ICML
International Conference
On Machine Learning

Action Dubber: Timing Audible Actions via Inflectional Flow

Wenlong Wan[1,2] Weiying Zheng[3] Tianyi Xiang[4,2,1] Guiqing Li[1] Shengfeng He[2]
[1]South China University of Technology
[2]Singapore Management University
[3]University of Hong Kong
[4]City University of Hong Kong

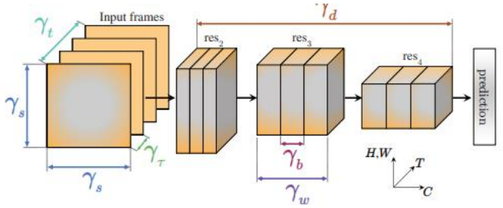
Video without audio

Problem Definition

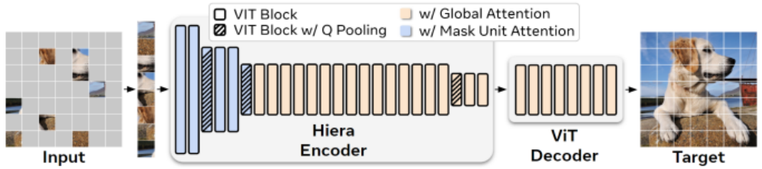
Video without audio

Problem Definition

Video Action Recognition

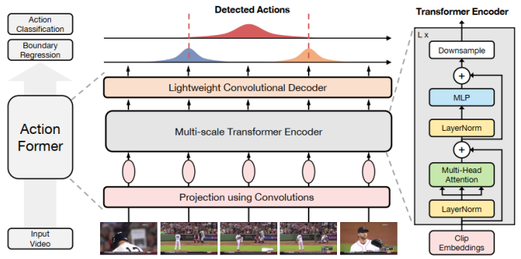


X3d[1] (CVPR 20’)

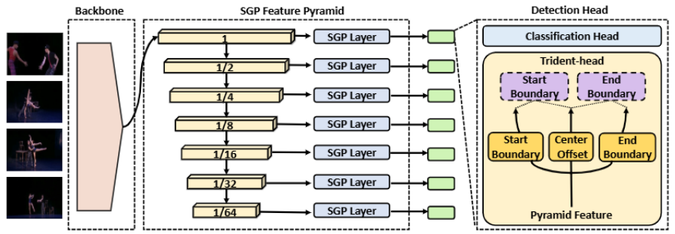


Hiera[2]

Temporal Action Localization



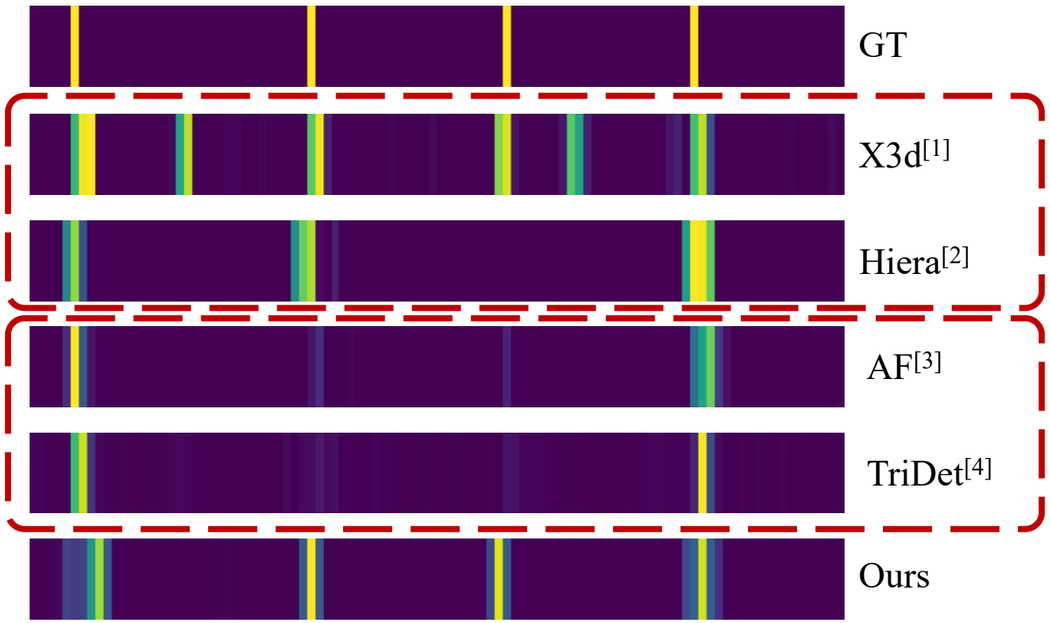
ActionFormer[3] (ECCV 22’)



