THE OHIO STATE UNIVERSITY
Photogrammetric Computer Vision Lab.
Research Lab. directed by Dr. Alper Yilmaz

CascadeFormer: A Family of Two-stage Cascading Transformers for Skeleton-based Human Action Recognition

(Session Code: **P1.T2**)

Yusen Peng, Alper Yilmaz
{peng.1007, yilmaz.15}@osu.edu

Github: https://github.com/Yusen-Peng/CascadeFormer
Huggingface: YusenPeng/CascadeFormerCheckpoints

Motivation

- Graph Convolutional Networks (GCNs) have been dominant in Skeleton-based action recognition [1]
- Transformer models and pretraining frameworks open new avenues for representation learning [2]
- we propose **CascadeFormer**, a family of two-stage cascading transformers

Feature Extraction Module

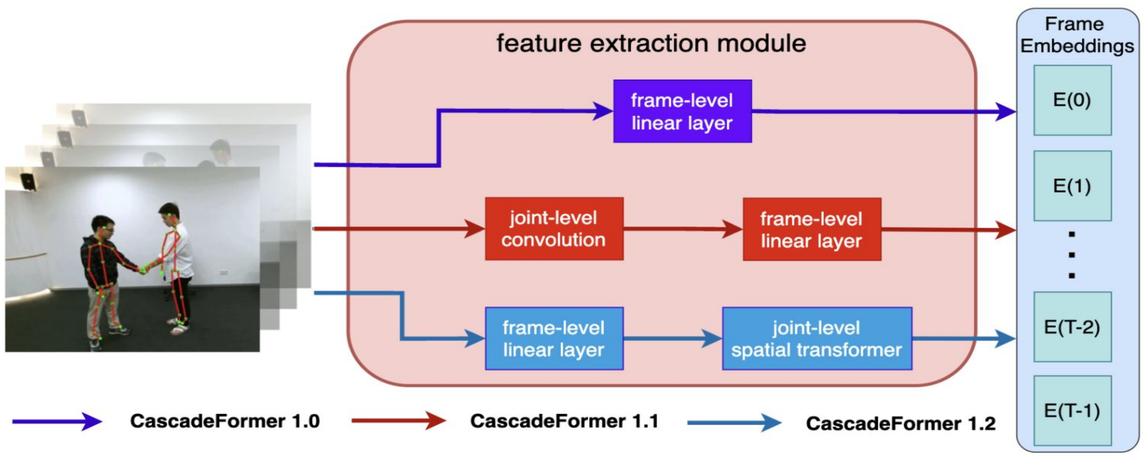

Fig. 3. Feature extraction module in CascadeFormer. All variants convert input skeleton sequences into frame-level embeddings for downstream temporal modeling. CascadeFormer 1.0 (purple) applies a simple frame-level linear projection. CascadeFormer 1.1 (red) enhances this by first applying a joint-level 1D convolution to capture spatial locality before linear projection. CascadeFormer 1.2 (blue) constructs joint-level embeddings via a linear layer, refines them using a joint-level spatial transformer, and aggregates the outputs into frame-level embeddings.

Main results

Model Variant	Penn Action	N-UCLA	NTU60/CS	NTU60/CV
accuracy				
CascadeFormer 1.0	**94.66%**	89.66%	**81.01%**	**88.17%**
CascadeFormer 1.1	94.10%	**91.16%**	79.62%	86.86%
CascadeFormer 1.2	94.10%	90.73%	80.48%	87.24%
precision				
CascadeFormer 1.0	**94.36%**	90.72%	**80.86%**	**88.34%**
CascadeFormer 1.1	93.63%	**91.63%**	79.53%	87.01%
CascadeFormer 1.2	93.94%	91.57%	80.33%	87.46%
F1 score				
CascadeFormer 1.0	**94.18%**	89.59%	**80.82%**	**88.18%**
CascadeFormer 1.1	93.66%	**91.12%**	79.40%	86.84%
CascadeFormer 1.2	93.88%	90.75%	80.28%	87.26%

Comparison against SOTA

Table 3. Penn Action comparison. CascadeFormer achieves the best accuracy compared to other methods.

Model	Accuracy
AOG [19]	85.5%
HDM-BG [36]	93.4%
CascadeFormer 1.0	**94.66%**
CascadeFormer 1.1	94.10%
CascadeFormer 1.2	94.10%

Table 4. Comparison on N-UCLA. CascadeFormer 1.1 achieves the best accuracy compared to other methods.

Model	Accuracy
ESV [15]	86.09%
E-TS-LSTM [9]	89.22%
CascadeFormer 1.0	89.66%
CascadeFormer 1.1	**91.16%**
CascadeFormer 1.2	90.73%

Model	NTU60/CV accuracy
ST-LSTM	77.7%
ST-GCN	**88.3%**
CascadeFormer 1.0	88.17%
CascadeFormer 1.1	86.86%
CascadeFormer 1.2	87.24%

Model	NTU60/CS accuracy
ST-LSTM	69.2%
ST-GCN	**81.5%**
CascadeFormer 1.0	81.01%
CascadeFormer 1.1	79.62%
CascadeFormer 1.2	80.48%

Table 5. Comparison on NTU60 (both Cross-Subject and Cross-View). CascadeFormer 1.0 achieves the same performance as ST-GCN without graph structures.

