessly integrate with different vision transformers

State-of-the-Art Vision Transformer EgoViT

Contribution of EgoViT:

- Incorporates hand-object interaction features.
- Dynamic class token (DCT): Focus on the informative parts
- Pyramid Architecture with Dynamic Merging (PADM): Effective processing large movements

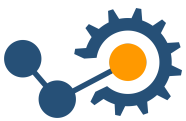


Table of Contents

Introduction

Related Works

- Video Swin Transformer
- State-of-the-Art Transformer EgoViT

Gaze Enhanced EgoViT

- The Structure of Gaze Enhanced EgoViT
- Key Components

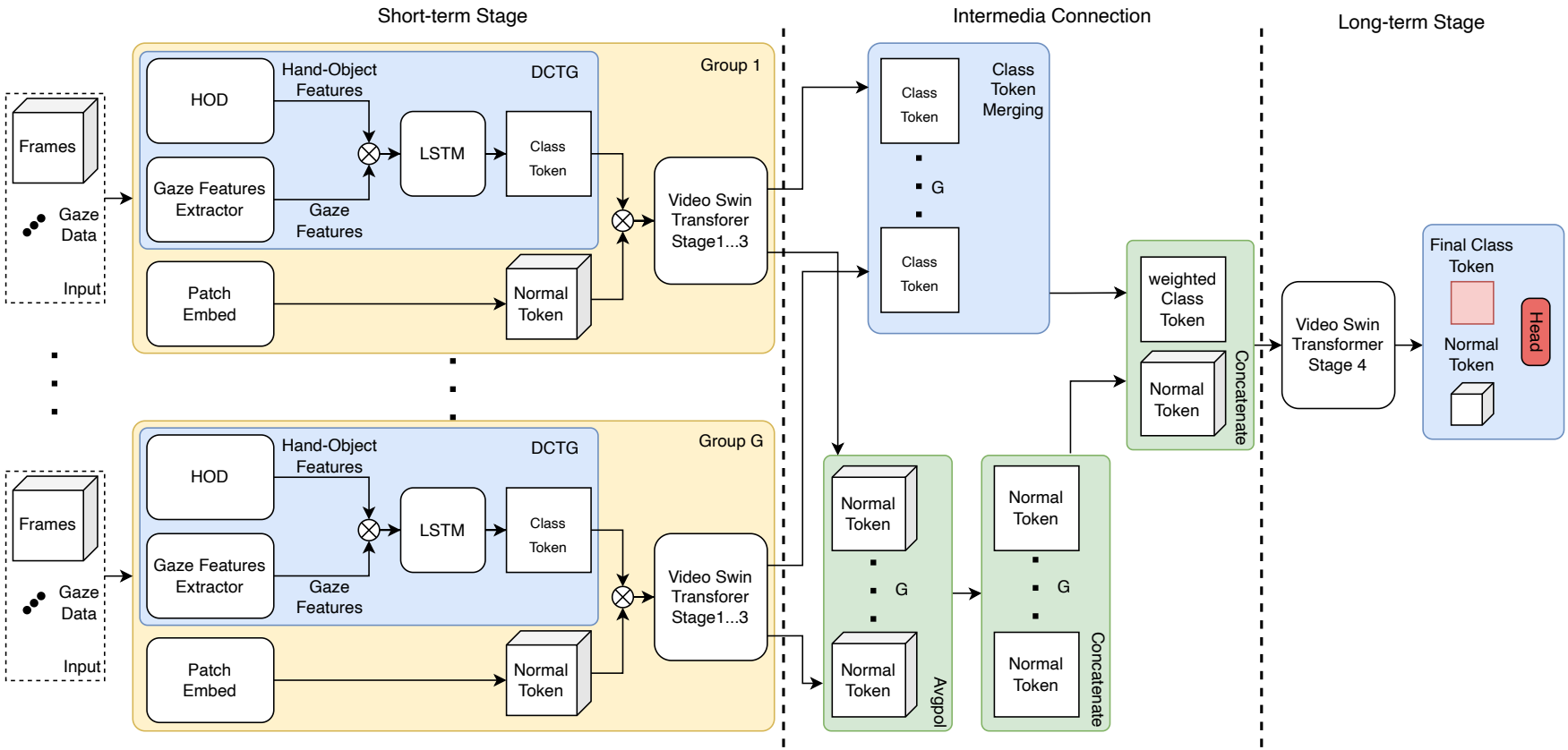
Experiments and Results

Conclusion and Future Works

Gaze Enhanced EgoViT

The Structure of Gaze Enhanced EgoViT

The Structure of Gaze Enhanced EgoViT



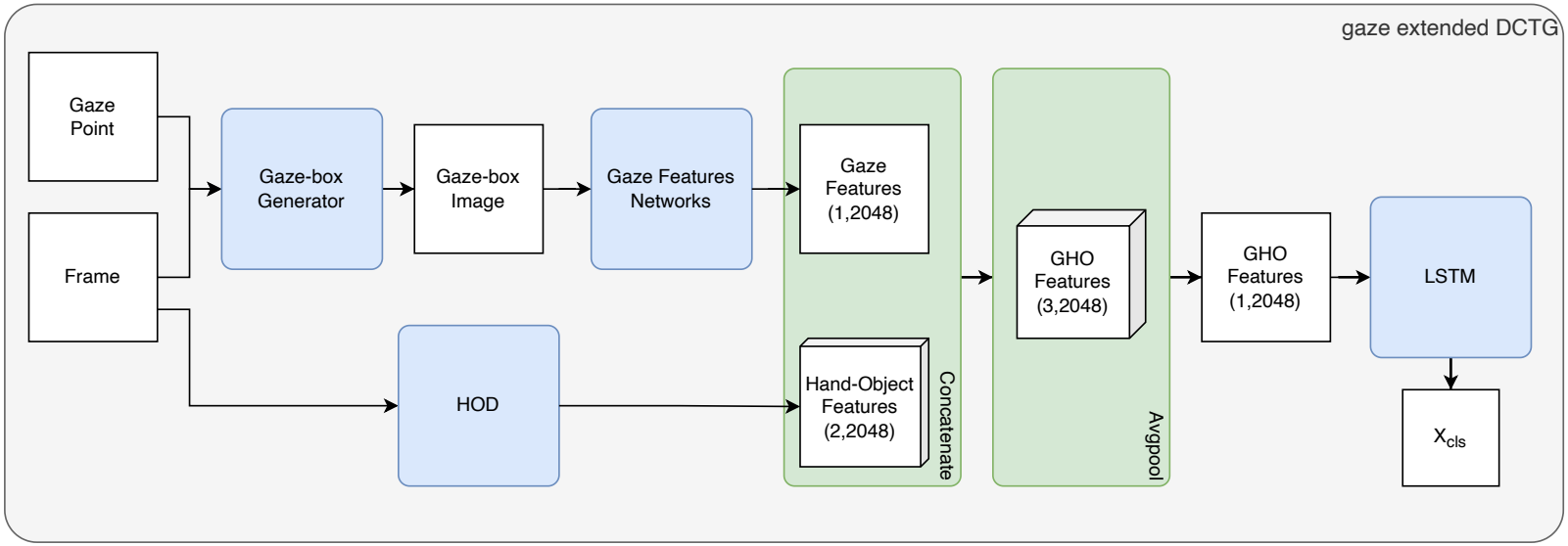
Source: The Structure of Gaze Enhanced EgoViT

- Architecture: Based on EgoViT
- DCTG extended with gaze features extract module

