Microsoft

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

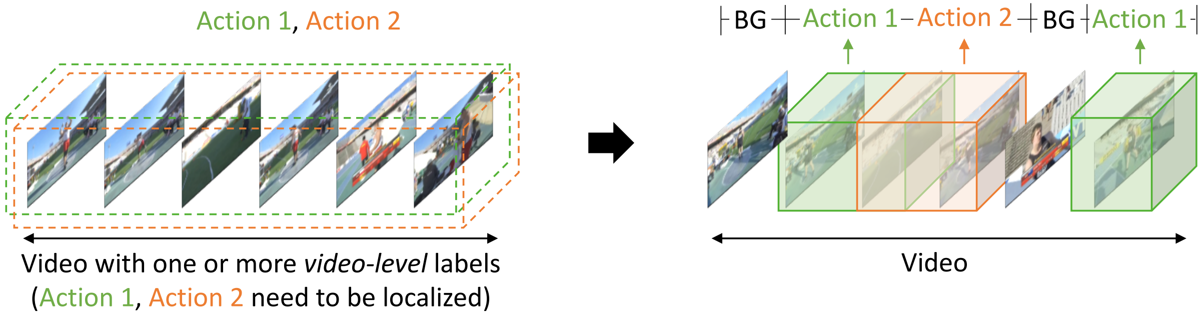
UCF CENTER FOR RESEARCH IN COMPUTER VISION

PivoTAL: Prior-Driven Supervision for Weakly-Supervised Temporal Action Localization

Mamshad Nayeen Rizve[*2], **Gaurav Mittal**[*1], Ye Yu[1], Matthew Hall[1], Sandra Sajeev[1], Mubarak Shah[2], Mei Chen[1]

[1]Microsoft [2]University of Central Florida

THU-PM-228



Action 1, Action 2

├ BG ┼ Action 1 ─ Action 2 ┼ BG ┤Action 1┤

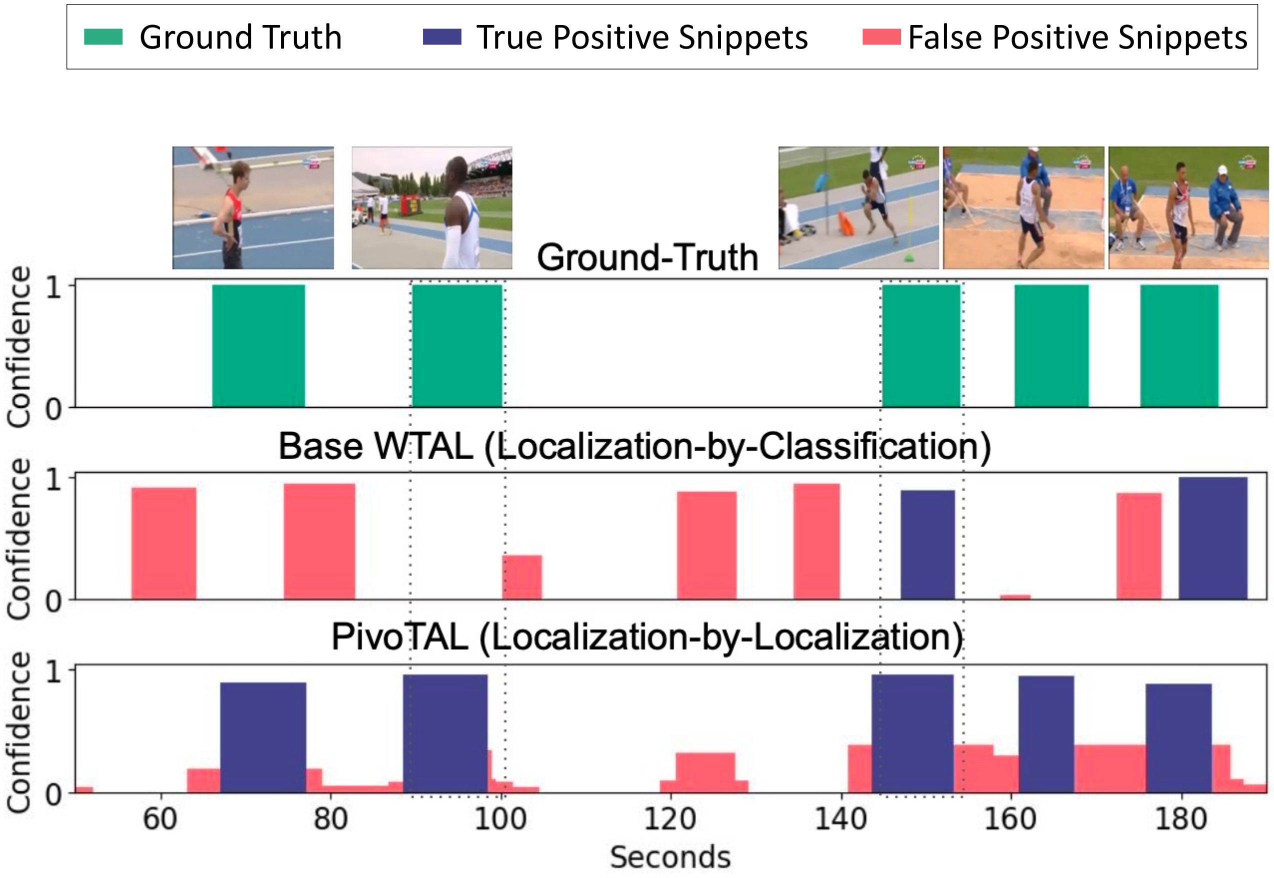
Video with one or more *video-level* labels
(Action 1, Action 2 need to be localized)

