Hierarchical Temporal Transformer for 3D Hand Pose Estimation and Action Recognition from Egocentric RGB Videos

Yilin Wen[1], Hao Pan[2], Lei Yang[3,1], Jia Pan[1], Taku Komura[1], Wenping Wang[4]

[1]The University of Hong Kong [2]Microsoft Research Asia [3]TransGP [4]Texas A&M University

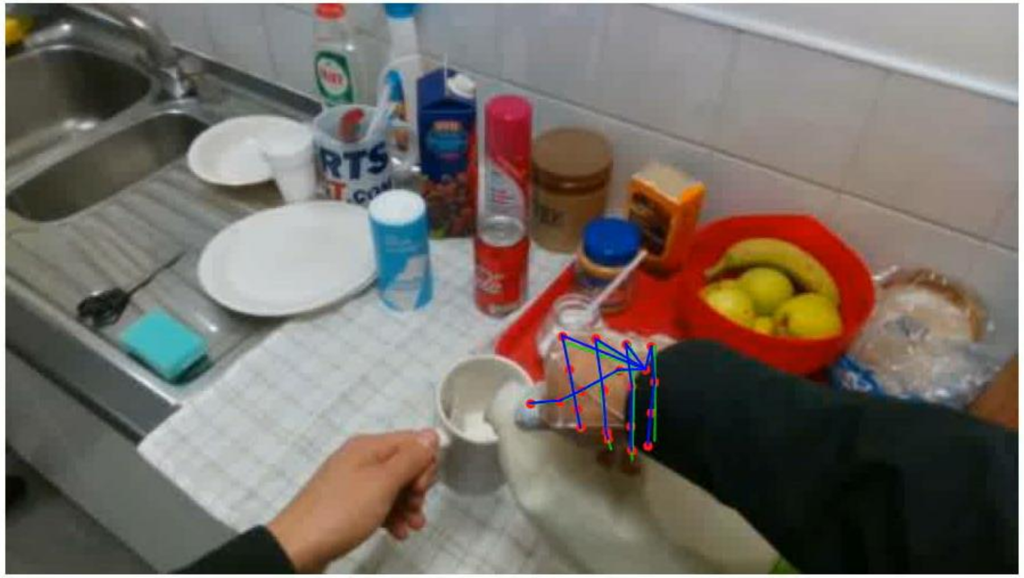
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

THU-PM-061

Tasks

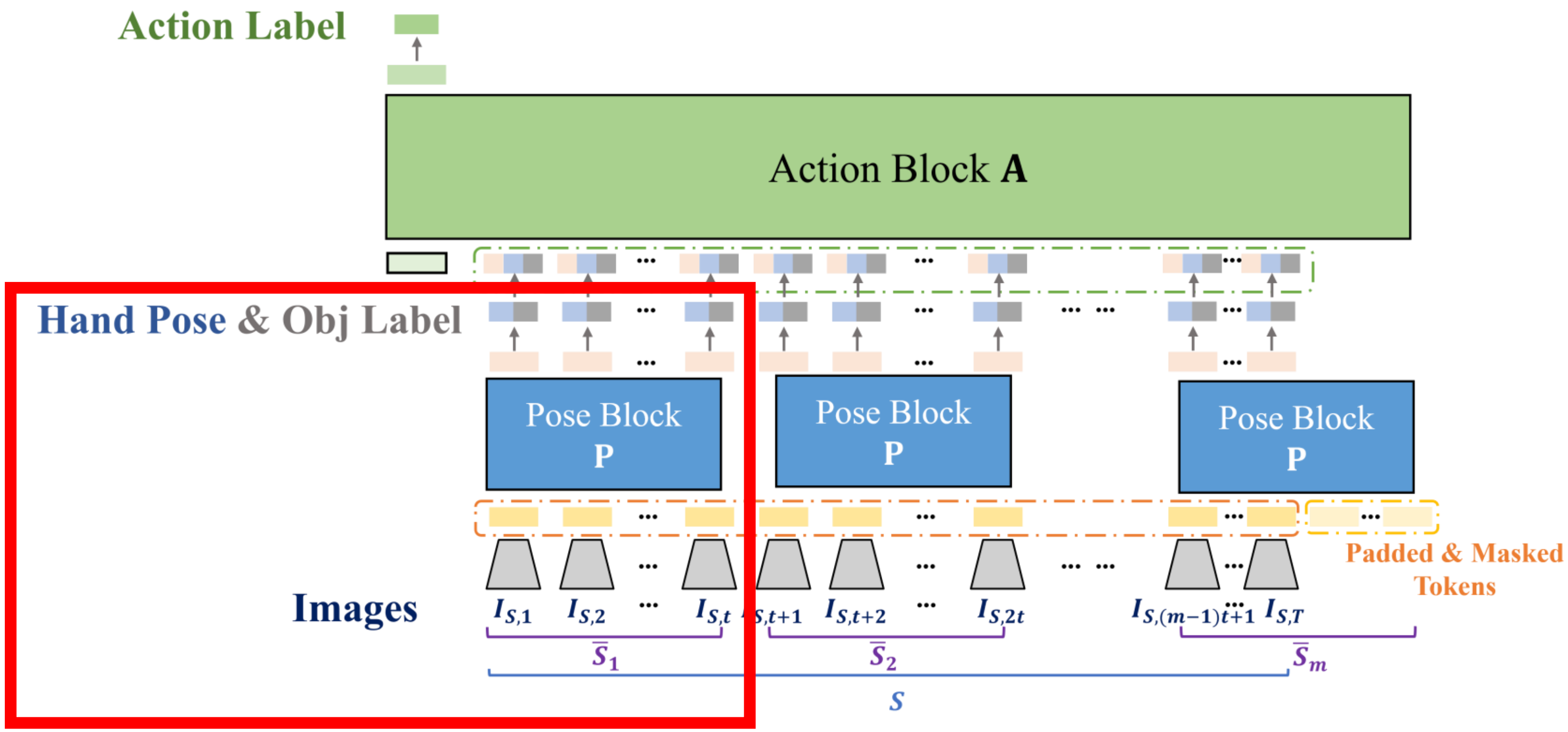
- **Input:** Observed RGB video under egocentric view
- **Output:** 1) Per-frame 3D hand joints position in the camera space;
  2) Performed action category.