Gaze Enhanced EgoViT

Key Components

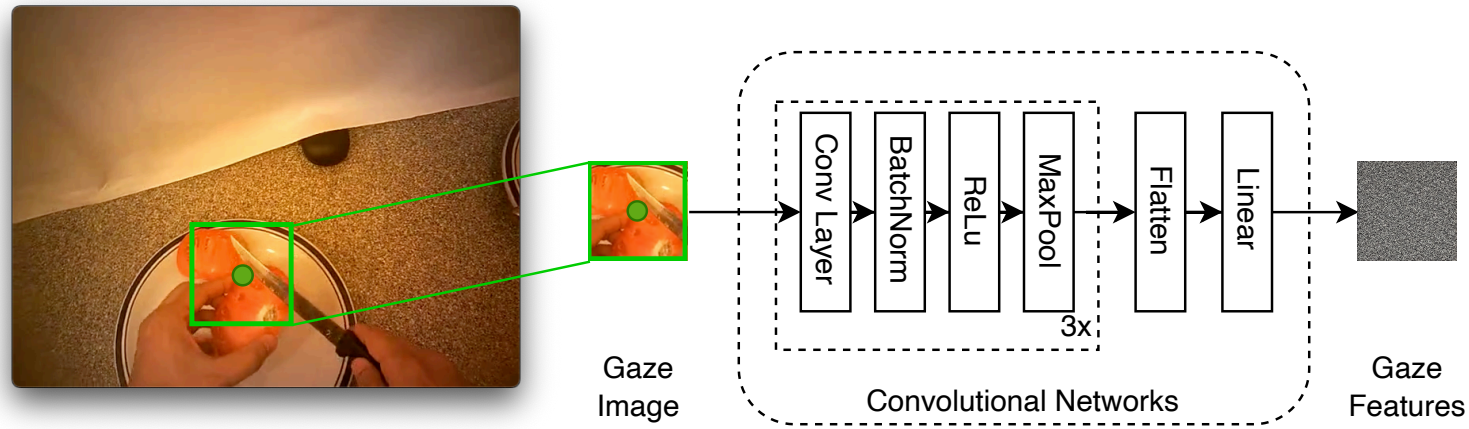
Gaze Enhanced DCTG



Source: Illustration of the Dynamic Class Token Generator with Gaze Information.

- G features and HO features are fused to form GHO features.
- Employs an LSTM to generate the class token, enhancing temporal coherence.

