Modeling Multi-Label Action Dependencies for Temporal Action Localization

Praveen Tirupattur, Kevin Duarte, Yogesh Rawat, Mubarak Shah

Center for Research in Computer Vision (CRCV)

University of Central Florida

CVPR VIRTUAL JUNE 19-25

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Problem

- Temporal Action Localization
  - Inputs → Untrimmed Videos
  - Task → Find action boundaries

- Real World Video
  - Multiple complex actions
  - Inherent relation between action classes

Proposed Approach

- Motivation - Model relationship between action classes to improve localization

- Relationship between actions
  - Co-occurrence (Overlapping activities)
  - Temporal Ordering

- Examples
  - Co-occurrence - Run and Pole Vault

Proposed Approach

- Motivation - Model relationship between action classes to improve localization

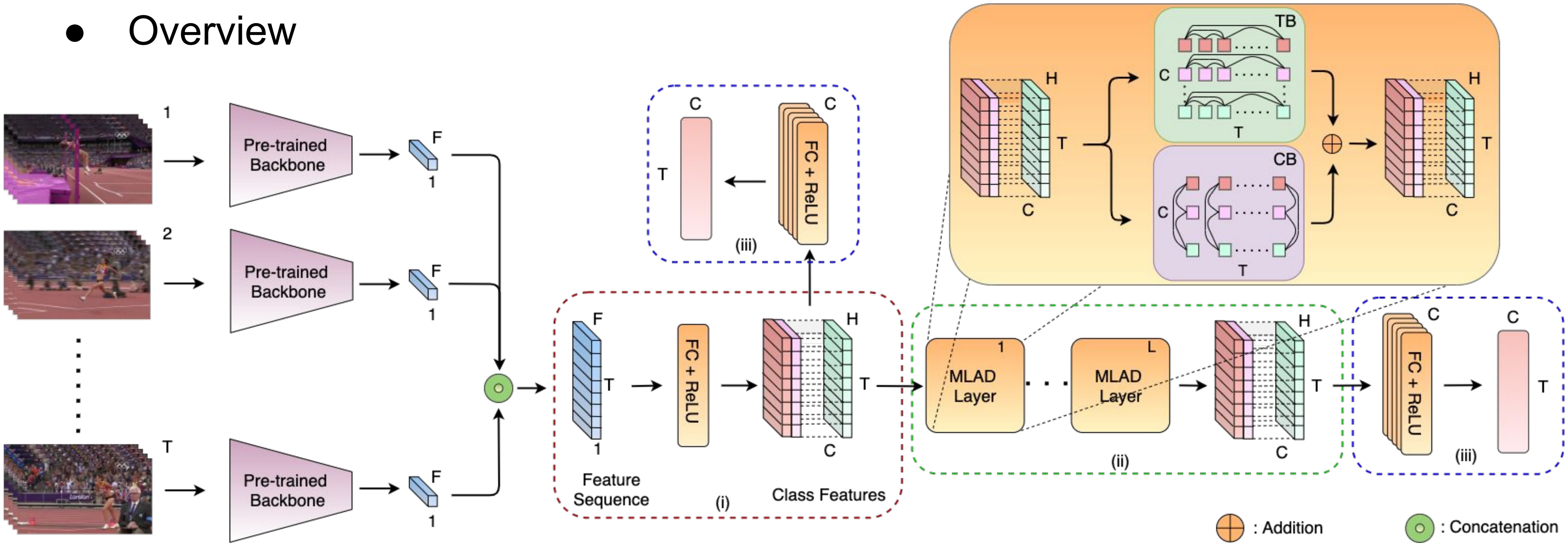
- Relationship between actions
    - Co-occurrence (Overlapping activities)
    - Temporal Ordering

- Examples
    - Co-occurrence - Run and Pole Vault
    - Temporal Ordering - Dribble precedes Dunk

Related Works:
1. Differentiable grammars for videos, AAAI 2020
2. Inferring temporal compositions of actions using probabilistic automata, CVPR 2020