Video from FPHA dataset[1]
Hand pose for the frames:
GT / Est.
Action for the sequence:
Pour milk (GT) / Pour milk (Est.)



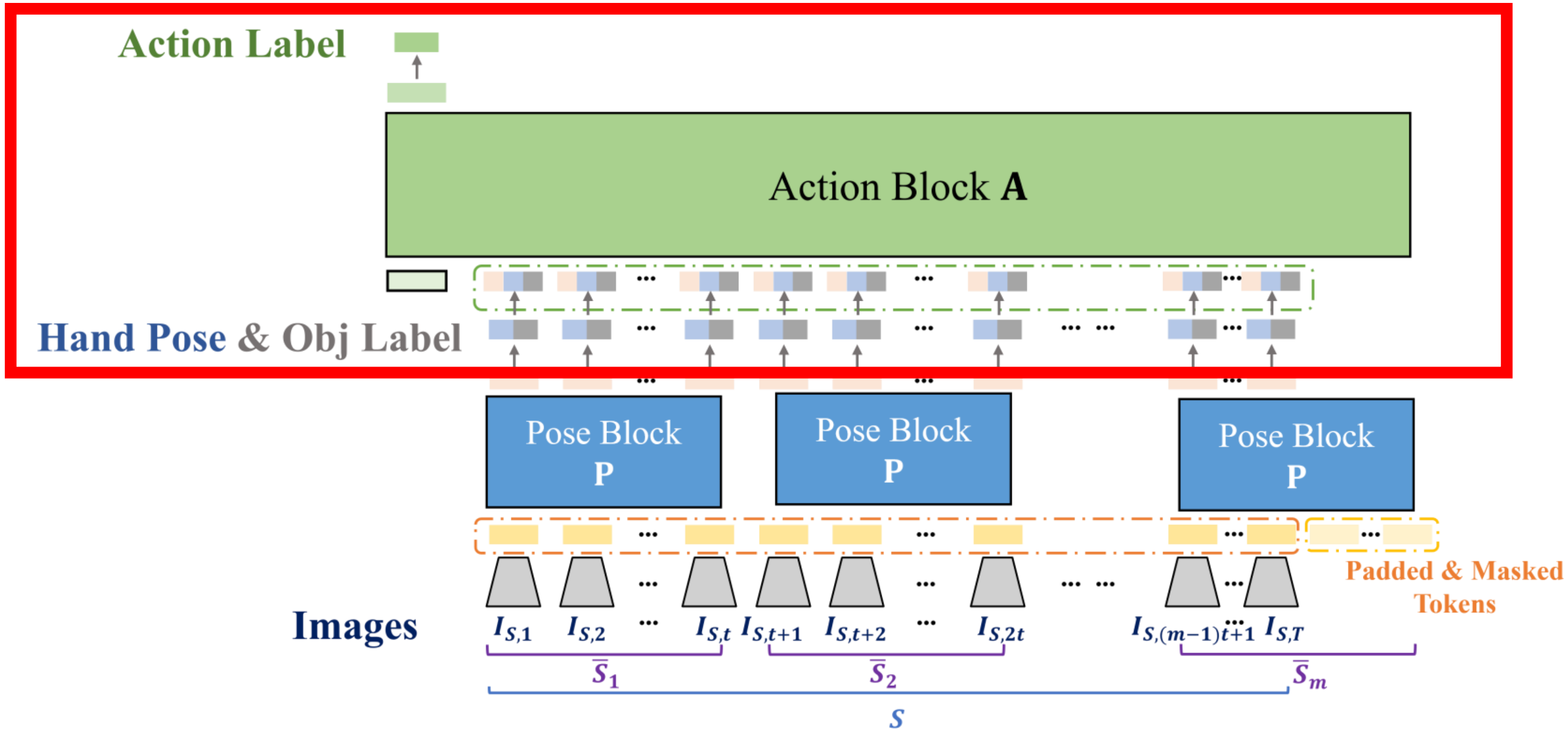
Overview

➢ A ***hierarchical temporal transformer*** with two cascaded blocks



Overview

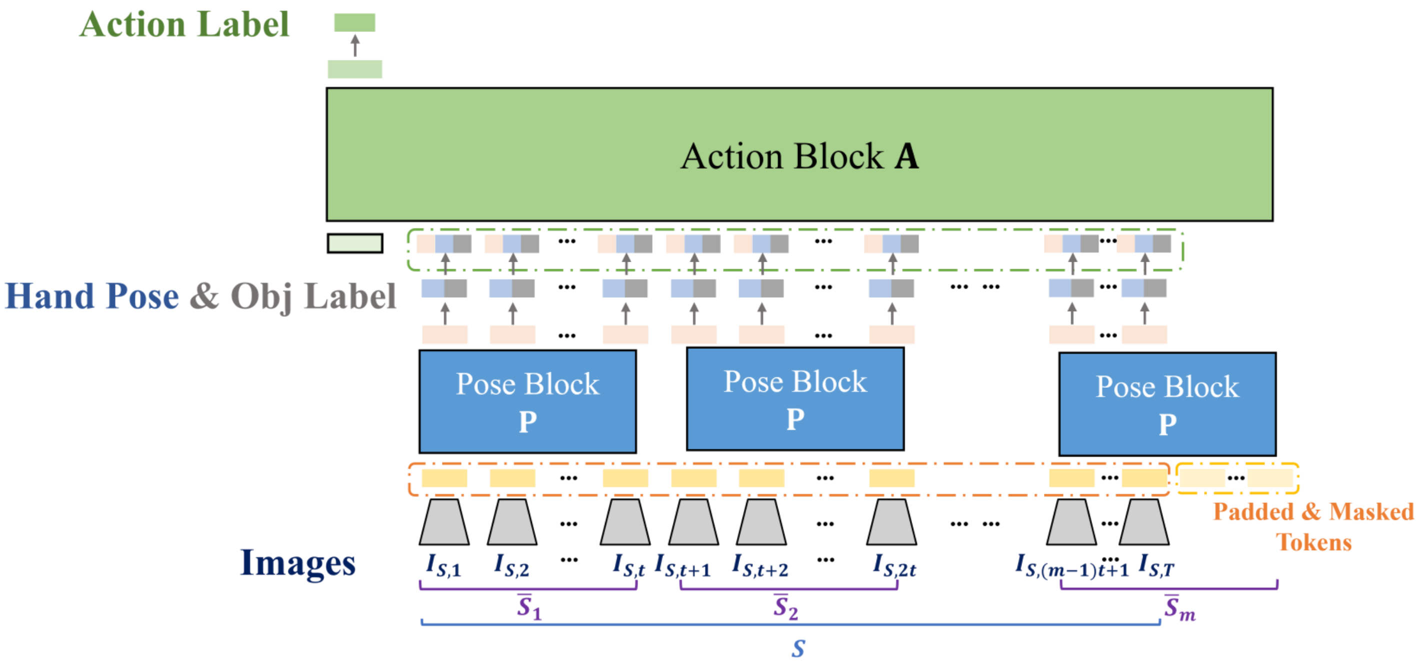
- A ***hierarchical temporal transformer*** with two cascaded blocks



Overview

➢ A ***hierarchical temporal transformer*** with two cascaded blocks, to:

1. ***leverage different time spans for pose and action estimation.***
2. ***model their semantic correlation*** by deriving the high-level action from the low-level hand motion and manipulated object label.

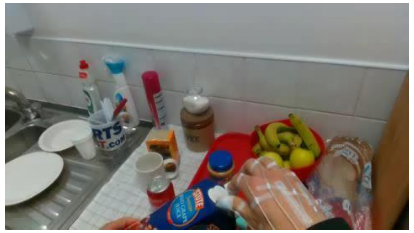


Motivation

Observations

- ➢Severe ambiguity of action types judged from individual frames
- ➢Frequent occlusion and truncations for per-frame hand pose.