TriDet[4] (CVPR 23’)

Silent Video with Audible Actions

Temporal Audio Position

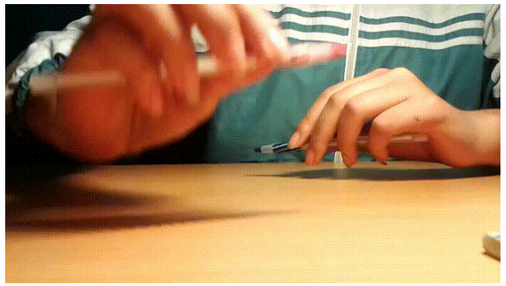
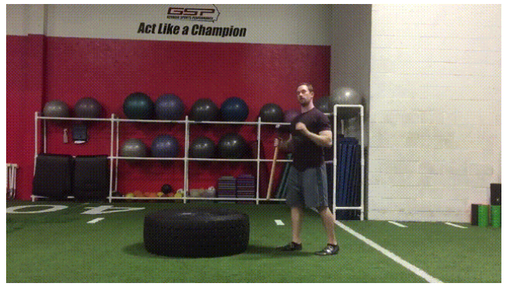


Motivation

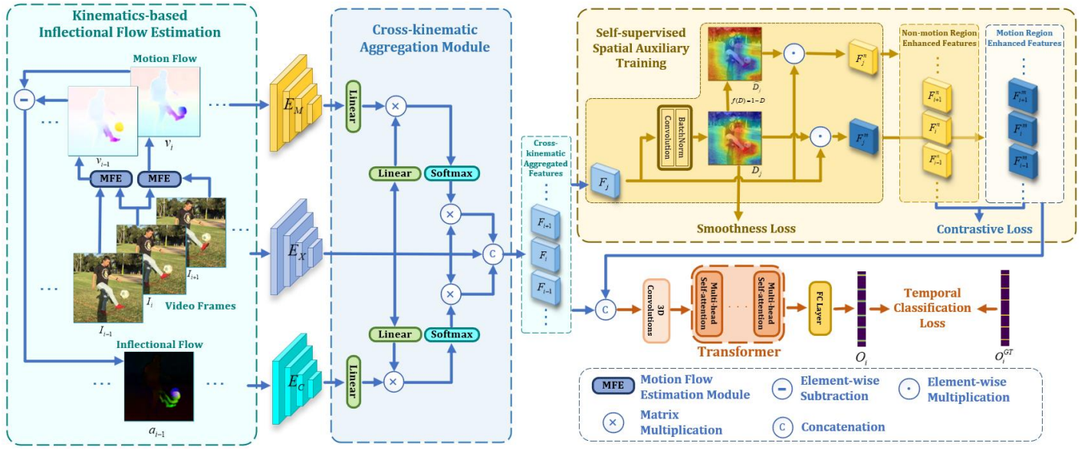
[1] Feichtenhofer, C. (2020). X3d: Expanding architectures for efficient video recognition. In *CVPR*. (pp. 203-213).
[2] Ryali, C., Hu, Y. T., Bolya, D., Wei, C., Fan, H., Huang, P. Y., ... & Feichtenhofer, C. (2023). Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles. arXiv preprint arXiv:2306.00989.
[3] Zhang, C. L., Wu, J., & Li, Y. (2022, October). Actionformer: Localizing moments of actions with transformers. In *ECCV*. (pp. 492-510).
[4] Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., & Tao, D. (2023). TriDet: Temporal action detection with relative boundary modeling. In *CVPR*. (pp. 18857-18866).

*Audible*623 Dataset



TA^2Net



Contributions

Basic Assumption

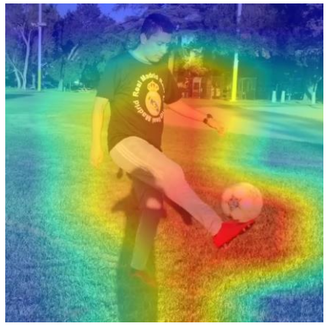
Audible actions are induced by sudden changes in the forces acting upon an object ( ).