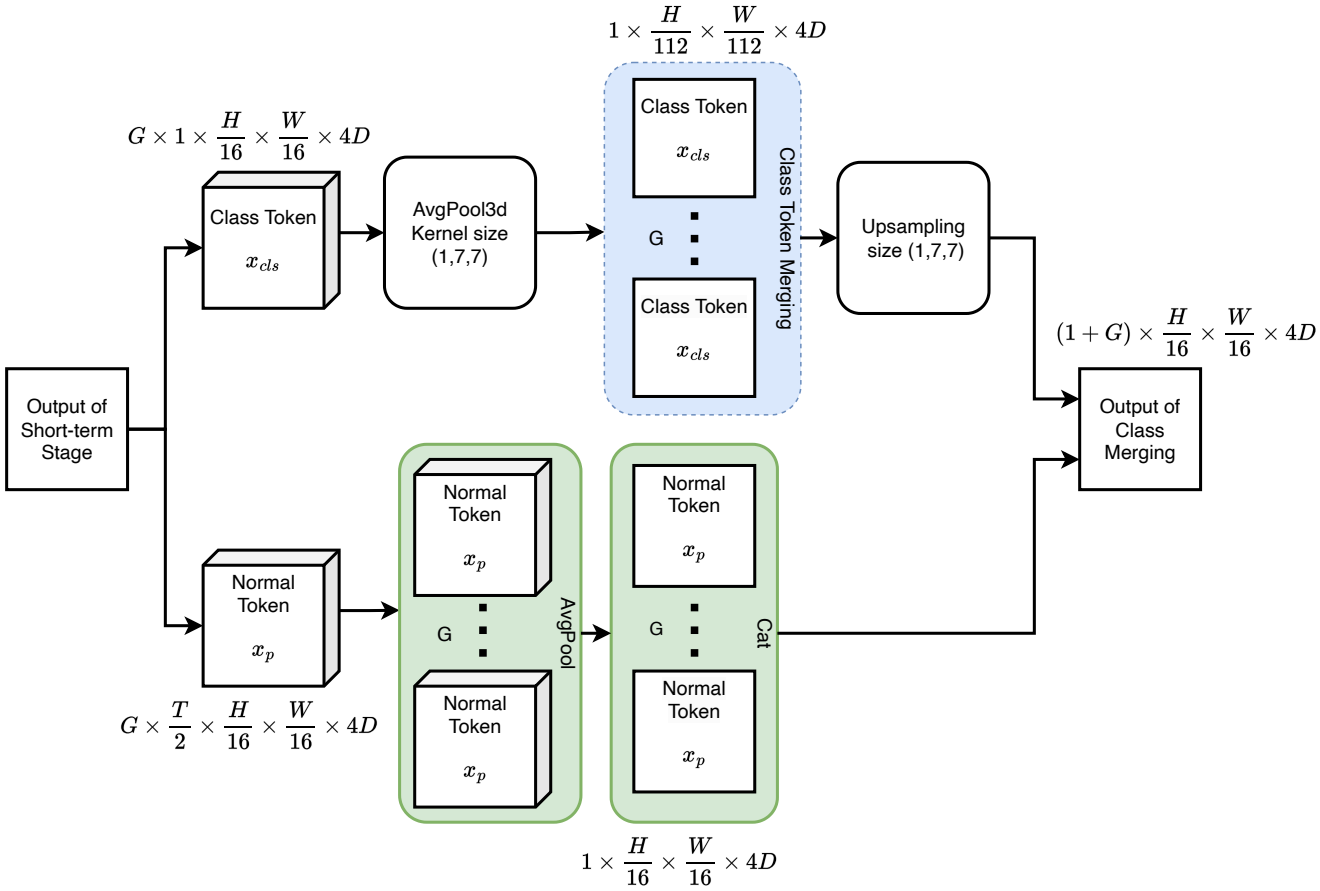
Gaze Features Extraction



Source: Illustration of the Gaze Features Extraction.

Gaze point → gaze-box image → Conv Networks → gaze features

Dynamic Merging Module



Source: Illustration of the Dynamic Merging Module.

- The class token and normal token are merged separately.
- Input $\rightarrow$ G-x_{cls}, G-x_p $\rightarrow$ Weighted x_{cls}, x_p

Table of Contents

Introduction

Related Works

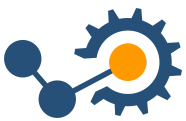
- Video Swin Transformer
- State-of-the-Art Transformer EgoViT

Gaze Enhanced EgoViT

- The Structure of Gaze Enhanced EgoViT
- Key Components

Experiments and Results

Conclusion and Future Works



Dataset

EGTEA Gaze+:A large-scale dataset for FPV actions and gaze analysis

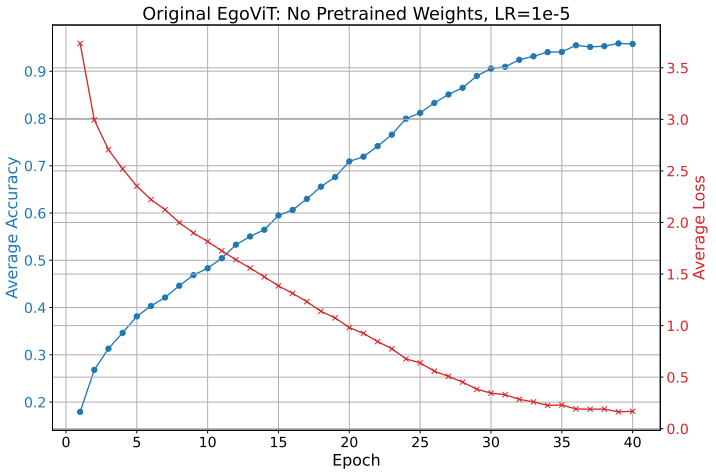
- Gaze tracking data available at 30Fps
- Frame-level action annotations

Source: Li et al. [2020] A frame from the dataset showing gaze point.