Close Juice Bottle



Open Juice Bottle

Pour Liquid Soap

Tear Paper

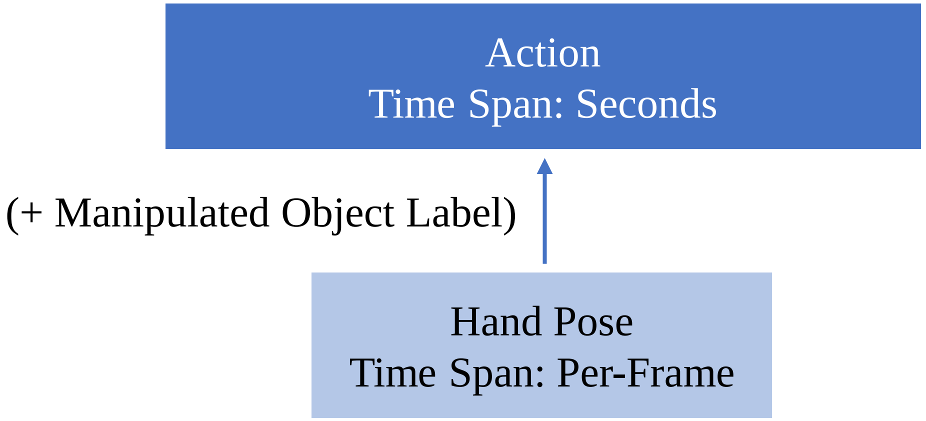
Our Key Designs

✓Leverage the temporal information for both pose and action

Motivation

Observations

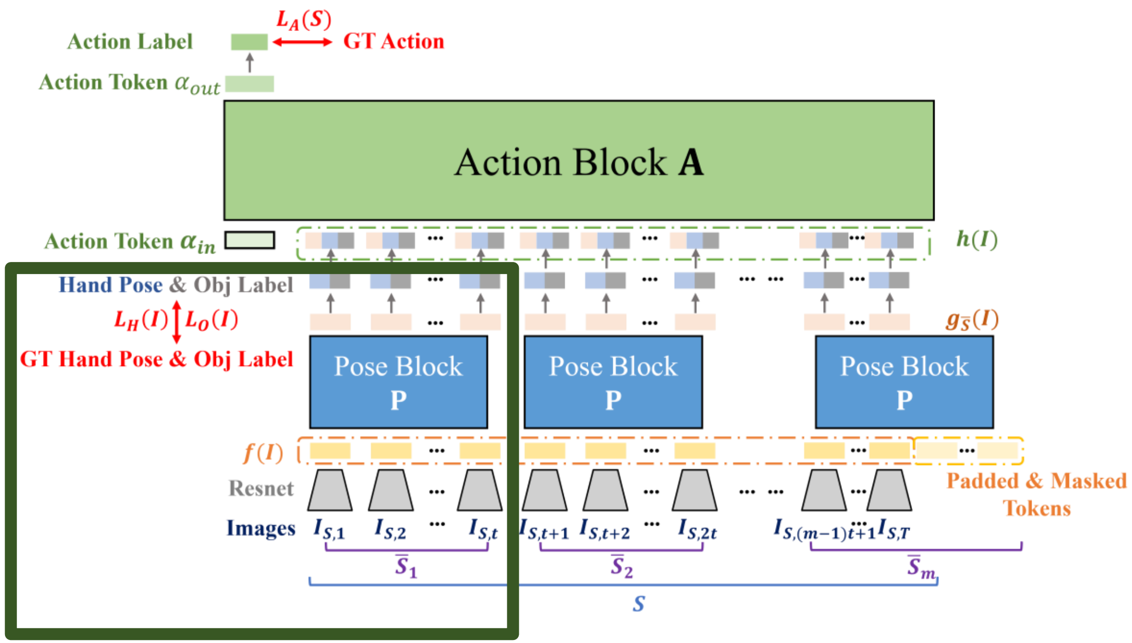
- ➢ Severe ambiguity of action types judged from individual frames
- ➢ Frequent occlusion and truncations for per-frame hand pose.
- ➢ Different temporal granularity and semantic correlation between hand-action.



Our Key Designs

- ✓ Leverage the temporal information for both pose and action
- ✓ Build a hierarchical temporal transformer with two cascaded blocks, to cope with the different temporal granularity and semantic correlation between hand-action.

Framework



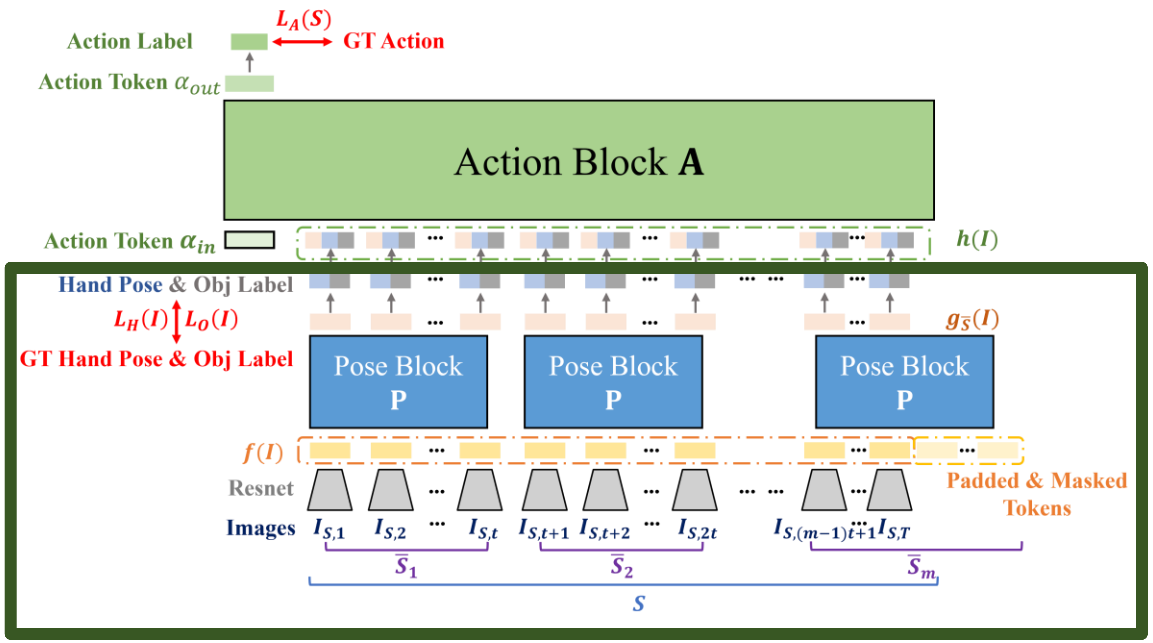
Hand Pose Estimation with Short-Term Temporal Cue

- ➢ Pose block ***P*** focuses on a narrower temporal receptive field to output per-frame 3D hand pose and manipulated object label.