Video

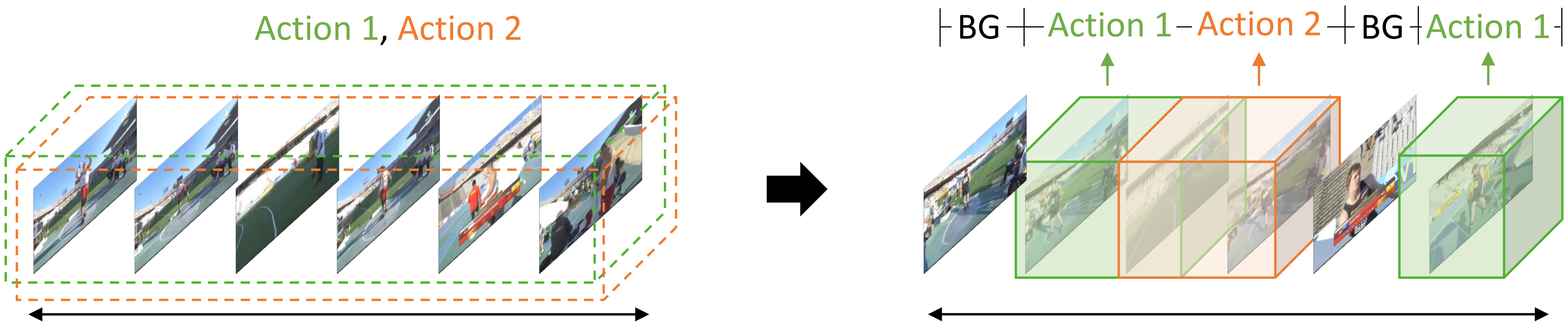
Weakly-supervised Temporal Action Localization (WTAL):
Localize actions using only video-level labels

Introducing **PivoTAL:**

- Generates pseudo-action snippets to localize actions directly.
- Unlike existing methods that classify each video frame and perform post-training aggregation into action snippets.



Weakly-supervised Temporal Action Localization (WTAL)

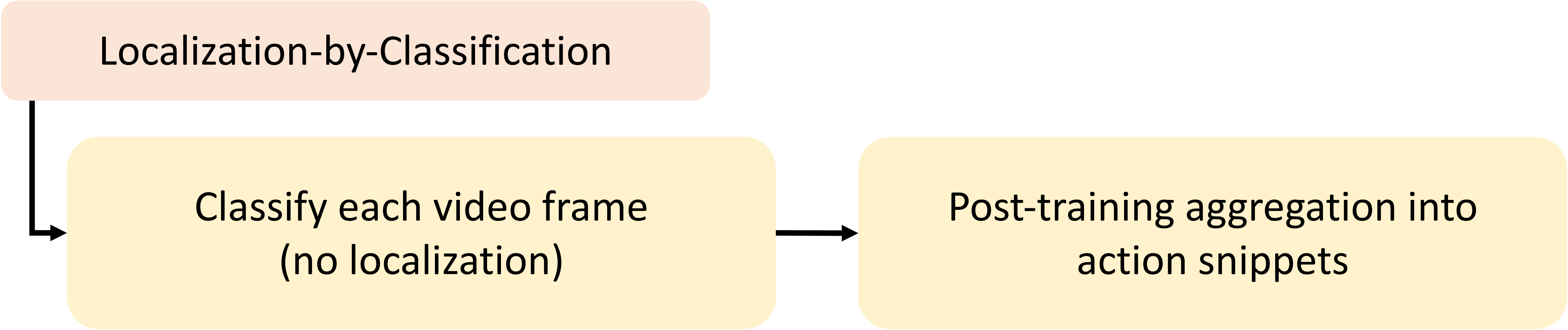


Video with one or more *video-level* labels
(Action 1, Action 2 need to be localized)

Video

- Localize actions using only *video-level* action labels.
- Challenging as **NO** dense, frame-level labels available for start/end of actions.
- Mitigates expensive and labor-intensive dense labels compared to supervised TAL.