Forces acted on the ball:
Gravity only

Forces acted on the ball:
Gravity + Sudden normal force
(Key frame of action)

Forces acted on the ball:
Gravity only

*Air resistance force is too small so it can be ignored.

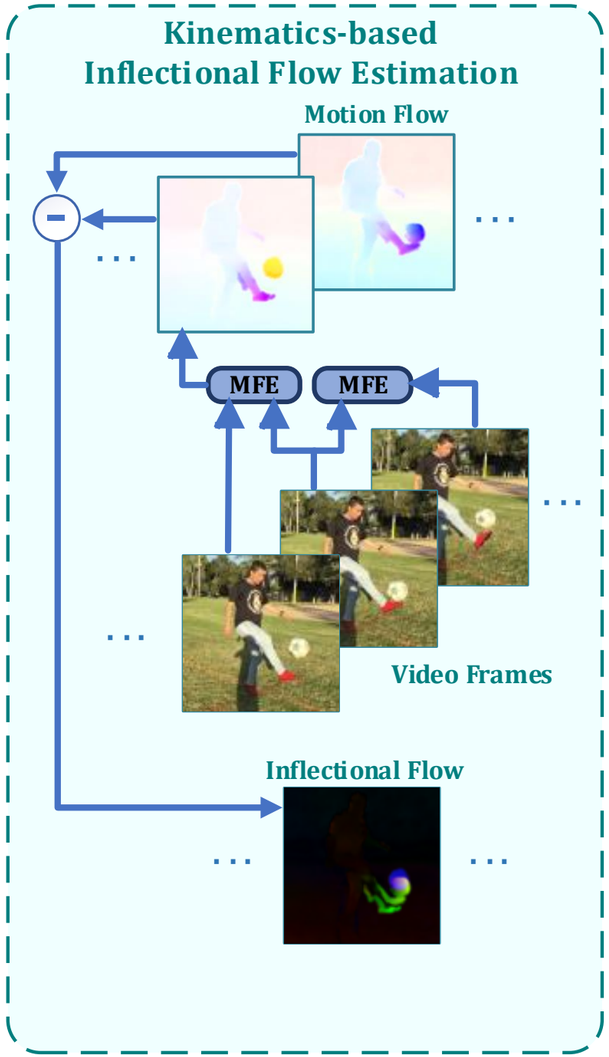


Discriminative maps

Self-supervised
Spatial Auxiliary
Training Strategy

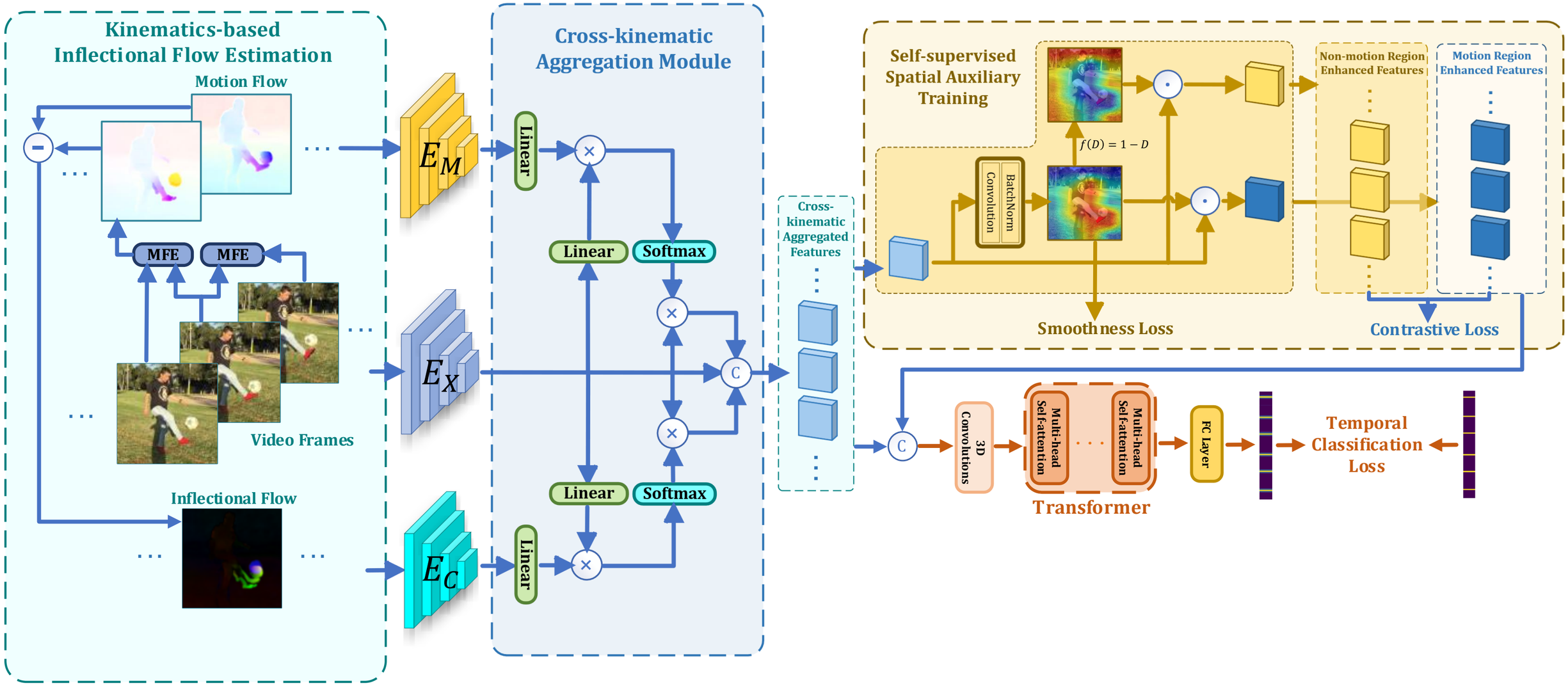
Spatial Modelling (Motion Localization)