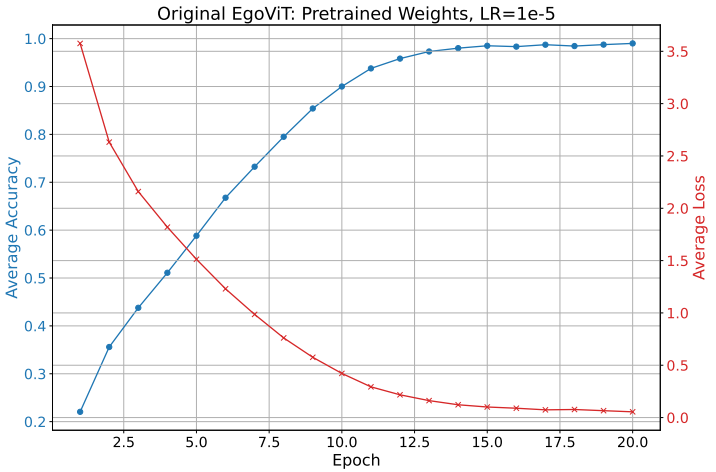
Dataset

- Gaze-Hand-Object (GHO) features are preprocessed offline.
- Each video is sampled at an average of 32 frames.
- Dataset structure:
  (Frames[32,3,224,224], GHO-Features[32,3,2048], Label[1])

Training the Original EgoViT



Source: Training accuracy and loss of the original EgoViT without pretrained weights.



Source: Training accuracy and loss of the original EgoViT with pretrained weights.

- Both models reached similar training accuracy and loss.
- The model with pretrained weights converges faster.

Test Results of original EgoViT with HO Features

Experiment ID	Top-1 Acc.(%)	Top-5 Acc.(%)	Mean Class Acc.(%)
Orig_HO_no_pretrain	51.5	78.5	38.8
Orig_HO	51.7	75.2	40.6

Discussion:

- The model with pretrained weights has higher top-1 and mean class accuracy
- Pretrained weights reduce the training time significantly

Test Results of Original EgoViT with HO Features

Experiment ID	Top-1 Acc.	Top-5 Acc.	Mean Class Acc.
Orig_HO	51.7	75.2	40.6
Enh_GHO	51.4	76.7	40.0
Enh_G	48.9	75.1	37.7

Discussion:

- Enh_GHO: slightly lower top-1 and mean class accuracy than Orig_HO.
- Enh_G: Lowest accuracy among the three models.
- The gaze features need further improvement.

Improve the quantity of the gaze data

- Gaze version 1 (GHO & G): Uses only fixation gaze tracking data.
- Gaze version 2 (GHO_v2 & G_v2): Uses both fixation and saccade gaze tracking data.
- Gaze version 2 has fewer random gaze features, enhancing the overall data quality.

Training the Gaze Enhanced EgoViT

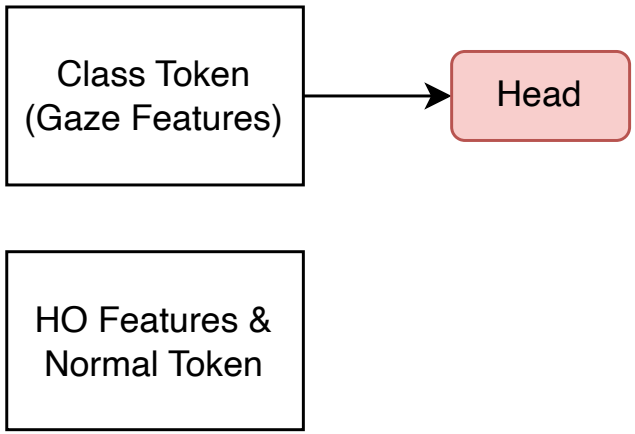
Test Results on Gaze Version 1 and Gaze Version 2

