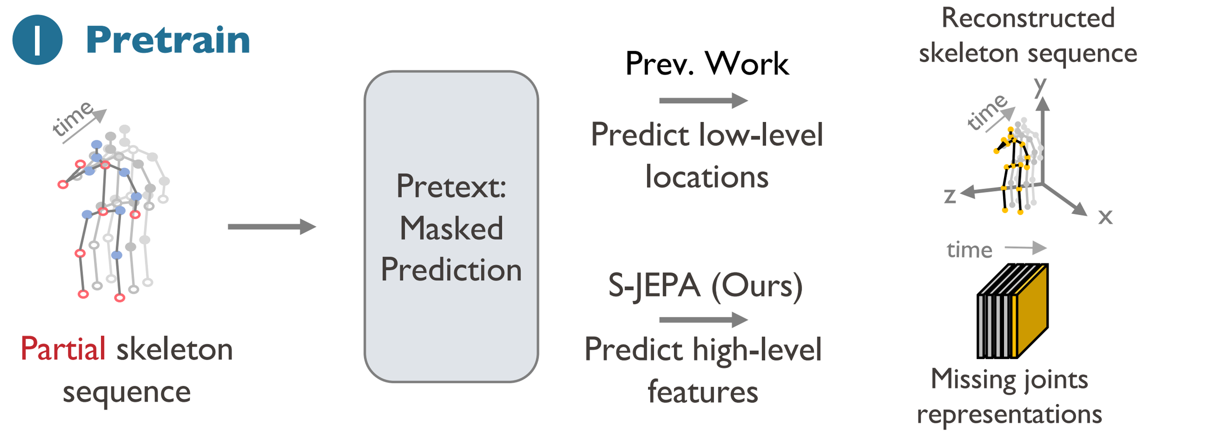
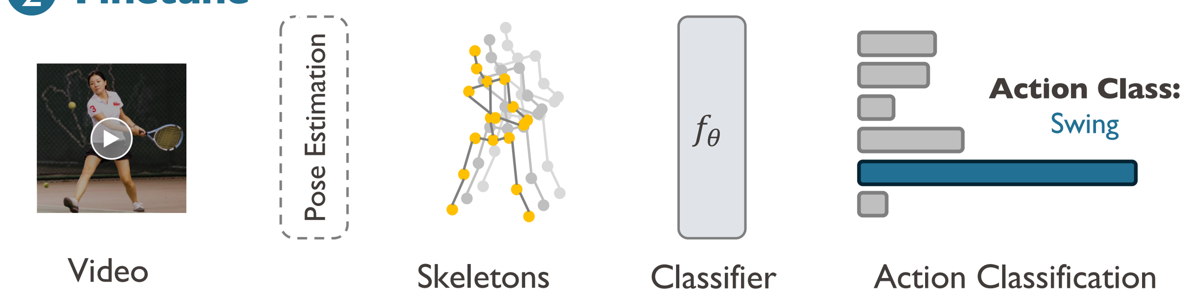
S-JEPA: A Joint Embedding Predictive Architecture for Skeletal Action Recognition