$L_H(I) = \left\| H_I^{2.5D} - H_{I,gt}^{2.5D} \right\|_1$. $H^{2.5D}$ is 2D+Depth for hand joints

$L_O(I) = CE(w_I^O, \Pr(O|I))$. w_I^O is a one-hot vector for the target distribution.

Framework



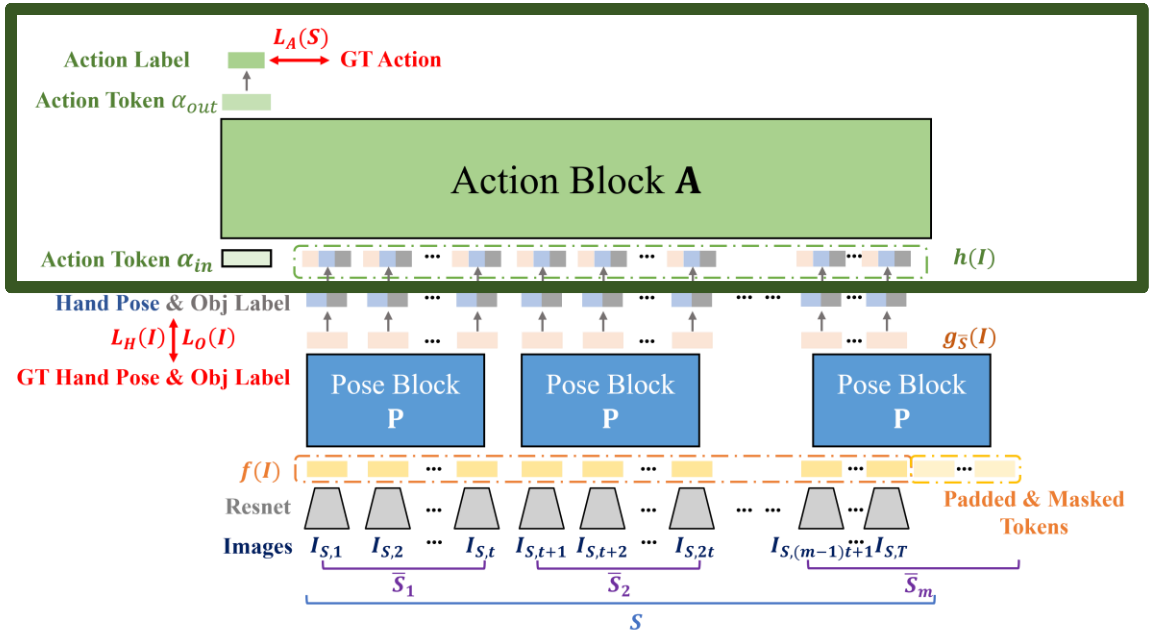
Hand Pose Estimation with Short-Term Temporal Cue

- Pose block ***P*** focuses on a narrower temporal receptive field to output per-frame 3D hand pose and manipulated object label.
- Input video is divided into consecutive segments by a shifting window strategy with window size t, segments are processed by ***P*** in parallel.

$L_H(I) = \left\|H_I^{2.5D} - H_{I,gt}^{2.5D}\right\|_1$. $H^{2.5D}$ is 2D+Depth for hand joints

$L_O(I) = CE(w_I^O, \Pr(O|I))$. w_I^O is a one-hot vector for the target distribution.