Motion Assodeltign

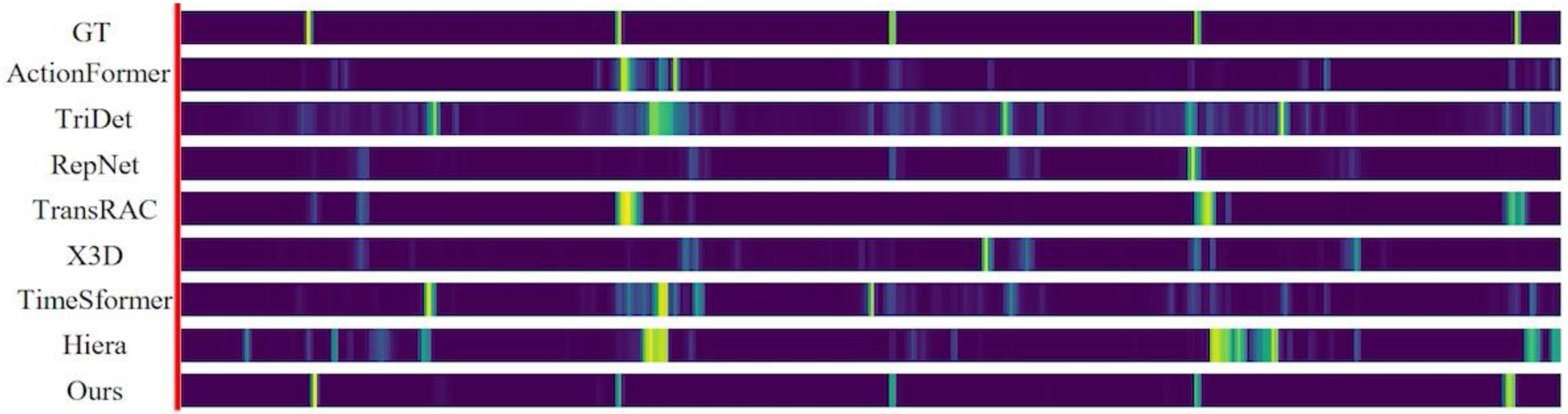

Kinematics-based
Inflectional Flow Estimation
Motion Flow
MFE
MFE
Video Frames
Inflectional Flow

Overview

Inflectional Flow Estimation


Kinematics-based
Inflectional Flow Estimation
Motion Flow
MFE
MFE
Video Frames
Inflectional Flow
E_M
E_X
E_C
Cross-kinematic
Aggregation Module
Linear
Linear
Softmax
Linear
Softmax
Linear
C
Cross-
kinematic
Aggregated
Features
Self-supervised
Spatial Auxiliary
Training
$f(D) = 1 - D$
Convolution
BatchNorm
Smoothness Loss
Non-motion Region
Enhanced Features
Motion Region
Enhanced Features
Contrastive Loss
C
3D
Convolutions
Multi-head
Self-attention
Multi-head
Self-attention
Transformer
FC Layer
Temporal
Classification
Loss