Mohamed Abdelfattah* and Alexandre Alahi

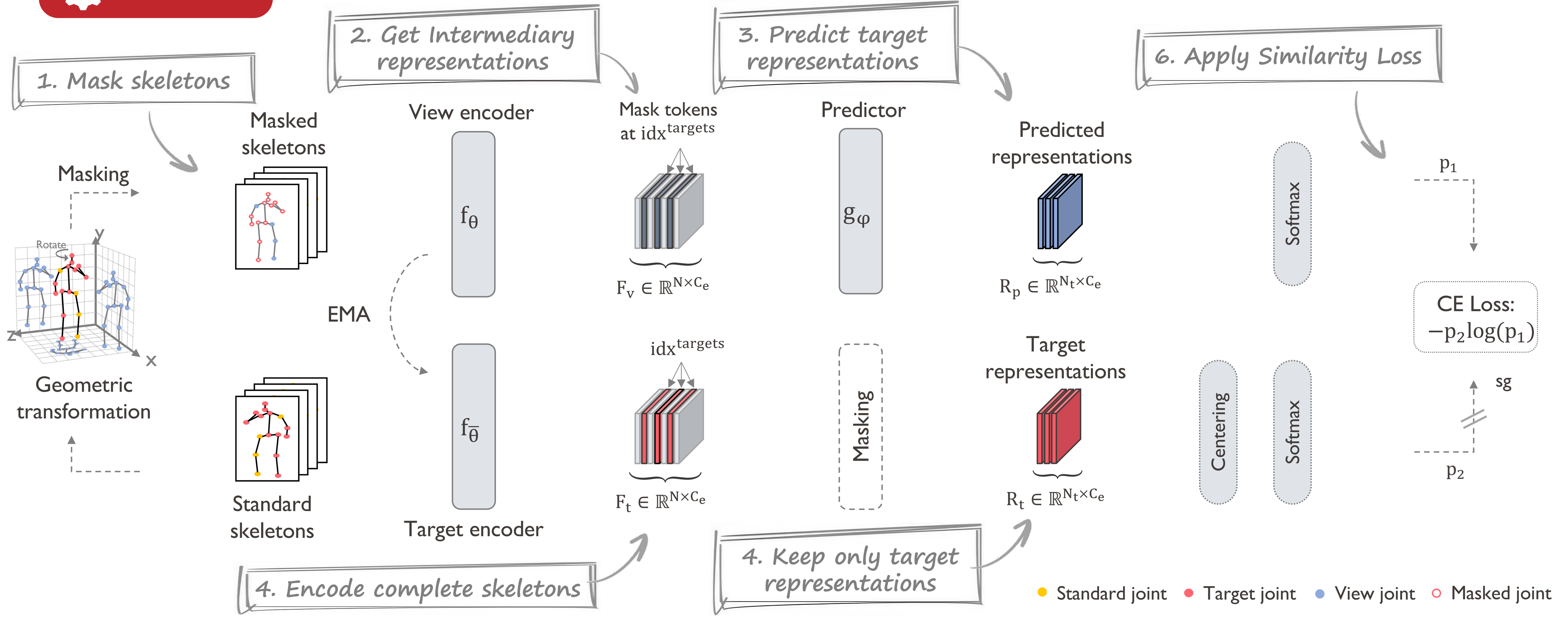
Overview

In one Sentence:
S-JEPA introduces a new pretext task for self-supervised skeletal action recognition based on masked self-reconstruction of latent *features*.