Stage 1: Masked Pretraining

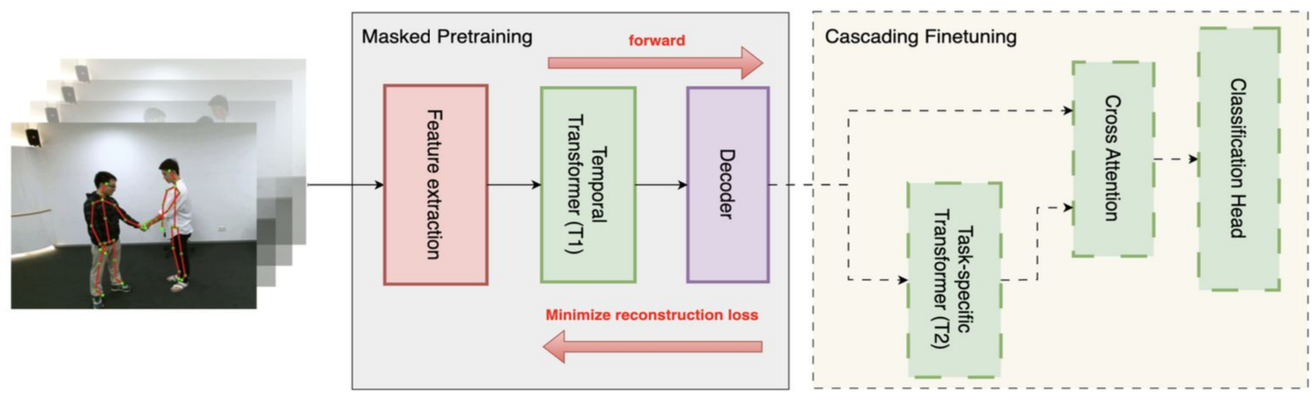

Fig. 1. Overview of the masked pretraining component in CascadeFormer. A fixed percentage of joints are randomly masked across all frames in each video. The partially masked skeleton sequence is passed through a feature extraction module to produce frame-level embeddings, which are then input into a temporal transformer (T1). A lightweight linear decoder is applied to reconstruct the masked joints, and the model is optimized using mean squared error over the masked positions. This stage enables the model to learn generalizable spatiotemporal representations prior to supervised finetuning.

Stage 2: Cascading Finetuning

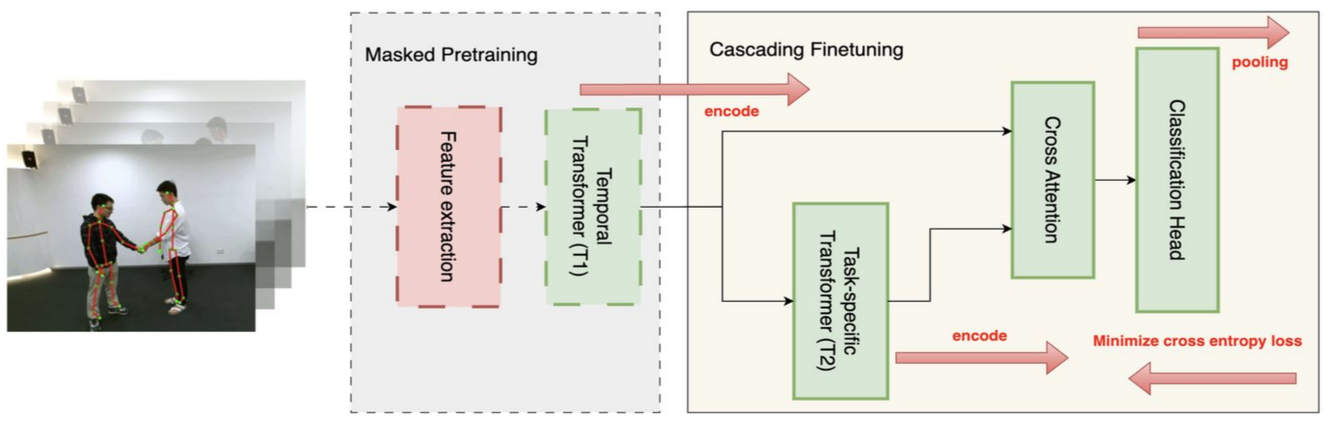

Fig. 2. Overview of the cascading finetuning component in CascadeFormer. The frame embeddings produced by the pretrained temporal transformer backbone (T1) are passed into a task-specific transformer (T2) for hierarchical refinement. The output of T2 is fused with the original embeddings via a cross-attention module. The resulting fused representations are aggregated through frame-level average pooling and passed to a lightweight classification head. The entire model, including T1, T2, and the classification head, is optimized using cross-entropy loss on labels during finetuning.

Multi-person Actions results

Action Type	NTU60/CS accuracy	NTU60/CV accuracy
Single-person	80.82%	**88.92%**
Two-persons	**81.81%**	84.86%
Overall	81.01%	88.17%

Table 2. Performance of CascadeFormer 1.0 on multi-person actions in NTU RGB+D 60. Accuracy is broken down into single-person actions and two-person interactions under both cross-subject and cross-view splits.

Ablation Highlights

# Epochs of Pretraining	# Epochs of Finetuning	NTU60/CS accuracy
1 epoch	100 epochs	76.38%
50 epoch	100 epochs	80.06%
100 epochs	100 epochs	**81.01%**
200 epochs	100 epochs	81.09%

Table 6. Effect of pretraining duration on NTU RGB+D 60 (cross-subject) performance. We report accuracy on the cross-subject split after pretraining **CascadeFormer 1.0** for 1, 50, 100, and 200 epochs. Performance improves significantly with longer pretraining, but plateaus after 100 epochs, indicating diminishing returns.

Backbone Freezing Decision	Accuracy
fully finetune	**94.66%**
fully freeze	85.11%
finetune the last layer	88.39%

Table 10. Effect of different backbone freezing strategies during cascading finetuning on Penn Action. Fully finetuning the transformer backbone yields the highest accuracy, while freezing all layers significantly degrades performance.

References

[1] Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition (2018), https://arxiv.org/abs/1801.07455.
[2] Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations (2023), https://arxiv.org/abs/2210.06551