Limitations of Existing Methods



Localization-by-Classification

Classify each video frame (no localization)

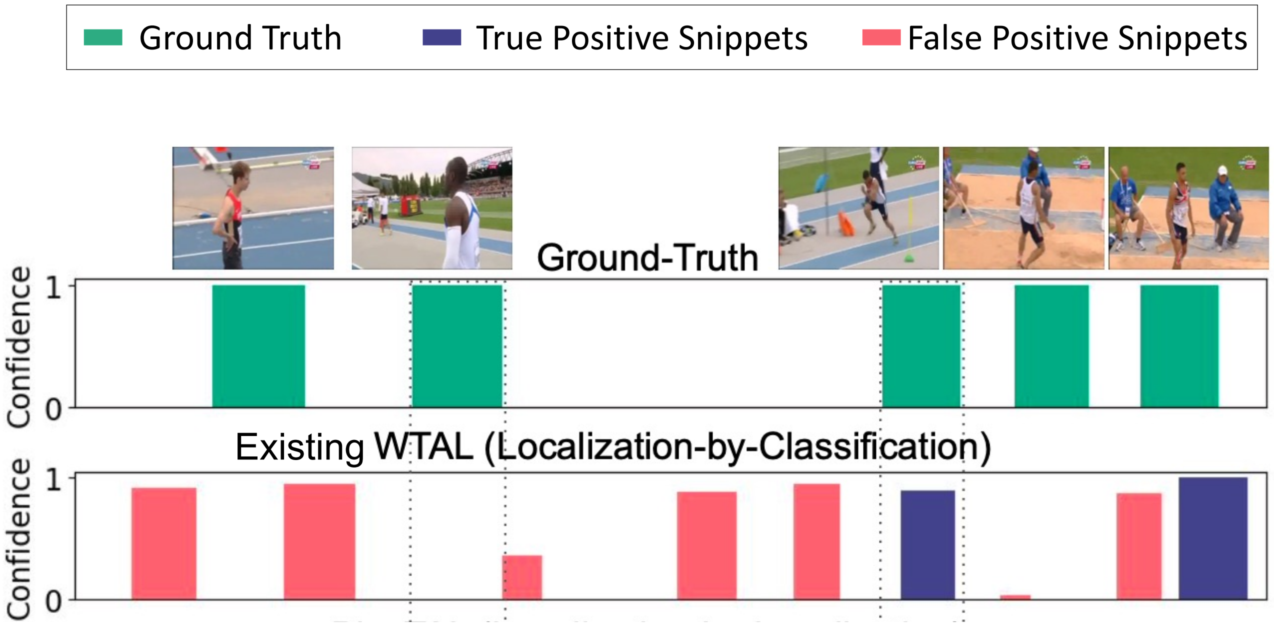
Post-training aggregation into action snippets

⚠ No explicit notion of temporal boundaries

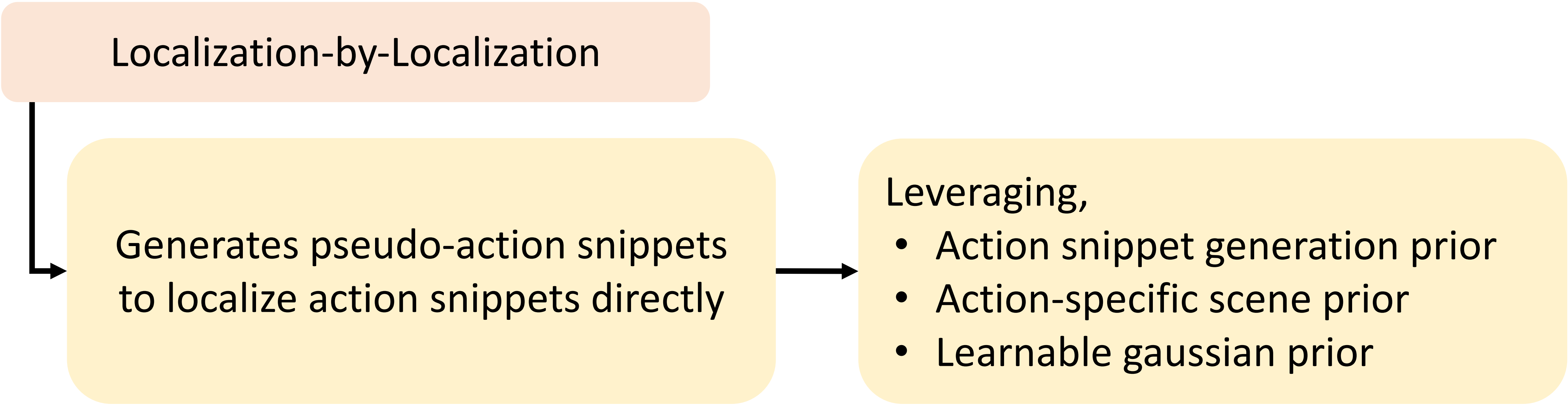
⚠ Post-training aggregation cannot influence training

⚠ Leads to incomplete action snippets and several high-confident false-positives

Long Jump


🟩 Ground Truth
🟦 True Positive Snippets
🟥 False Positive Snippets
Ground-Truth
Confidence
1
0
Existing WTAL (Localization-by-Classification)
Confidence
1
0

PivoTAL (Prior-driven Supervision for WTAL)



