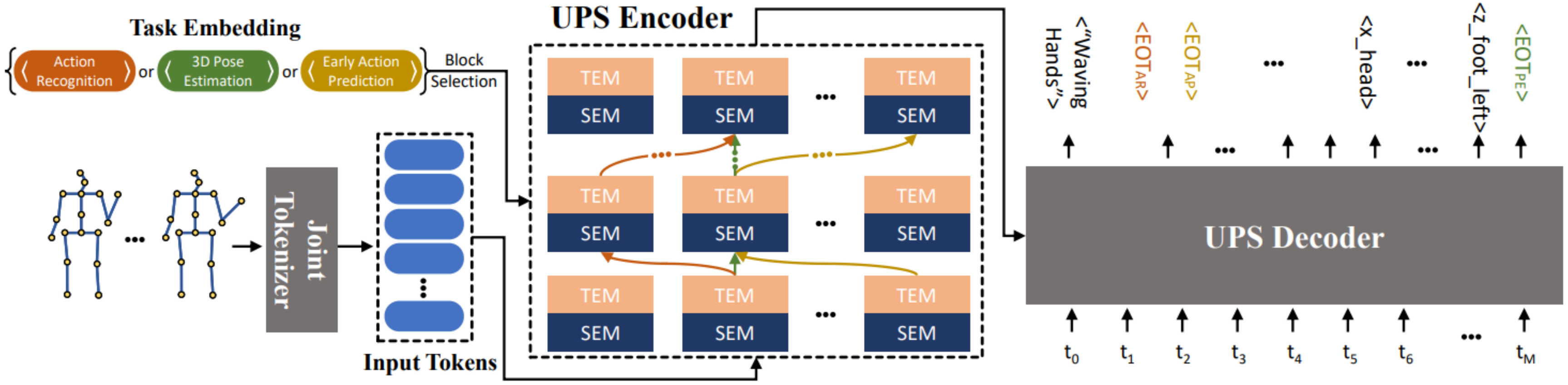
SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

Unified Pose Sequence Modeling

Lin Geng Foo[1*] Tianjiao Li[1*] Hossein Rahmani[2] Qiuhong Ke[3] Jun Liu[1†]
[1]Singapore University of Technology & Design [2]Lancaster University [3]Monash University
{lingeng_foo,tianjiao_li}@mymail.sutd.edu.sg, h.rahmani@lancaster.ac.uk,
qiuhong.ke@monash.edu, jun_liu@sutd.edu.sg

Paper Tag: WED-PM-063

TL;DR



Pose-based tasks, e.g., action recognition, early action prediction and pose estimation, are hot research topics and have been well explored in the deep learning community. However, the existing methods for the aforementioned tasks still often require task-specific architectures, for instance, hourglass networks for pose estimation and specialized GCN for action recognition and early action prediction. All these task-specific models can be inconvenient and inefficient.

Therefore, in this work, we propose the UPS, Unified Pose Sequence Modeling, to unify heterogeneous output formats for different pose-based tasks (action recognition, early action prediction and 3D pose estimation) by considering the text-based action labels and coordinate-based human poses as a form of unified language sequences.

Foo, L. G., Li, T., Rahmani, H., Ke, Q., & Liu, J. (2023). Unified pose sequence modeling. In *Proceedings of the IEEE/CVF CVPR*.

Motivation

Why do we need a unified pose sequence model?

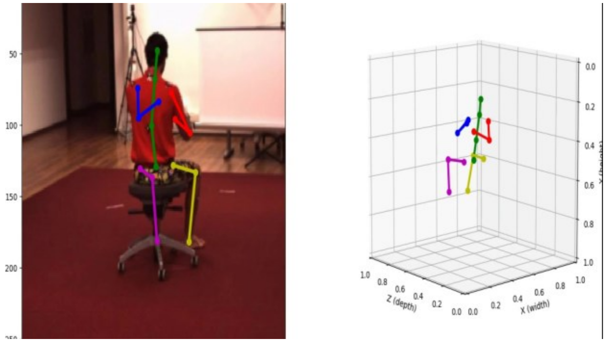
Action Recognition

Monitoring Pedestrian Movements

Early Action Prediction

Anticipating danger in advance

3D Pose Estimation



Perceiving Spatial Environments

These people are crossing the street.

The man is going to take out a gun.

Head locates at …
Hands locate at …

Foo, L. G., Li, T., Rahmani, H., Ke, Q., & Liu, J. (2023). Unified pose sequence modeling. In *Proceedings of the IEEE/CVF CVPR*.