Overview

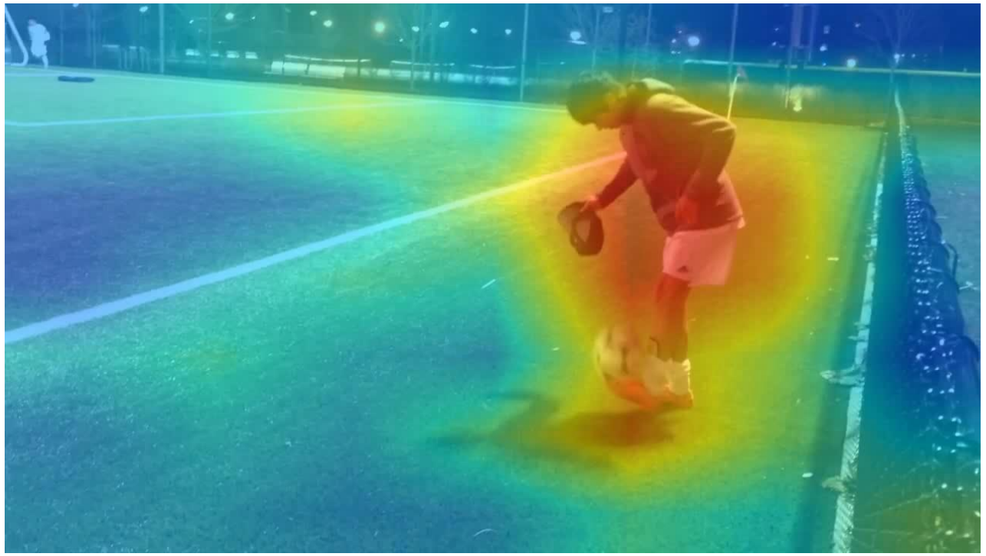


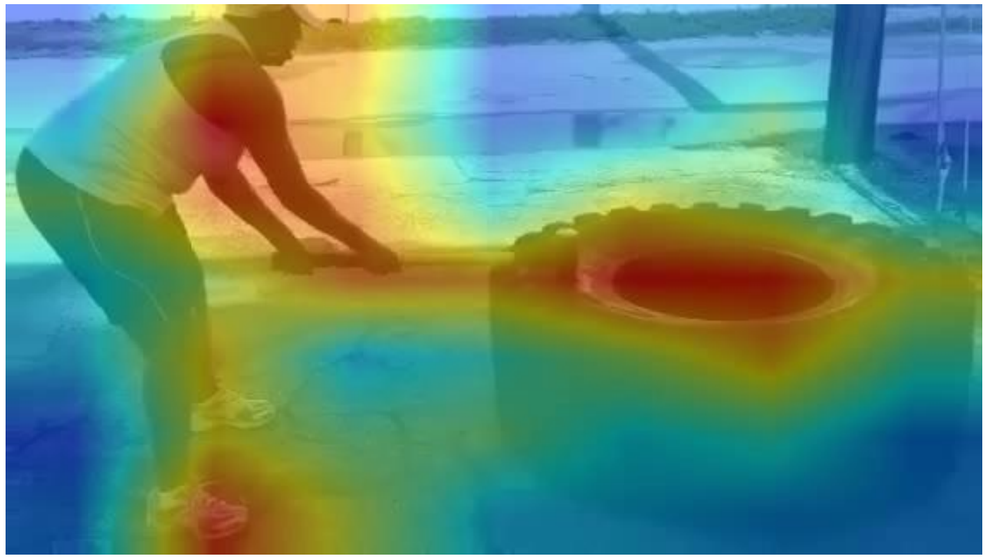
Temporal Probabilities Predictions

Method	UCFRep		CountixAV	
	MAE↓	OBO↑	MAE↓	OBO↑
RepNet	0.915	0.074	0.749	0.231
TransRAC	<u>0.594</u>	**0.222**	<u>0.686</u>	<u>0.255</u>
X3d	1.245	0.037	0.876	0.192
TimeSformer	0.832	0.0	1.551	0.185
Hiera	0.791	0.152	3.648	0.231
Ours	**0.588**	<u>0.185</u>	**0.549**	**0.346**

Table 3: Counting performance on UCFRep and CountixAV datasets. Top-2 results are marked in **bold** and <u>underlined</u>. Our method has not been trained on any counting dataset.