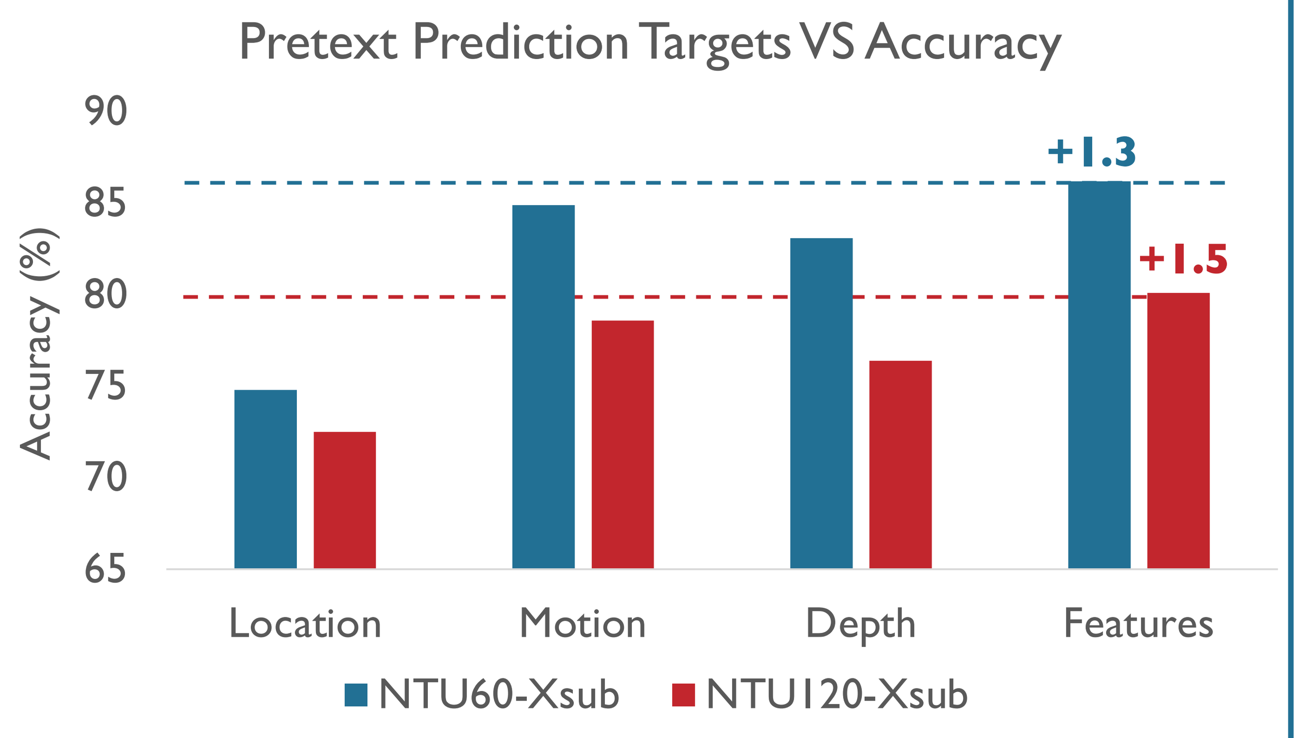




Method



Features VS Joints



Predicting the latent features results in superior linear evaluation performance compared to predicting in the joint space (location, motion, depth).

Linear Evaluation

Method	Input	NTU60 XSub	NTU60 XView	NTU120 XSub	NTU120 XSet
Other pretext task:					
LongTGAN	J	39.1	48.1	-	-
P&C	J	50.7	75.3	42.7	41.7
Contrastive learning:					
CrosSCLR	J+B+M	77.8	83.4	67.9	66.7
AimCLR	J+B+M	78.9	83.8	68.2	68.8
SkeAttnCLR	J+B+M	82.0	86.5	77.1	80.0
CMD	J	79.4	86.9	70.3	71.5
HaLP	J	79.7	86.8	71.1	72.2
ActCLR	J+B+M	84.3	88.8	74.3	75.7
Masked prediction:					
SkeletonMAE	J	74.8	77.7	72.5	73.5
MAMP	J	84.9	89.1	78.6	79.1
S-JEPA	J	**85.3**	**89.8**	**79.6**	**79.9**

Finetuning Evaluation

Method	Backbone	NTU60 XSub	NTU60 XView	NTU120 XSub	NTU120 XSet
Other pretext task:					
Colorization	DGCNN	88.0	94.9	-	-
Hi-TRS	Transformer	90.0	95.7	85.3	87.4
Contrastive learning:					
CPM	ST-GCN	84.8	91.1	78.4	78.9
CrosSCLR	ST-GCN	86.2	92.5	80.5	80.4
AimCLR	ST-GCN	86.9	92.8	80.1	80.9
ActCLR	ST-GCN	88.2	93.9	82.1	84.6
HYSP	ST-GCN	89.1	95.2	84.5	86.3
Masked prediction:					
SkeletonMAE	STRL	92.8	96.5	84.8	85.7
MotionBERT	DSTformer	93.0	97.2	84.8	86.4
MAMP	Transformer	**93.1**	97.5	90.0	**91.3**
S-JEPA	Transformer	**93.1**	**97.6**	**90.3**	**91.3**

Semi-supervised Evaluation

Method	Input	NTU60-XSub 1%	NTU60-XSub 10%	NTU60-XView 1%	NTU60-XView 10%
Other pretext task:					
Colorization	J+B+M	48.3	71.7	52.5	78.9
Hi-TRS	J+B+M	49.3	77.7	51.5	81.1
Contrastive learning:					
ISC	J	35.7	65.9	38.1	72.5
CPM	J	56.7	73.0	57.5	77.1
CrosSCLR	J+B+M	51.1	74.4	50.0	77.8
AimCLR	J+B+M	54.8	78.2	54.3	81.6
CMD	J+B+M	55.6	79.0	55.5	82.4
SkeAttnCLR	J+B+M	59.6	81.5	59.2	83.8
Masked prediction:					
SkeletonMAE	J	54.4	80.6	54.6	83.5
MAMP	J	66.0	88.0	68.7	91.5
S-JEPA	J	**67.5**	**88.4**	**69.1**	**91.4**

mohamed.abdelfattah@epfl.ch
* corresponding author

EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024