Localization-by-Localization

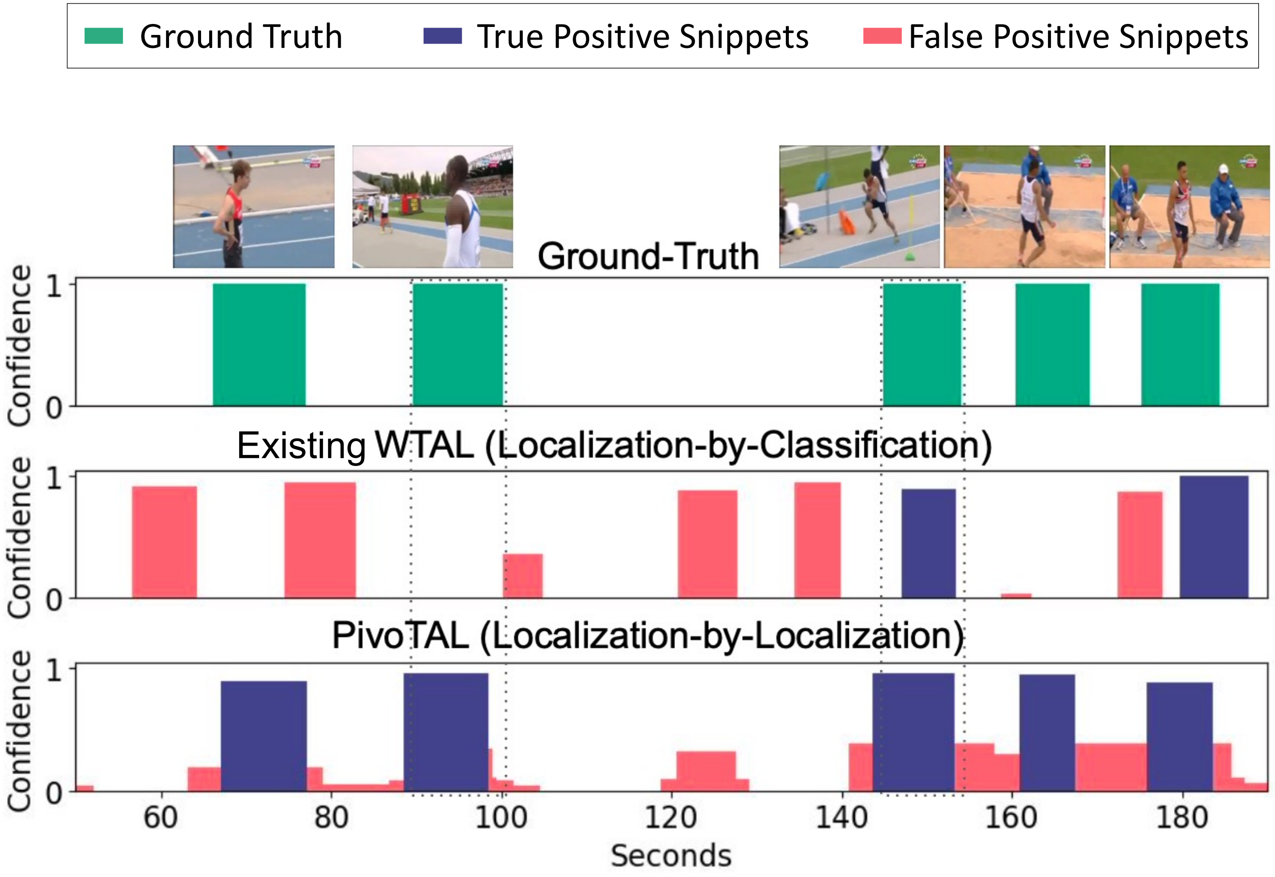
Generates pseudo-action snippets to localize action snippets directly

Leveraging,

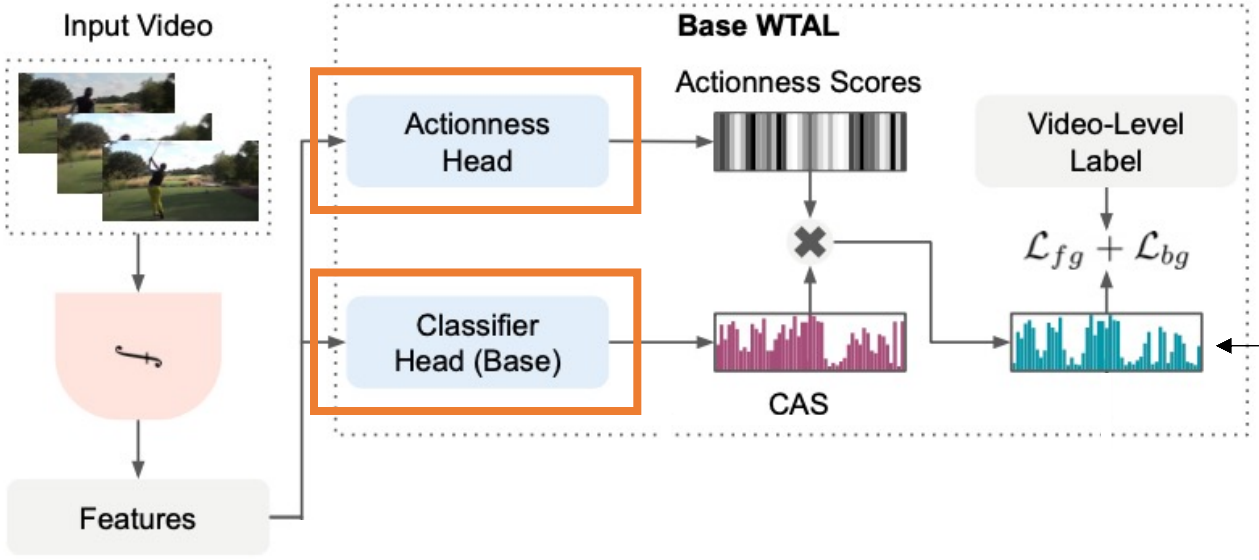
- Action snippet generation prior
- Action-specific scene prior
- Learnable gaussian prior