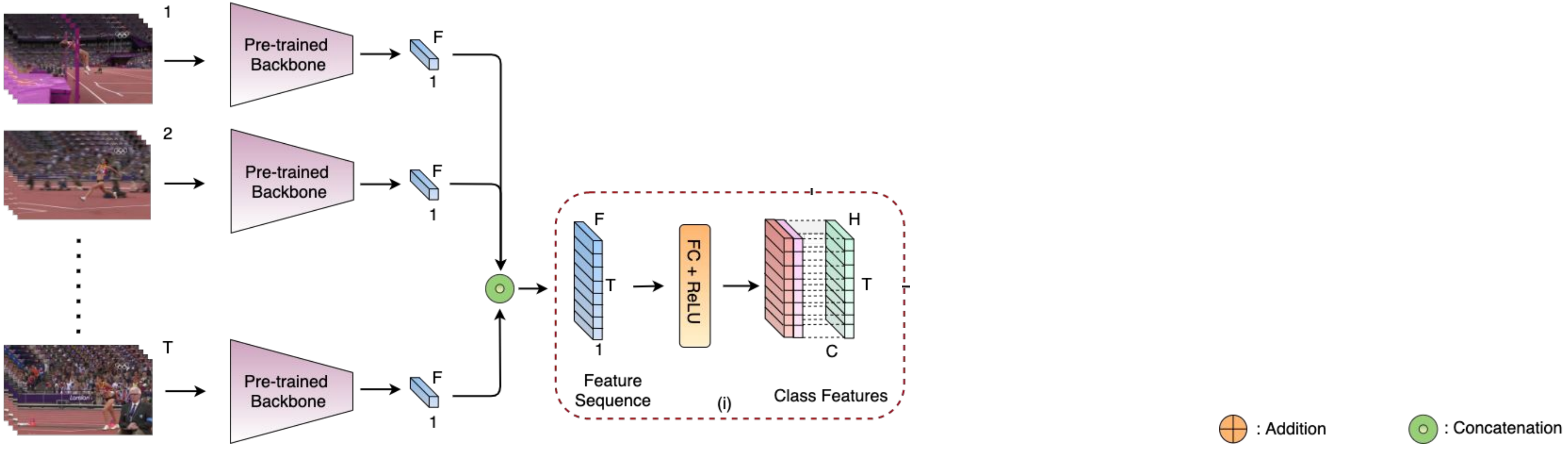
Architecture

- Overview



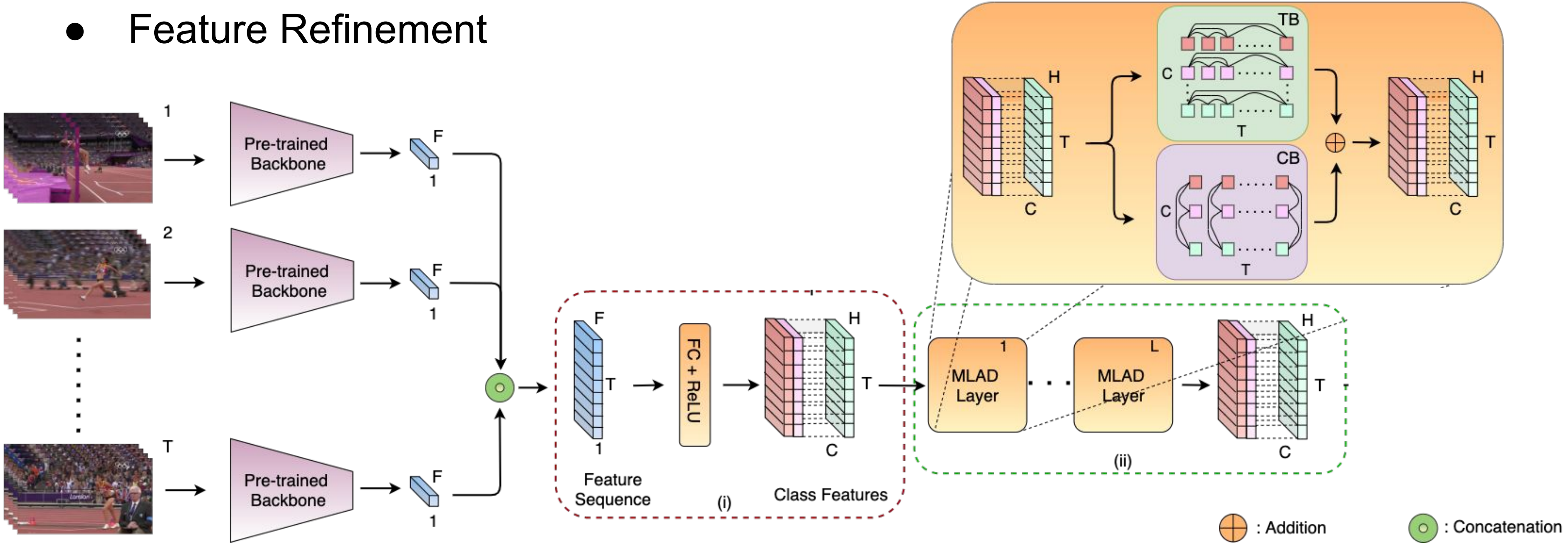
Architecture

- Feature Extraction



Architecture

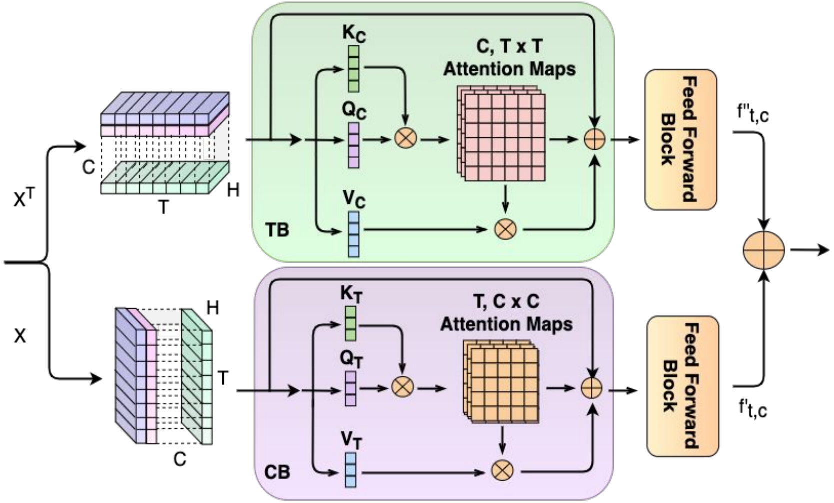
- Feature Refinement



MLAD Layer

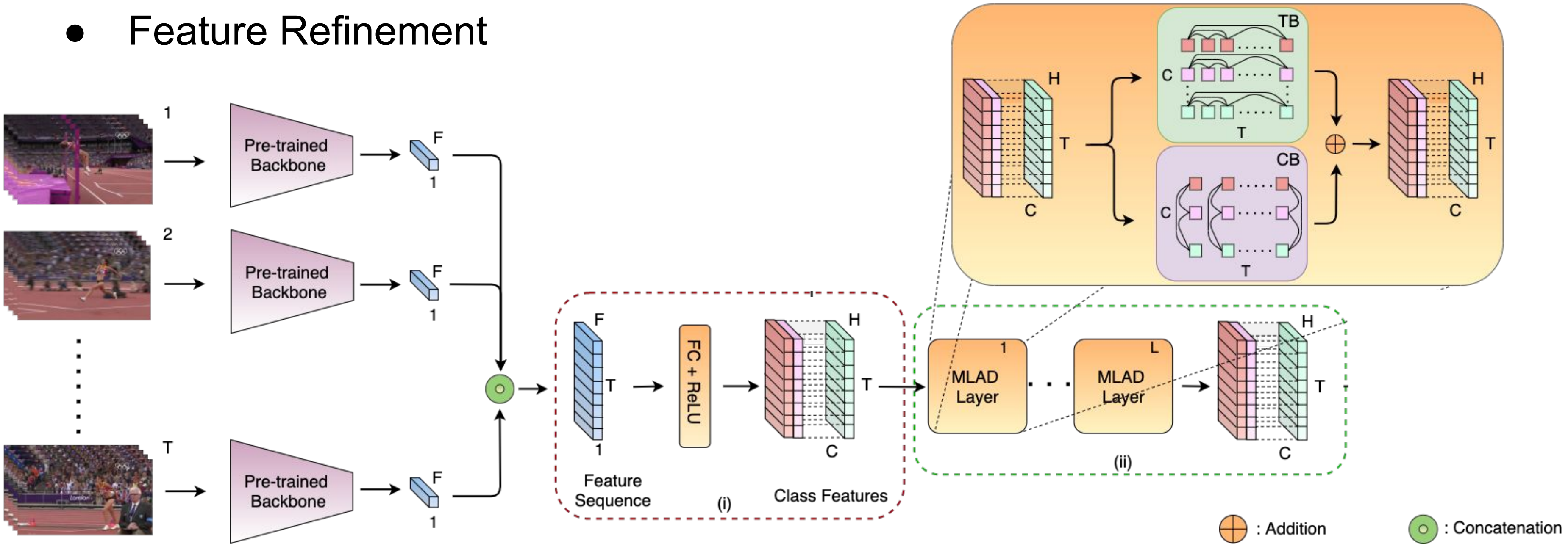
- Co-occurrence Dependency
  - For each timestep T, learn class-wise relation
  - T attention maps of shape C x C
- Temporal Dependency
  - For each class C, learn relation across time
  - T attention maps of shape C x C
- Weighted average of learned features