Framework

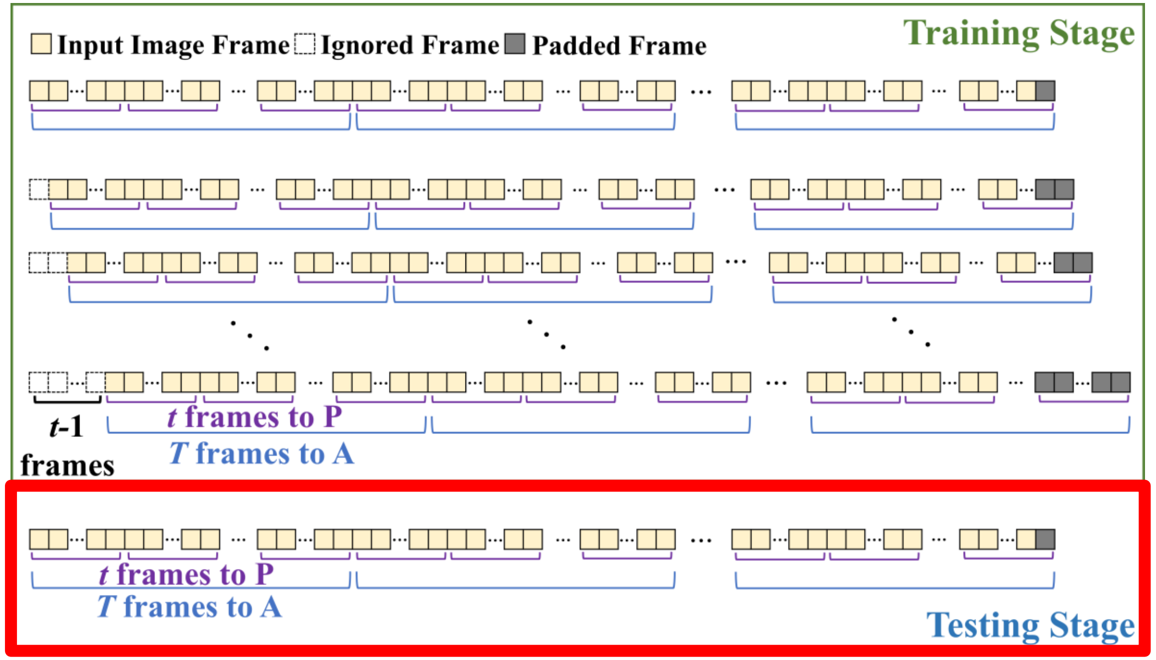


Action Recognition with Long-Term Temporal Cue

- Action block $\boldsymbol{A}$ uses the full video to predict the action label.
- The input of $\boldsymbol{A}$ leverages the per-frame predicted hand pose, object label and image feature.

$L_A(S) = CE(w_S, \Pr(A|I))$. w_S is a one-hot vector for the target distribution.

Framework

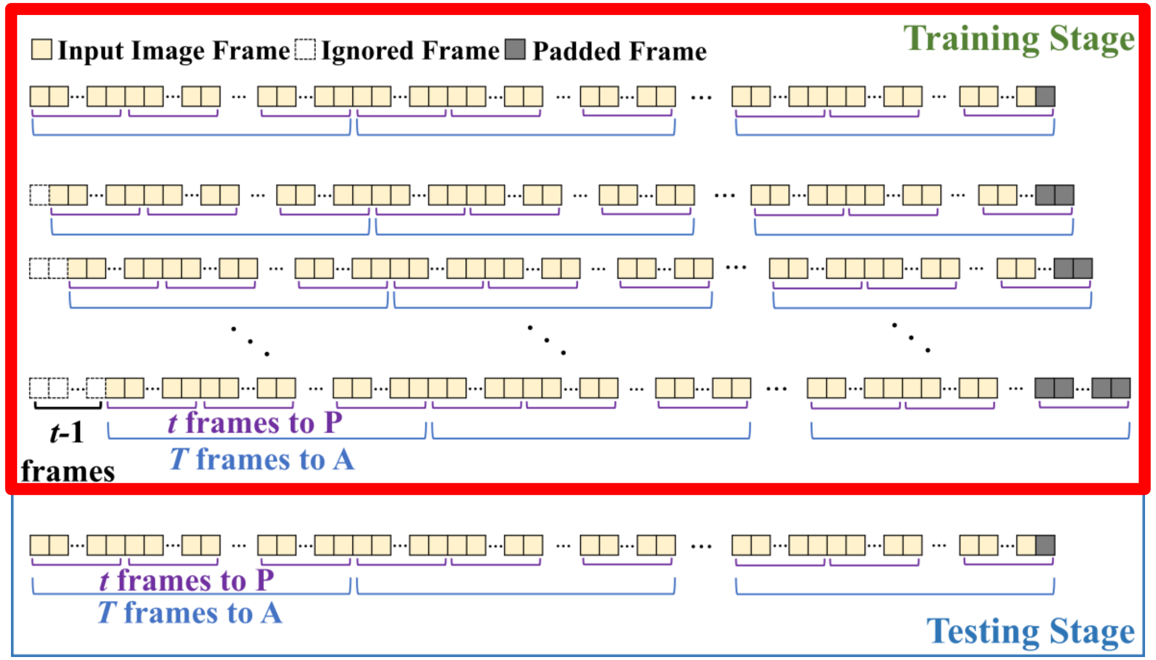


Segmentation Strategy to Divide Long Videos into HTT Inputs

Videos longer than T frames are divided into consecutive clips by adopting the shifting window strategy with a window size T:

- In testing stage, the hand pose are estimated by $\boldsymbol{P}$, the action category is voted from the predictions among segmented clips.

Framework



Segmentation Strategy to Divide Long Videos into HTT Inputs

Videos longer than T frames are divided into consecutive clips by adopting the shifting window strategy with a window size T:

- In testing stage, the hand pose are estimated by ***P***, the action category is voted from the predictions among segmented clips.
- In training stage, for data augmentation, the starting frame for shifting window is offset to each of the first t frames.

Results (FPHA[1])

Action (GT / Est.) :
Scratch Sponge / Scratch Sponge

Action (GT / Est.):
Close Peanut Butter / Close Peanut Butter

Hand pose: GT / Est.