Counting Performance Evaluation

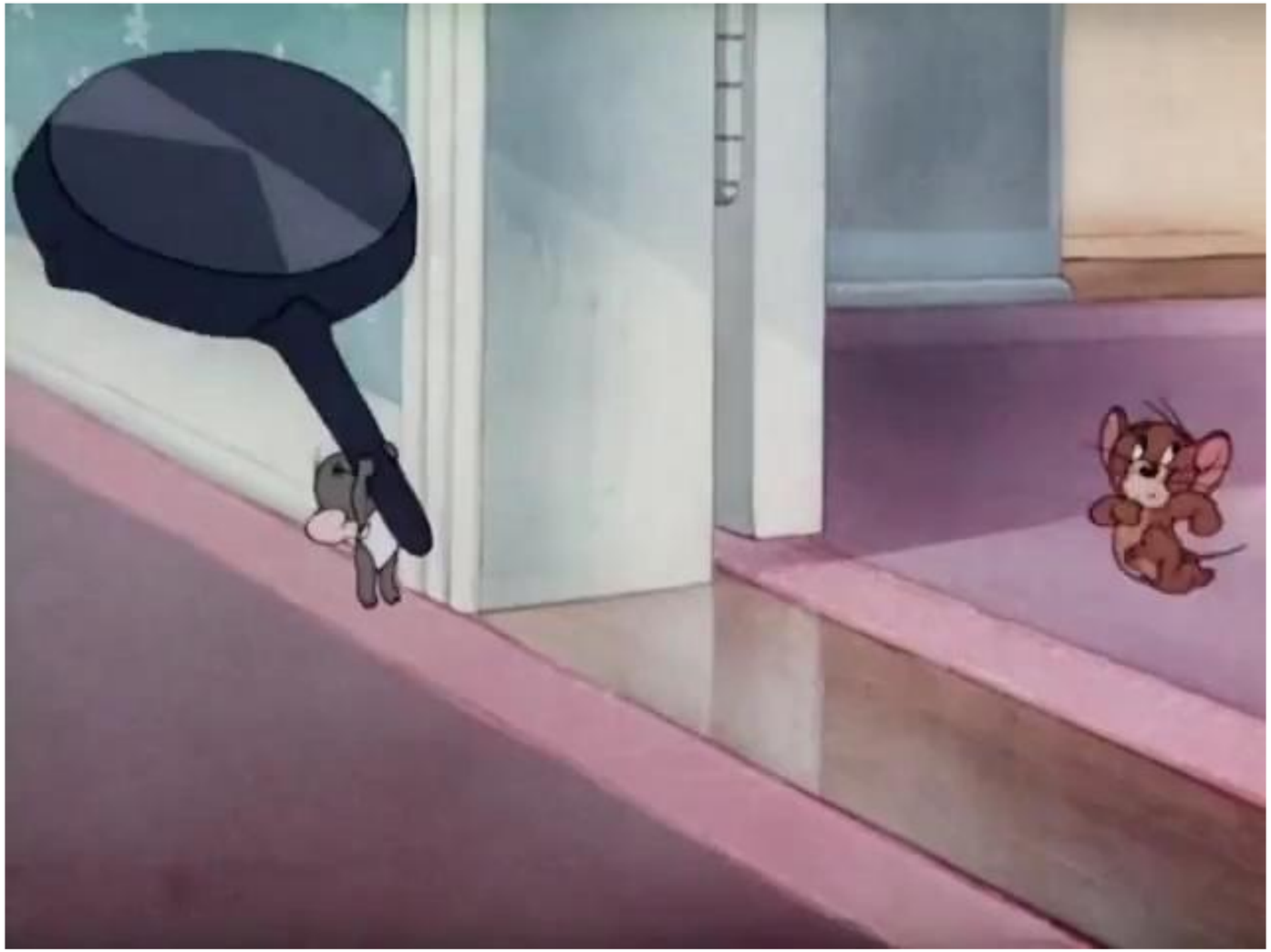




Self-supervised Spatial Localization Visualization

Video Re-dubbing Result

Video Re-dubbing Result



Video Re-dubbing Result

Thanks for watching!