Motivation

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

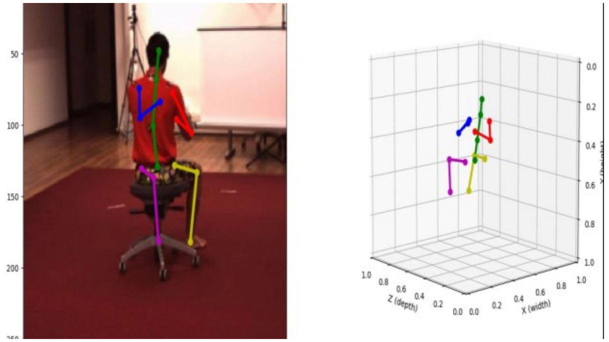
Action Recognition

Monitoring Pedestrian Movements

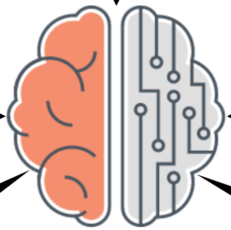
Early Action Prediction

Anticipating danger in advance

3D Pose Estimation



Perceiving Spatial Environments



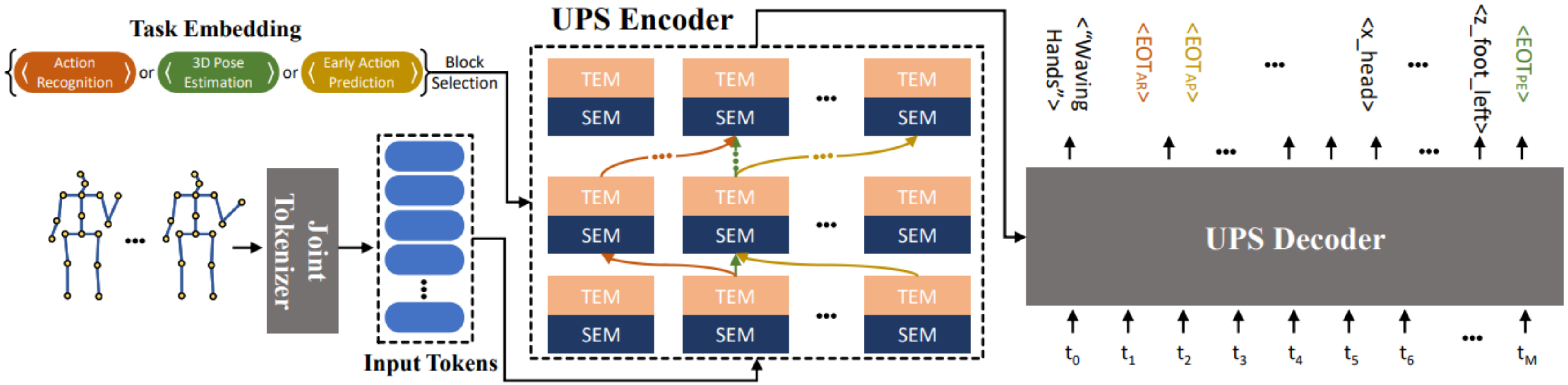
These people are crossing the street.

The man is going to take out a gun.

Head locates at …
Hands locate at …

Foo, L. G., Li, T., Rahmani, H., Ke, Q., & Liu, J. (2023). Unified pose sequence modeling. In *Proceedings of the IEEE/CVF CVPR*.

UPS



Network Components:

- Joint Tokenizer
- UPS Encoder with Dynamic Routing Mechanism
- UPS Decoder with Unified Vocabulary

Input for UPS:

- Human poses in coordinates format

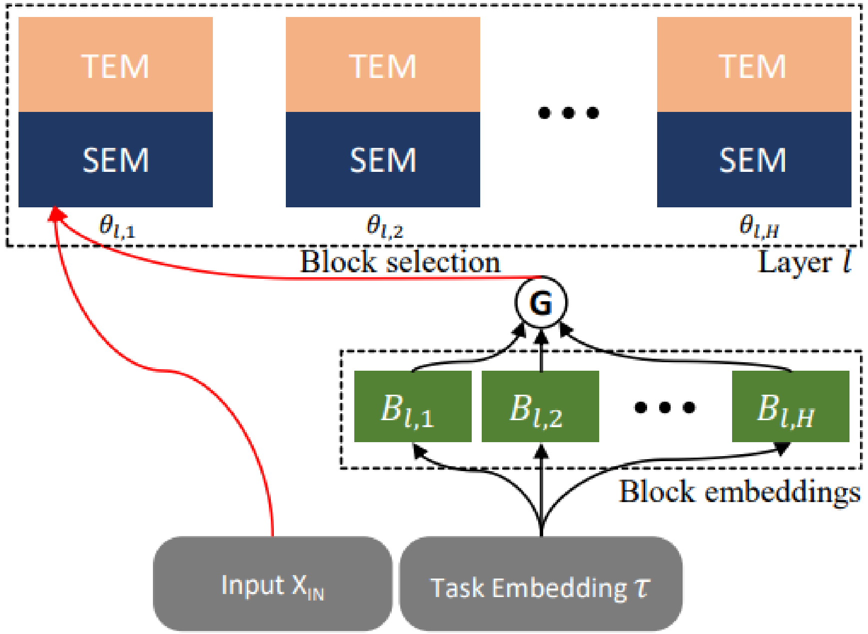
Output for UPS:

- Unified language sequence

Foo, L. G., Li, T., Rahmani, H., Ke, Q., & Liu, J. (2023). Unified pose sequence modeling. In *Proceedings of the IEEE/CVF CVPR*.

Dynamic Routing

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

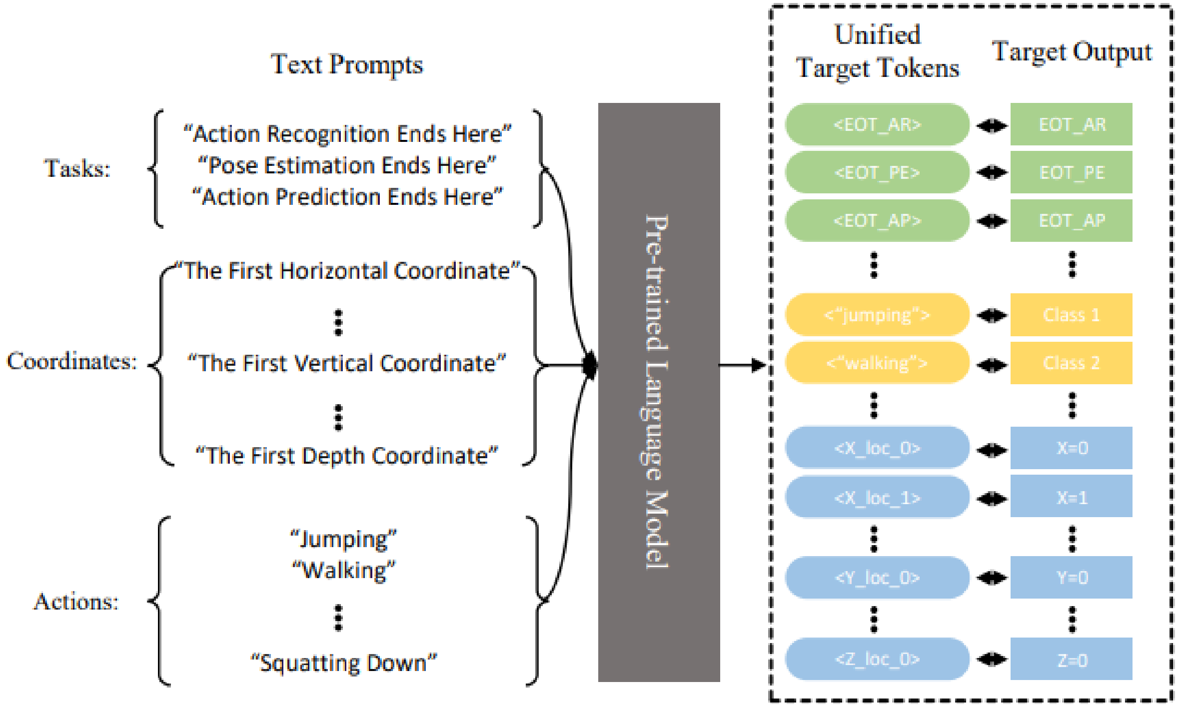


The dot products between task embedding τ and block embeddings $\{B_{l,h}\}_{h=1}^{H}$ are computed and send into the Gumbel-Softmax to select the best-matched block in each layer from $\{\theta_{l,h}\}_{h=1}^{H}$.

Note that τ and $\{B_{l,h}\}_{h=1}^{H}$ are learnable parameters during training and are optimized. Therefore, by iteratively updating, our dynamic routing mechanism can select the most suitable block conditioned on the input tasks.