$$g_{t,c} = \alpha f'_{t,c} + (1 - \alpha) f''_{t,c}.$$



Architecture

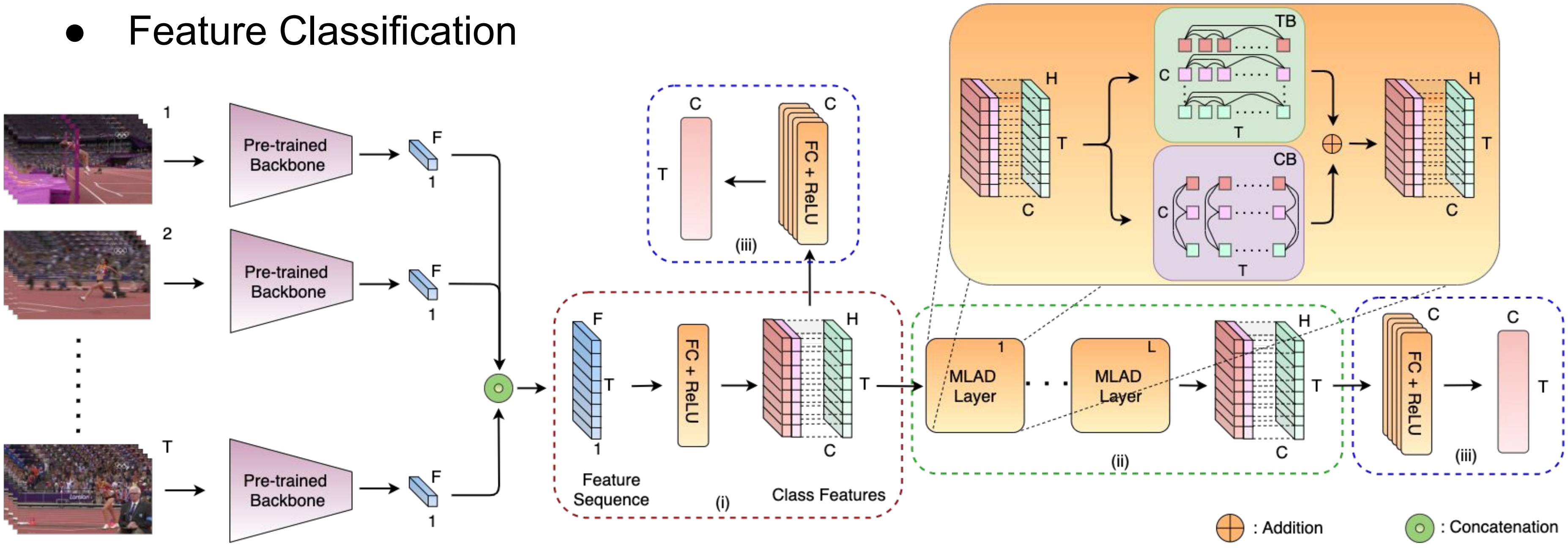
- Feature Refinement



CENTER FOR RESEARCH IN COMPUTER VISION

Architecture

- Feature Classification



Multi-label Metrics

- Existing Multi-label metrics
    - Hamming Loss (HL)
    - Zero-One Loss (ZL)
    - Ranking Loss (RL)
    - Coverage- Loss (CE)
    - Jaccard Score (JS)
    - Label Ranking Average Precision (LRAP)

- Existing evaluation metric treat each timestep as an individual sample.
- Each class within a timestep is evaluated independently.