👍 High-confident true positives, more complete action snippets, low-confident false positives

Long Jump

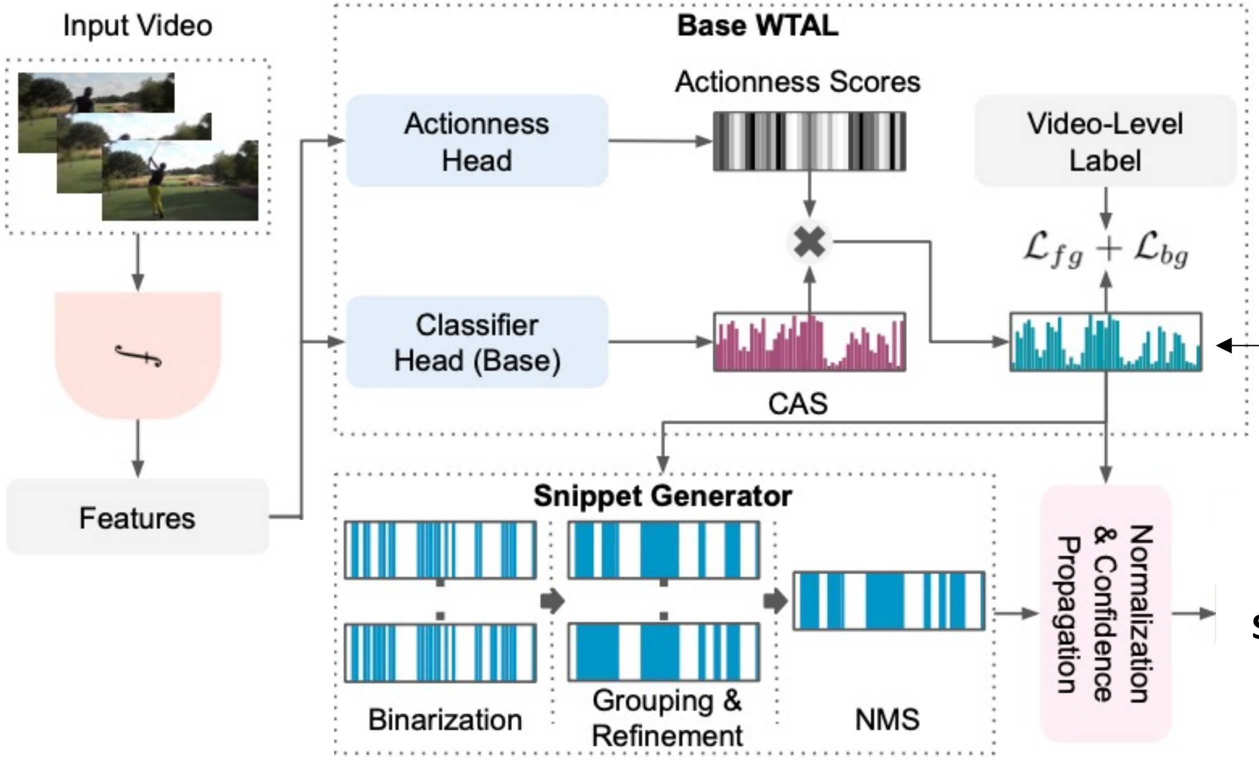

Ground Truth
True Positive Snippets
False Positive Snippets
Ground-Truth
Confidence
1
0
Existing WTAL (Localization-by-Classification)
Confidence
1
0
PivoTAL (Localization-by-Localization)
Confidence
1
0
60
80
100
120
140
160
180
Seconds

Base WTAL



Action prediction per frame
(no explicit localization)