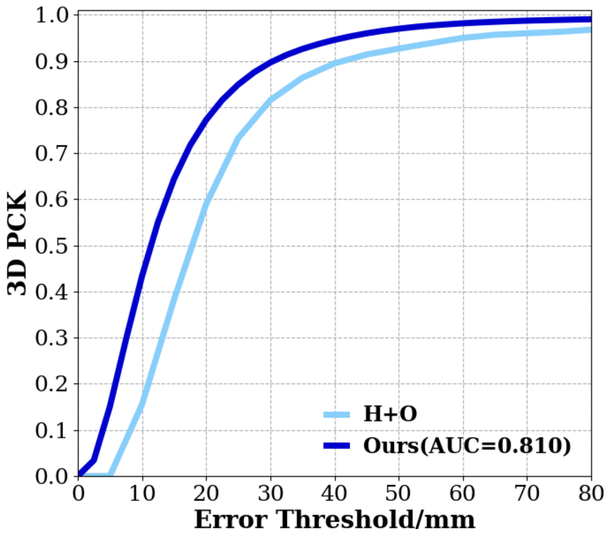
Results (FPHA[1])

Action Recognition

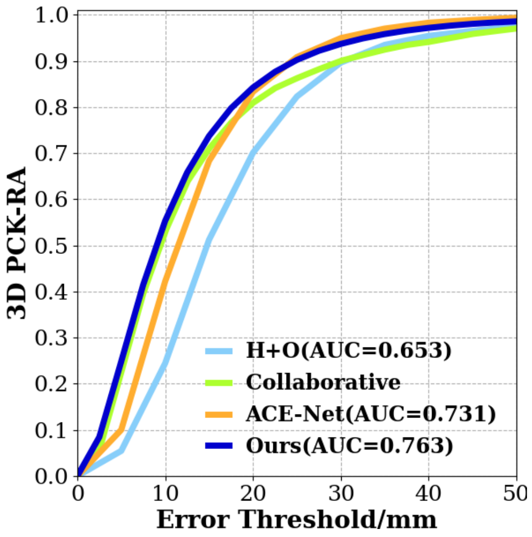
	Joule-color [2]	Two-Stream [3]	H+O [4]	Collaborative [5]	Ours
Process both hand-action			√	√	√
Acc.	66.78	75.30	82.43	85.22	**94.09**

3D Hand Pose Estimation

Camera Space



Root-Aligned Space

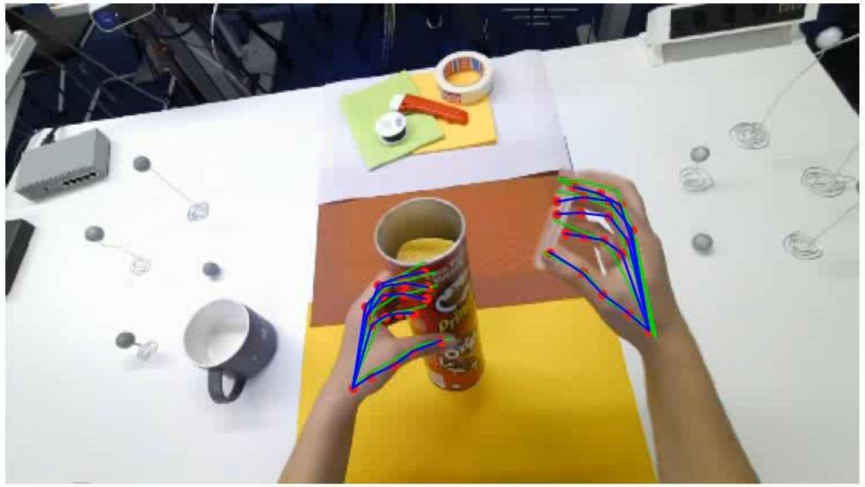


	Input	**Process both hand-action**
H+O [4]	Image	√
Collaborative [5]	Video	√
ACE-Net[6]	Video	
Ours	Video	√

Results (H2O[7])

Action (GT / Est.) :
Put in Cocoa / Put in Cocoa

Action (GT / Est.) :
Close Chips / Close Chips



Hand pose: GT / Est.

Results (H2O[7])

Action Recognition

	Process both hand-action	**Acc.**
C2D [8]		70.66
I3D [9]		75.21
SlowFast [10]		77.69
H+O [4]	√	68.88
H2O w/ ST-GCN [7]	√	73.86
H2O w/ TA-GCN [7]	√	79.25
Ours	√	**86.36**

3D Hand Pose Estimation

		LPC [11]	H+O [4]	H2O [7]	Ours
Input		Image	Image	Image	Video
Process both hand-action			√	√	√
MEPE (in *mm*) **Camera Space**	**Left**	39.56	41.42	41.45	**35.02**
	Right	41.87	38.86	37.21	**35.63**