Proposed Metric

VIRTUAL JUNE 19-25

$$Precision(c) = \frac{N_{\text{correct}}(c)}{N_{\text{predict}}(c)} \qquad Recall(c) = \frac{N_{\text{correct}}(c)}{N_{\text{gt}}(c)}$$

$$Precision(c_i|c_j) = \frac{N_{\text{correct}}(c_i|c_j)}{N_{\text{predict}}(c_i|c_j)} \qquad Recall(c_i|c_j) = \frac{N_{\text{correct}}(c_i|c_j)}{N_{\text{gt}}(c_i|c_j)}$$

$$Precision(c_i|c_j,\tau) = \frac{N_{\text{correct}}(c_i|c_j,\tau)}{N_{\text{predict}}(c_i|c_j,\tau)} \qquad Recall(c_i|c_j,\tau) = \frac{N_{\text{correct}}(c_i|c_j,\tau)}{N_{\text{gt}}(c_i|c_j,\tau)}$$

CENTER FOR RESEARCH IN COMPUTER VISION

Proposed Metric

- Results on MultiTHUMOS

Action-Conditional Metrics ↑

	$\tau = 0$				$\tau = 20$			
	$\mathbf{P}_{AC}$	$\mathbf{R}_{AC}$	$\mathbf{F1}_{AC}$	$\mathbf{mAP}_{AC}$	$\mathbf{P}_{AC}$	$\mathbf{R}_{AC}$	$\mathbf{F1}_{AC}$	$\mathbf{mAP}_{AC}$
I3D	33.63	15.23	18.65	32.58	37.88	18.01	21.96	35.53
CF	36.73	21.39	23.71	35.00	41.95	23.91	27.22	38.42
TGM [33]	34.59	17.21	20.14	36.90	39.27	20.13	23.86	40.18
Our	**39.22**	**28.33**	**29.37**	**40.15**	**42.89**	**30.27**	**32.18**	**43.76**

Experiments

- Qualitative Results - fmAP Scores

Method	MultiTHUMOS	Charades
I3D Baseline* [33]	29.7	17.2
CF Baseline	42.6	14.8
Super-events* [34]	36.4	19.4
TGMs* [33]	44.3	21.5
TGMs + SE* [33]	46.4	22.3
TGMs + DG* [32]	48.2	22.9
Our Approach	**51.5**	**23.7**

Ablations

	MultiTHUMOS	Charades
L = 1	48.55	20.48
L = 3	50.30	23.15
L = 5	51.52	23.74

Features	**MultiTHUMOS**	**Charades**
RGB	42.24	18.40
Flow	48.77	20.10
Late Fusion	49.58	22.93
Early Fusion	51.52	23.74

Eval. Length	**Fixed Tr. Length**	**Var. Tr. Length**
T = 32	49.90	50.20
T = 64	51.14	51.01
T = 96	51.31	51.31
T = 128	50.59	51.52

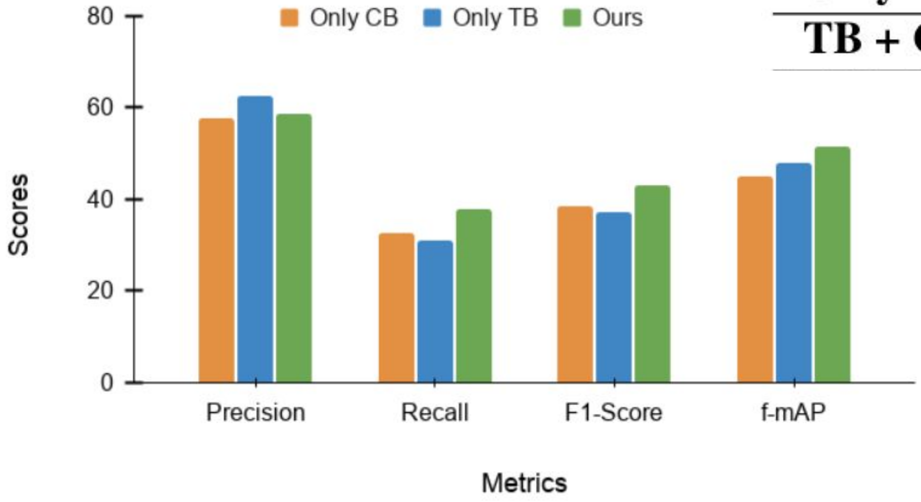
	W/O Initial Loss	**With Initial Loss**
f-mAP	49.96	51.52