Base WTAL

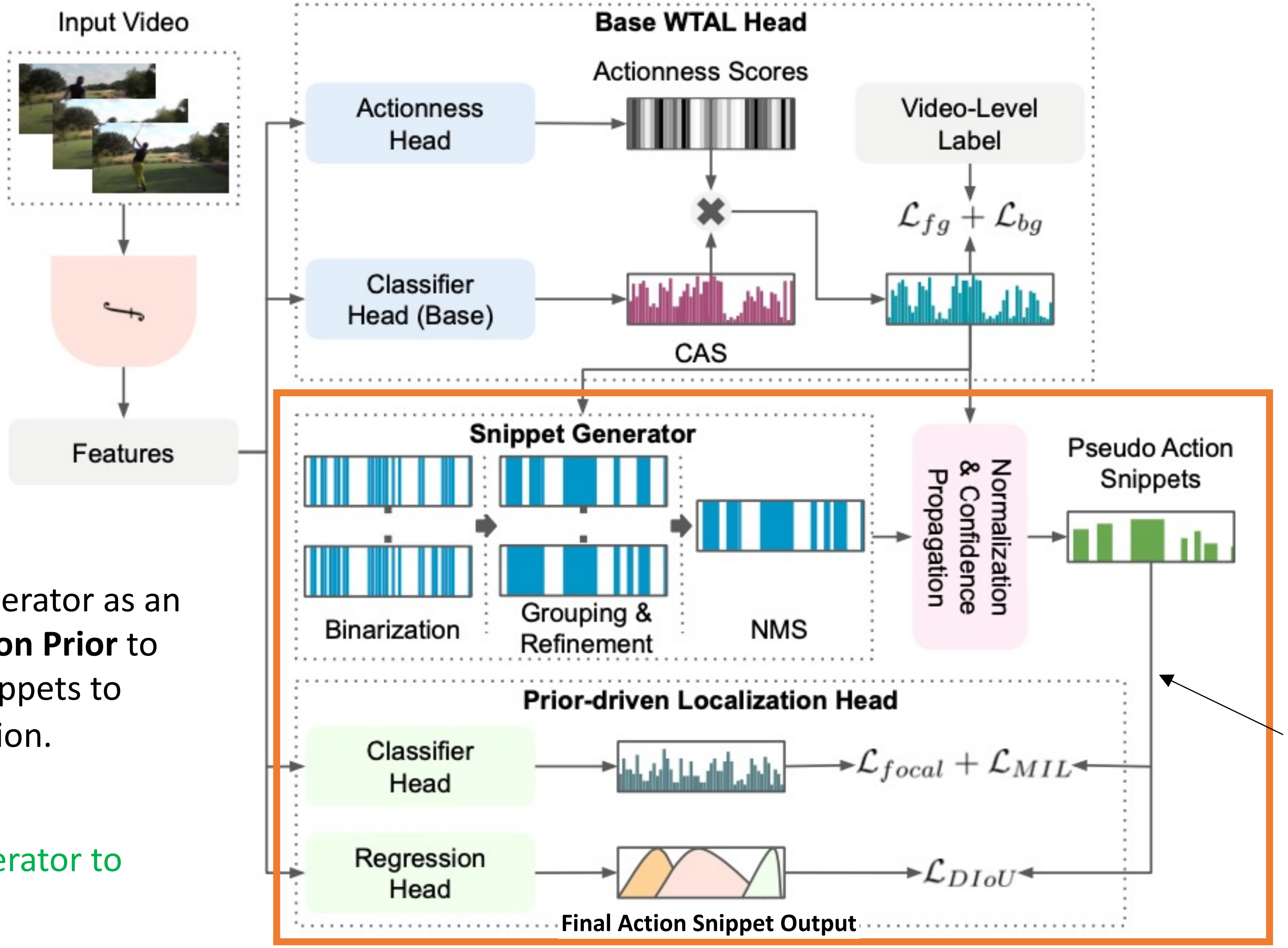


Action prediction per frame (no explicit localization)

Final Action Snippet Output

Post-training Aggregation (no influence on model training)

PivoTAL: Action Snippet Generation Prior



Providing supervision for direct localization

PivoTAL uses snippet generator as an **Action Snippet Generation Prior** to obtain pseudo-action snippets to supervise direct localization.

It also helps snippet generator to influence training.

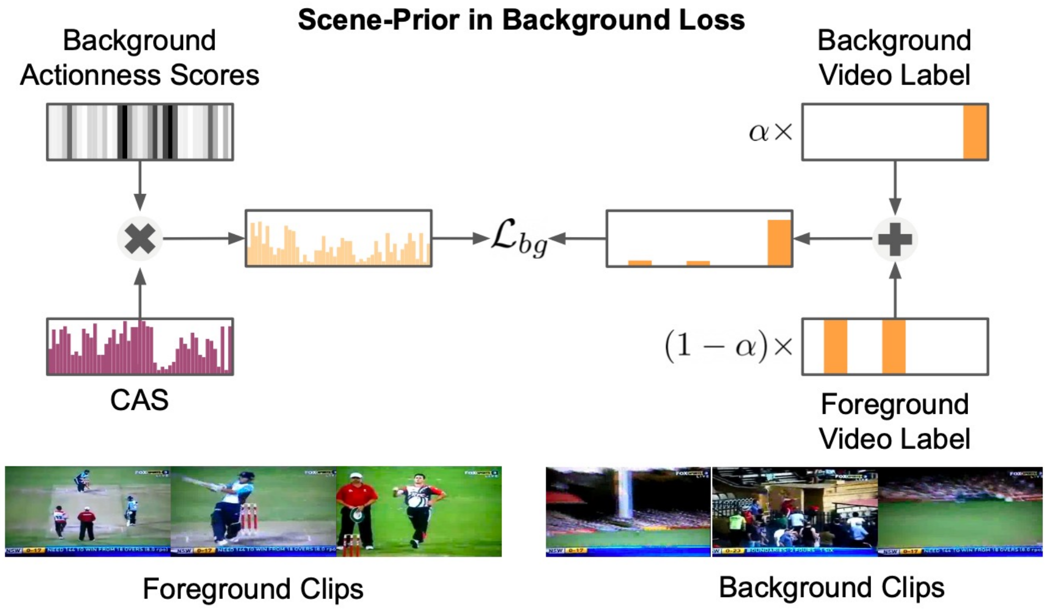
PivoTAL: Action-specific Scene Prior

- Spatial information overlaps between action (foreground) and no-action (background) video frames.