Ablation and Visualization

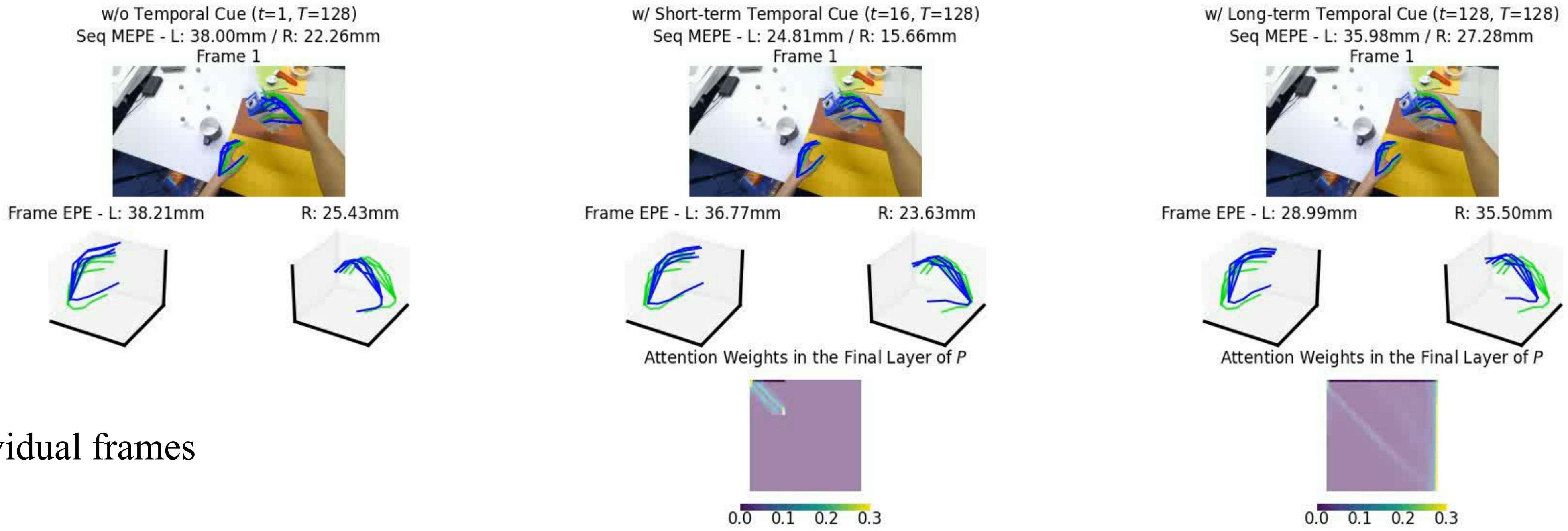
➢ *Hand Pose Estimation with Short-Term Temporal Cue.*

Est. vs **GT**

2D Projection

In Camera Space

Attention weights for
current frame → individual frames
is highlighted



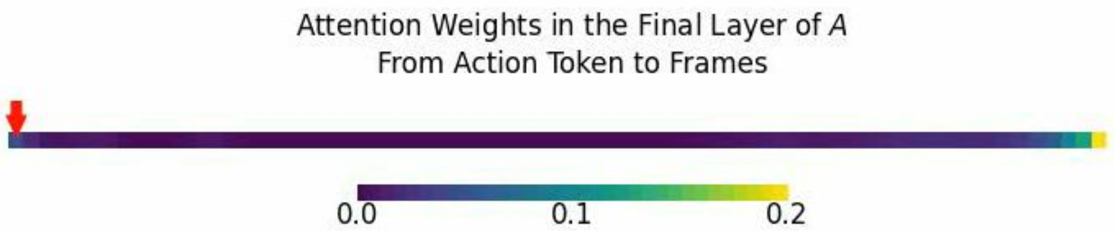
- ➢ Compared with *t=1*, we show enhanced robustness under occlusion and truncation.
- ➢ Compared with *t=128*, we avoid over-attending to distant frames, therefore ensuring sharp local motion.

Video from H2O dataset [7]

Ablation and Visualization

➢ ***Action Recognition with Long-Term Temporal Cue***

The arrow indicates the attention to the current frame



➢ The last few frames are the key for recognizing the action of *take out espresso.*

➢ In response our network pays most attention to these frames.

Video from H2O dataset [7]

Concluding Remarks

- Task
  - 3D hand pose estimation and action recognition from egocentric RGB videos

- Key ideas
  - Leverage the temporal information for both pose and action.
  - Build a hierarchical temporal transformer with two cascaded blocks, to cope with the different temporal granularity and semantic correlation between hand-action.

- Results
  - State-of-the-art results on FPHA and H2O datasets.

Reference

[1] Garcia-Hernando, G., Yuan, S., Baek, S., & Kim, T. K. (2018). First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 409-419).

[2] Hu, J. F., Zheng, W. S., Lai, J., & Zhang, J. (2015). Jointly learning heterogeneous features for RGB-D activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5344-5352).

[3] Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1933-1941).

[4] Tekin, B., Bogo, F., & Pollefeys, M. (2019). H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4511-4520).

[5] Yang, S., Liu, J., Lu, S., Er, M. H., & Kot, A. C. (2020). Collaborative learning of gesture recognition and 3D hand pose estimation with multi-order feature analysis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16* (pp. 769-786). Springer International Publishing.

[6] Fan, Z., Liu, J., & Wang, Y. (2020). Adaptive computationally efficient network for monocular 3d hand pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16* (pp. 127-144). Springer International Publishing.

[7] Kwon, T., Tekin, B., Stühmer, J., Bogo, F., & Pollefeys, M. (2021). H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10138-10148).

[8] Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794-7803).

[9] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299-6308).

[10] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6202-6211).

[11] Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., & Schmid, C. (2020). Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 571-580).