	Fixed Alpha	**Learned Alpha**
f-mAP	50.95	51.52

Analysis

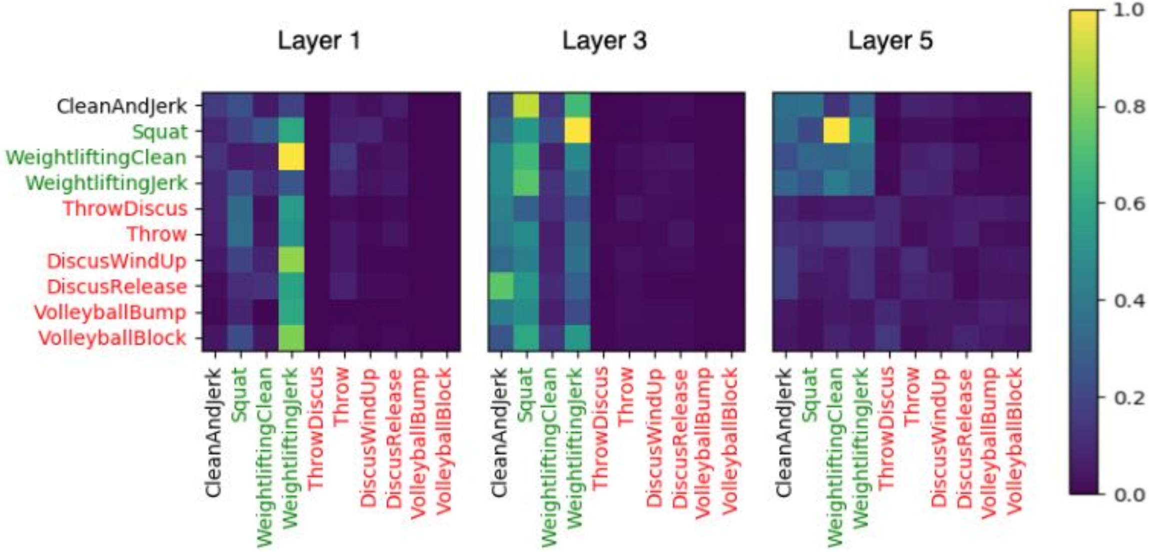
- Effect of CB and TB

	MultiTHUMOS	Charades
No CB, No TB	42.60	14.80
Only CB	44.98	20.3
Only TB	48.03	21.1
TB + CB	51.52	23.5