- So, background frames can be related to neighboring foreground frames.

- **Action-specific Scene Prior:** Loss modified in Base WTAL to assign background frames a label which is a linear combination of background class and neighboring foreground action class.

- Improves quality of pseudo-action snippets, thus improves localization

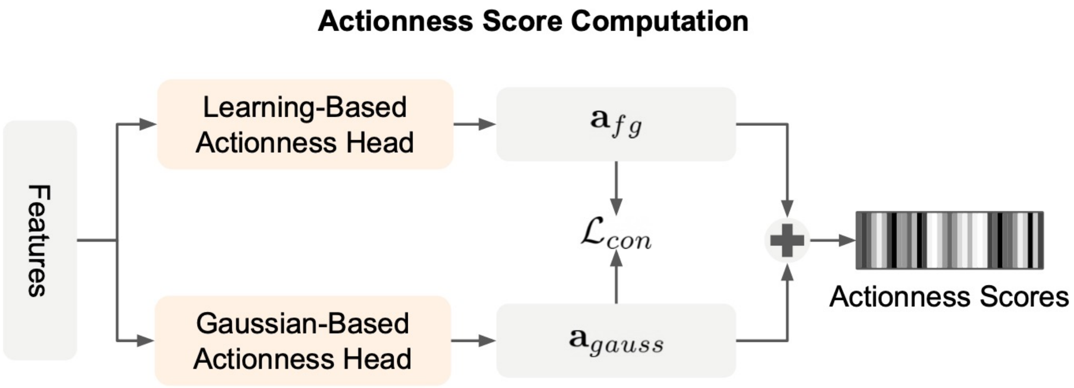


PivoTAL: Gaussian Prior

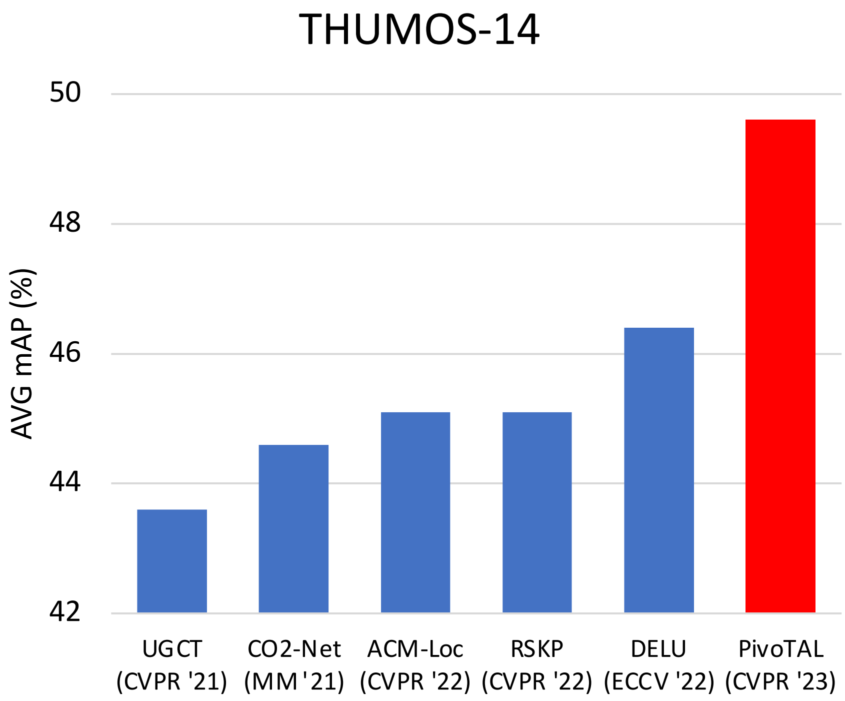
- Per-frame Base WTAL predictions need to be locally consistent to improve quality of pseudo-action snippets.

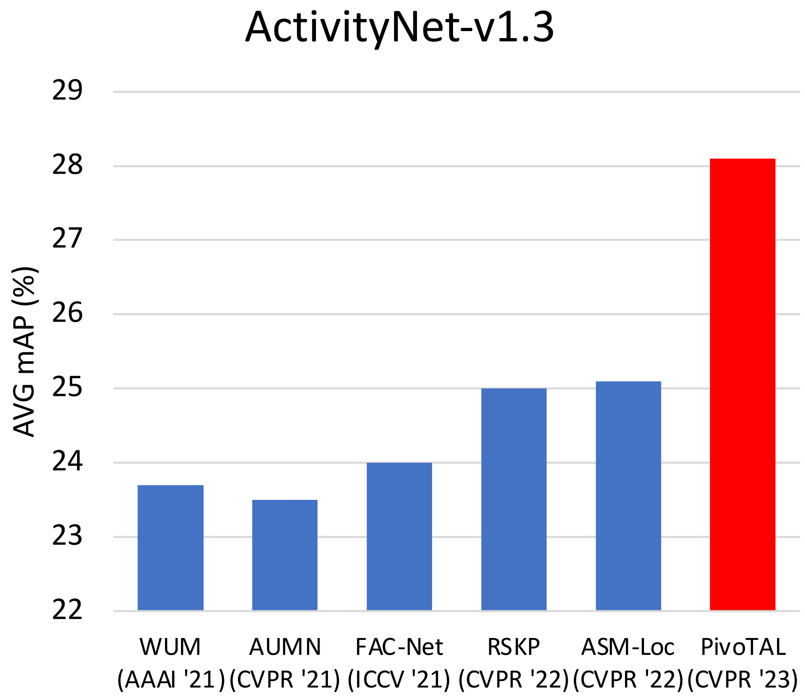
- **Gaussian Prior:** Actionness scores augmented with a Gaussian Mask to incorporate local video context.