Experiment ID	Top-1 Acc.(%)	Top-5 Acc.(%)	Mean Class Acc.(%)
Enh_GHO	51.4	76.7	40.0
Enh_G	48.9	75.1	37.7
Enh_GHO_v2	52.0	76.3	38.7
Enh_G_v2	50.0	76.2	38.0

Discussion:

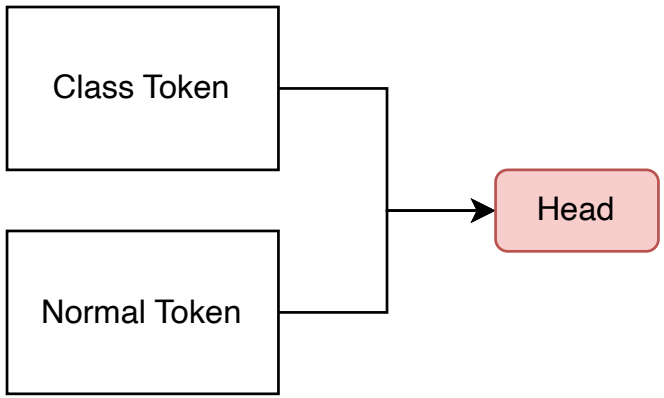
- Enh_G_v2 shows significant improvement in all three metrics.
- The quality of gaze features is important for action recognition.

Training EgoViT Variants



Source: Enhanced EgoViT Version 2

- Enh_v2_GHO_: Uses only gaze features for classification.



Source: EgoViT Version 3

- Orig_v3, Enh_v3: Use class token and normal token for classification.