Foo, L. G., Li, T., Rahmani, H., Ke, Q., & Liu, J. (2023). Unified pose sequence modeling. In *Proceedings of the IEEE/CVF CVPR*.

Unified Vocabulary



Unified Vocabulary

We use text descriptions to represent (1) action category labels, (2) joint coordinate values and (3) task-ending indicators. These text descriptions are sent to the off-the-shelf RoBERTa to extract text features as our target tokens.

Therefore, the heterogeneous output formats are unified in the formation of language sequences.

Foo, L. G., Li, T., Rahmani, H., Ke, Q., & Liu, J. (2023). Unified pose sequence modeling. In *Proceedings of the IEEE/CVF CVPR*.

Experiments

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Methods	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
Pavllo *et al.* [54]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Lin *et al.* [40]	42.5	44.8	42.6	44.2	48.5	57.1	42.6	41.4	56.5	64.5	47.4	43.0	48.1	33.0	35.1	46.6
Cai *et al.* [5]	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
Xu *et al.* [81]	37.4	43.5	42.7	42.7	46.6	59.7	41.3	45.1	52.7	60.2	45.8	43.1	47.7	33.7	37.1	45.6
Wang *et al.* [74]	41.3	43.9	44.0	42.2	48.0	57.1	42.2	43.2	57.3	61.3	47.0	43.5	47.0	32.6	31.8	45.6
Liu *et al.* [46]	41.8	44.8	41.1	44.9	47.4	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.2	45.1
Zeng *et al.* [91]	46.6	47.1	43.9	41.6	45.8	49.6	46.5	40.0	53.4	61.1	46.1	42.6	43.1	31.5	32.6	44.8
Zheng *et al.* [96]	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
Chen *et al.* [7]	41.4	43.5	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	41.7	46.0	31.5	32.7	44.1
Shan *et al.* [65]	38.4	42.1	39.8	40.2	45.2	48.9	40.4	38.3	53.8	57.3	43.9	41.6	42.2	29.3	29.3	42.1
$\text{UPS}_{\text{separate}}$	39.4	44.2	38.0	42.5	43.6	52.5	40.9	49.2	53.6	70.5	43.5	45.3	48.1	30.0	31.9	44.9
UPS	37.5	39.2	36.9	40.6	39.3	46.8	39.0	41.7	50.6	63.5	40.4	37.8	44.2	26.7	29.1	**40.8**

Methods	NTU60		NTU120	
	xsub	xview	xsub	xset
ST-GCN [88]	81.5	88.3	70.7	73.2
2s-AGCN [67]	88.5	95.1	82.2	84.1
Shift-GCN [12]	90.7	96.5	85.9	87.6
MS-G3D [48]	91.5	96.2	86.9	88.4
DSTA-Net [68]	91.5	96.4	86.6	89.0
CTR-GCN [11]	92.4	96.8	88.9	90.6
PoseConv3D [21]	**94.1**	**97.1**	86.9	90.3
InfoGCN [13]	93.0	**97.1**	**89.8**	**91.2**
$\text{UPS}_{\text{separate}}$	89.6	93.1	85.1	87.8
UPS	92.6	97.0	89.3	91.1

Methods	Observation Ratios on NTU60		
	20%	40%	60%
Jain *et al.* [29]	7.07	18.98	44.55
Ke *et al.* [34]	8.34	26.97	56.78
Weng *et al.* [78]	35.56	54.63	67.08
Aliakbarian *et al.* [63]	27.41	59.26	72.43
Wang *et al.* [77]	35.85	58.45	73.86
Pang *et al.* [52]	33.30	56.94	74.50
Tran *et al.* [72]	24.60	57.70	76.90
Ke *et al.* [35]	32.12	63.82	77.02
Li *et al.* [39]	42.39	72.24	82.99
Foo *et al.* [23]	**53.98**	74.34	85.03
$\text{UPS}_{\text{separate}}$	50.11	69.84	82.59
UPS	53.25	**75.06**	**85.35**

Here, note that $\textbf{UPS}_{\textbf{separate}}$ is optimized separately on each task, and **UPS** represents our full model which is optimized on all tasks.

Foo, L. G., Li, T., Rahmani, H., Ke, Q., & Liu, J. (2023). Unified pose sequence modeling. In *Proceedings of the IEEE/CVF CVPR*.

Fin & Thanks!

Foo, L. G., Li, T., Rahmani, H., Ke, Q., & Liu, J. (2023). Unified pose sequence modeling. In *Proceedings of the IEEE/CVF CVPR*.