- Improves quality of pseudo-action snippets, thus improves localization

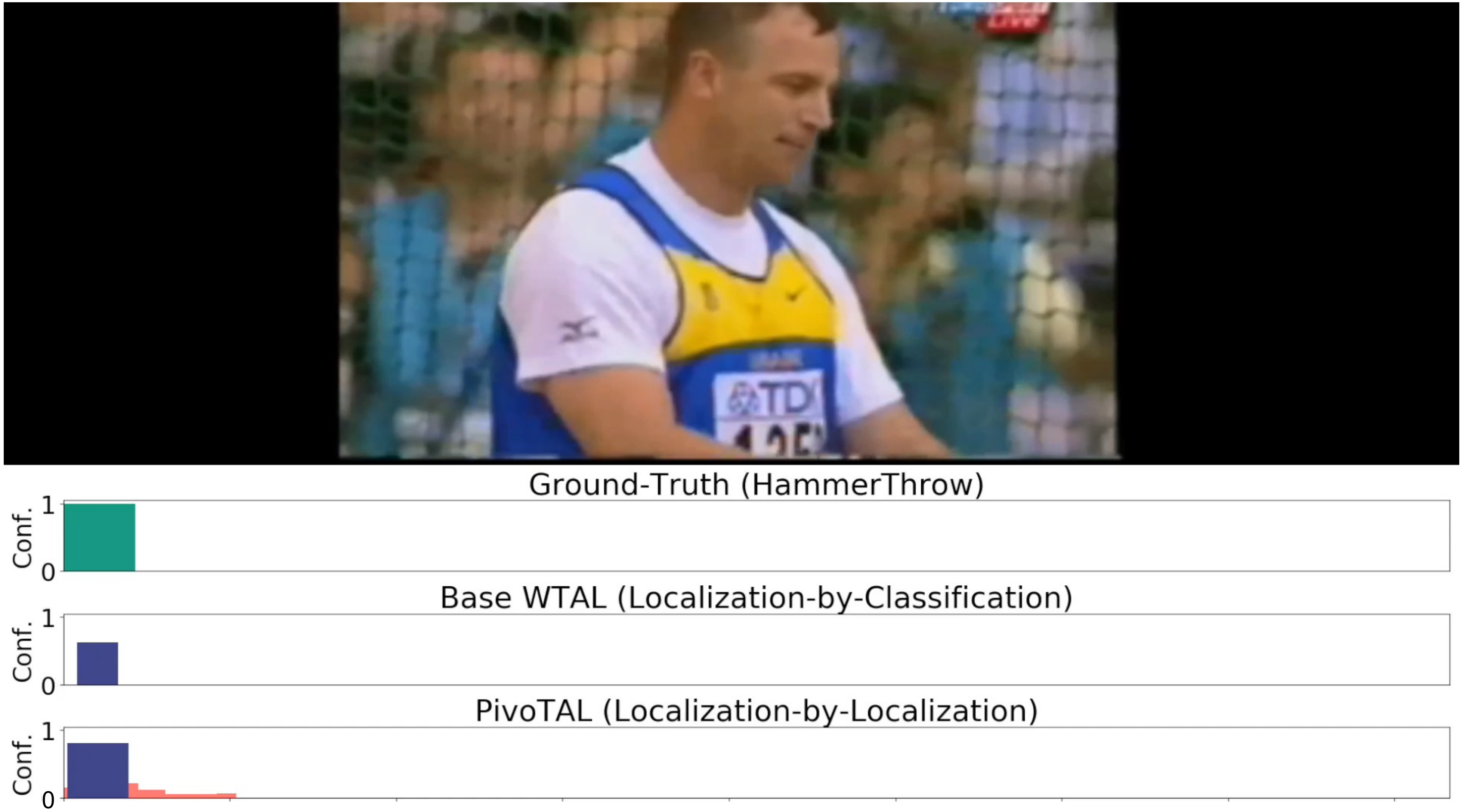

Actionness Score Computation
Features
Learning-Based Actionness Head
Gaussian-Based Actionness Head
$\mathbf{a}_{fg}$
$\mathcal{L}_{con}$
$\mathbf{a}_{gauss}$
Actionness Scores

PivoTAL: Significantly outperforms all existing methods on public benchmarks





PivoTAL: Localization Predictions



Conclusion

- PivoTAL: Localization-by-localization method for WTAL that generates pseudo-action snippets to localize actions directly.
- Introduces and leverages action-snippet generation prior, action-specific scene prior, and learnable gaussian prior.
- Achieves at least 3% higher average mAP than any existing WTAL method on THUMOS-14 and ActivityNet-v1.3.
- **Hope to see you at our poster THU-PM-228!** 😊