Training EgoViT Variants

Test Results on EgoViT Variants

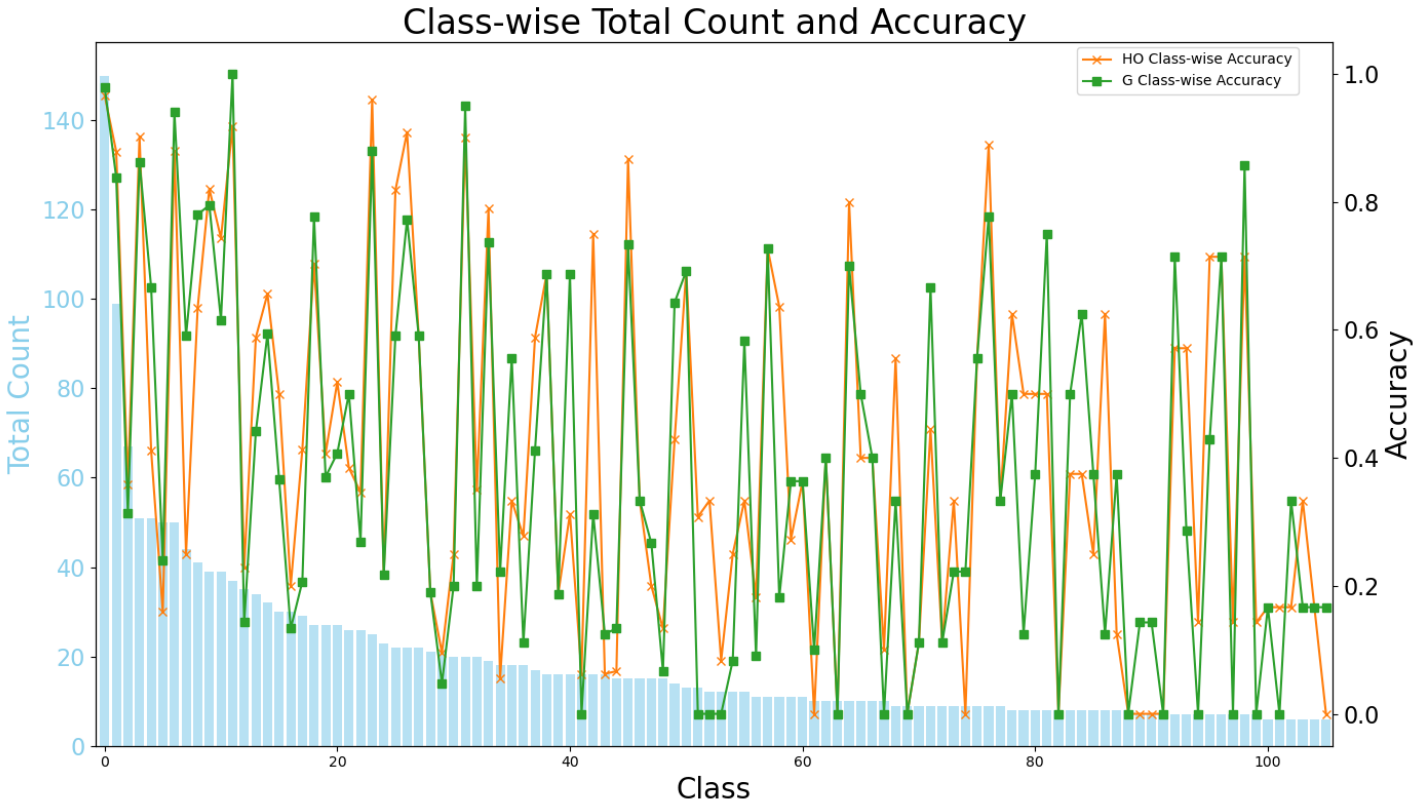
Training Method	Top-1 Acc.(%)	Top-5 Acc.(%)	Mean Class Acc.(%)
Enh_GHO_v2	52.0	76.3	38.7
Enh_v2_GHO_v2	50.4	75.8	38.3
Enh_v3_GHO_v2	51.2	77.9	39.2
Orig_HO	51.7	75.2	40.6
Orig_v3_HO	51.1	77.4	40.5
Enh_G_v2	50.0	76.2	38.0
Enh_v3_G_v2	50.8	75.9	37.8

Discussion:

- Enh_v2 shows lower performance
- Enh_v3 and Orig_v3 improve only top-5 accuracy

Possible Reason: Information in the class token and normal token is already exchanged in 3D Shifted Window Self-Attention.

Distribution of Dataset



Source: Visualization of the dataset and model predictions.

- The dataset is imbalanced, leading to overfitting on the majority class.
- The model with gaze features performs poorly in predicting movement actions.

Table of Contents

Introduction

Related Works

- Video Swin Transformer
- State-of-the-Art Transformer EgoViT

Gaze Enhanced EgoViT

- The Structure of Gaze Enhanced EgoViT
- Key Components

Experiments and Results

Conclusion and Future Works

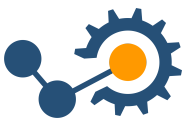
Future Works

Conclusion:

- Additional gaze features could improve the model's performance.
- The quality of the gaze features is Crucial for improving the accuracy of action recognition.

Future Works:

- Study the impact of gaze region size.
- Use a pretrained network, such as an Autoencoder, to generate gaze features.



References i

Y. Li, M. Liu, and J. M. Rehg. In the eye of the beholder: Gaze and actions in first person video. *arXiv:2006.00626 [cs]*, Oct 2020. URL `http://arxiv.org/abs/2006.00626`.

Z. Liu et al. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. URL `http://arxiv.org/abs/2106.13230`.

C. e. a. Pan. Egovit: Pyramid video transformer for egocentric action recognition. *arXiv:2303.08920 [cs]*, 2023. URL `http://arxiv.org/abs/2303.08920`.