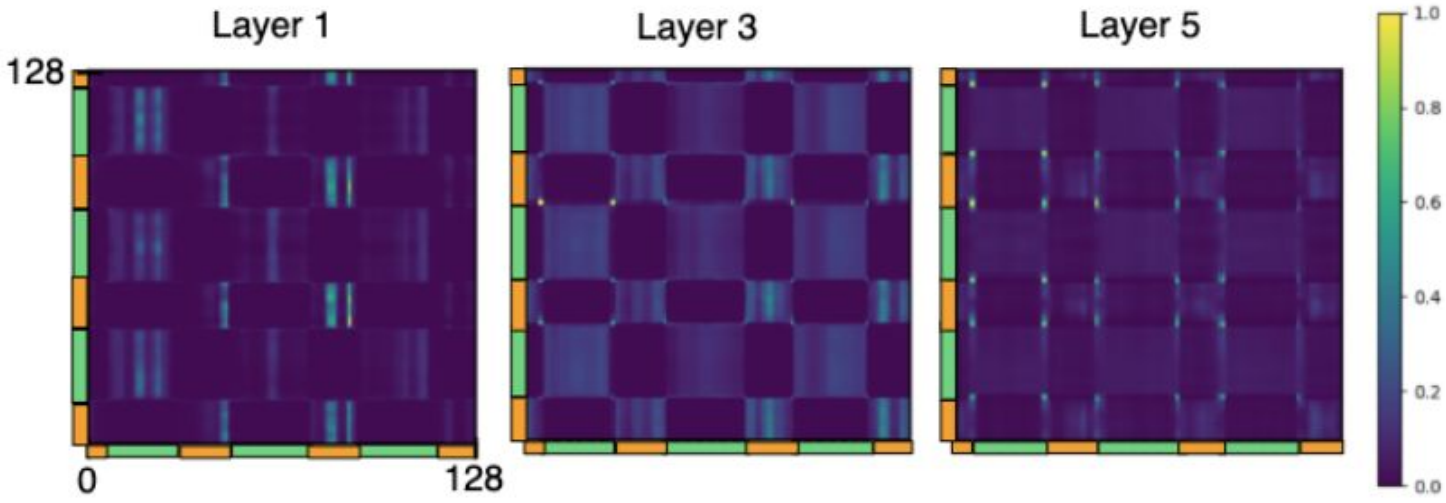
Interpretability of MLAD Layer

- Visualization from CB

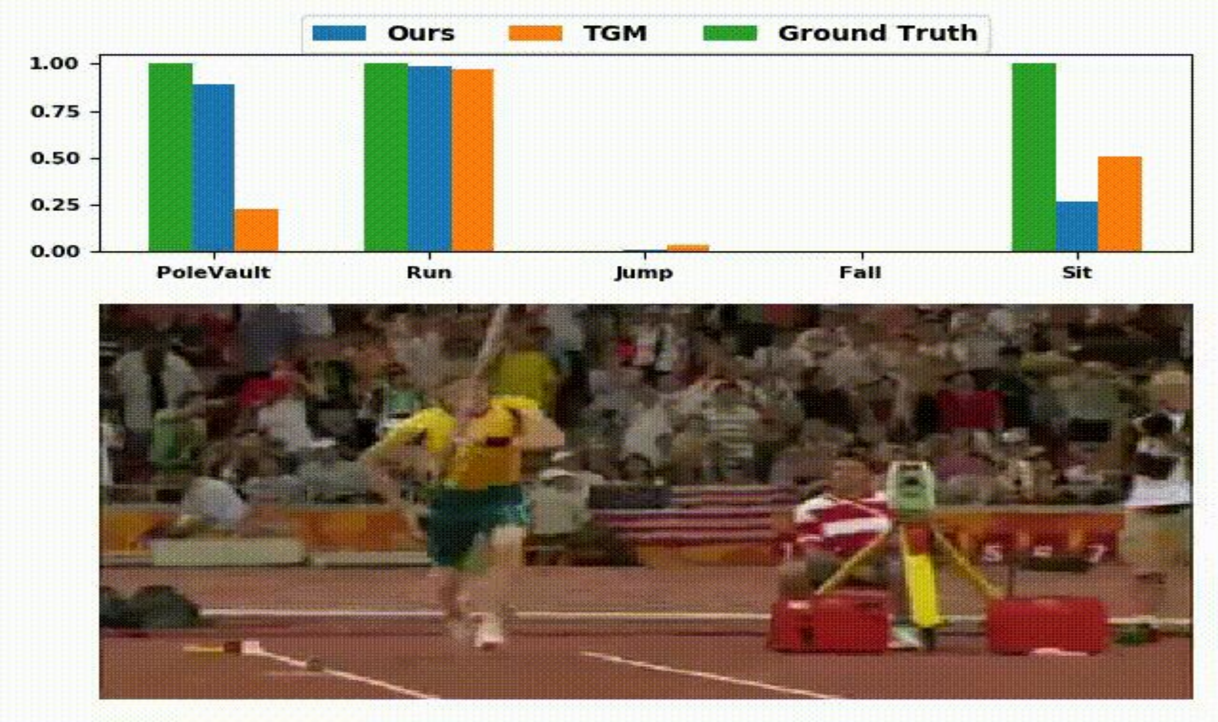


Interpretability of MLAD Layer

- Visualization from TB



Qualitative Results



CVPR VIRTUAL JUNE 19-25

Modeling Multi-Label Action Dependencies for Temporal Action Localization

Praveen Tirupattur, Kevin Duarte, Yogesh Rawat, Mubarak Shah

Center for Research in Computer Vision (CRCV)

University of Central Florida

Code available at
https://github.com/ptirupat/MLAD

UCF CENTER FOR RESEARCH IN COMPUTER VISION