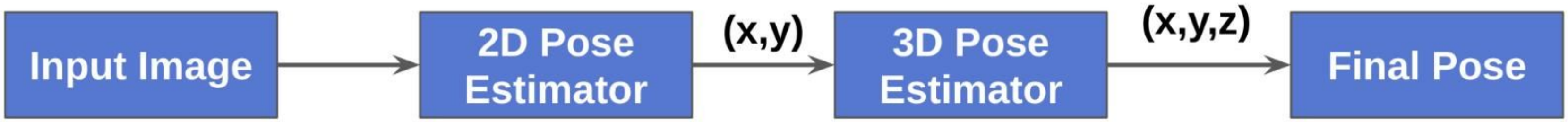
End-to-End Pose Estimation from monocular Ice Hockey Videos

Bavesh Balaji, Jerrin Bright, Harish Prakash
Systems Design Engineering Department
University of Waterloo
Waterloo, Ontario, Canada

MOTIVATION

- Pose estimation helps in better action recognition.
- Can be used to assess performance and strategy.

OVERALL FRAMEWORK



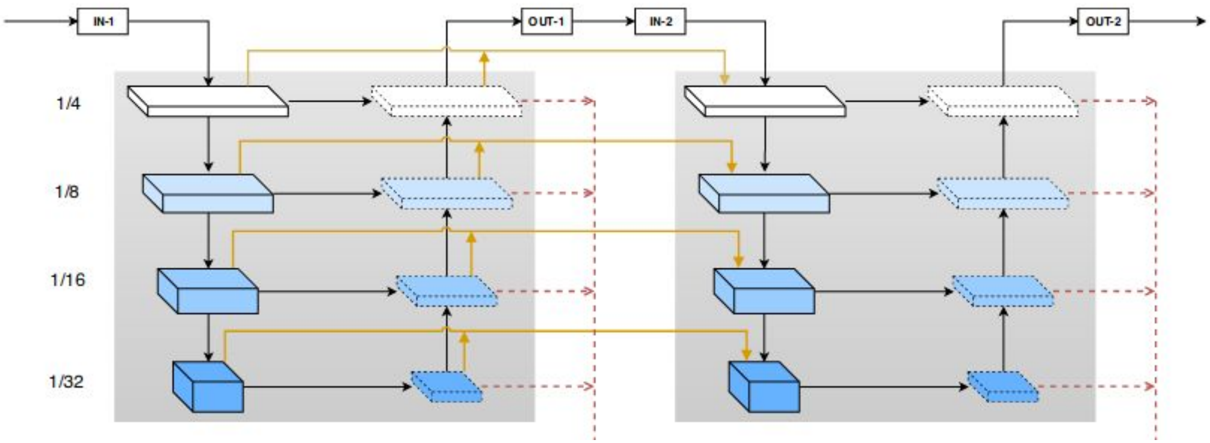
DATA

- Consists of 10 broadcast NHL videos.
- A total of 11,661 frames with ground-truth poses and bbox.

2-D Pose Estimation

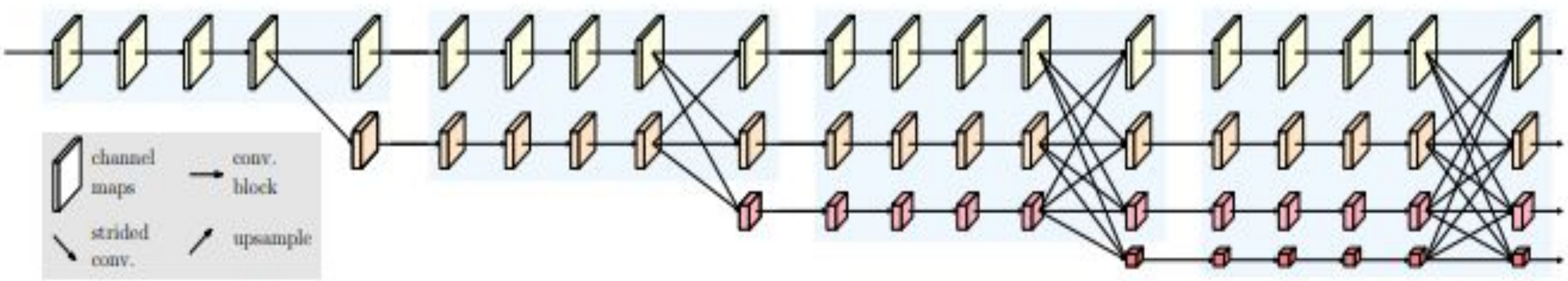
MULTI STAGE POSE NETWORKS

- An encoder-decoder structure following a top-down approach.
- Addresses the design flaws in previous multi stage networks.
  - equal channel width design.
  - cross stage feature aggregation.



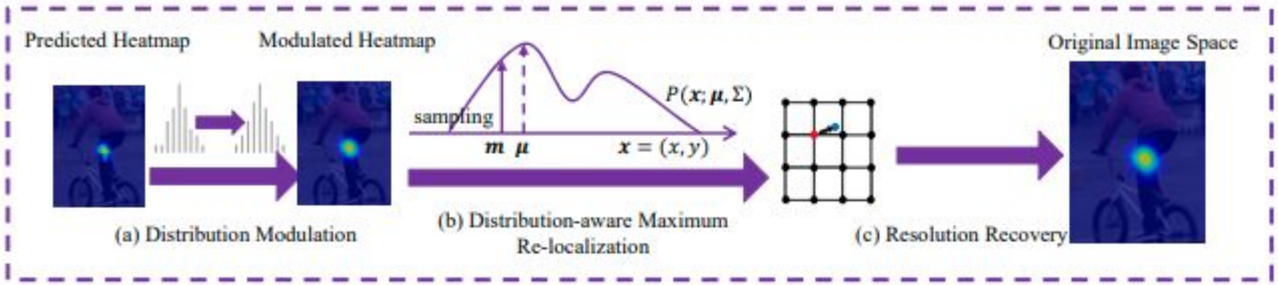
HIGH RESOLUTION NETWORKS

- Multi-stage network
- Parallel connections from high-to-low resolution.
- Repeated multi-scale fusion across parallel convolutions



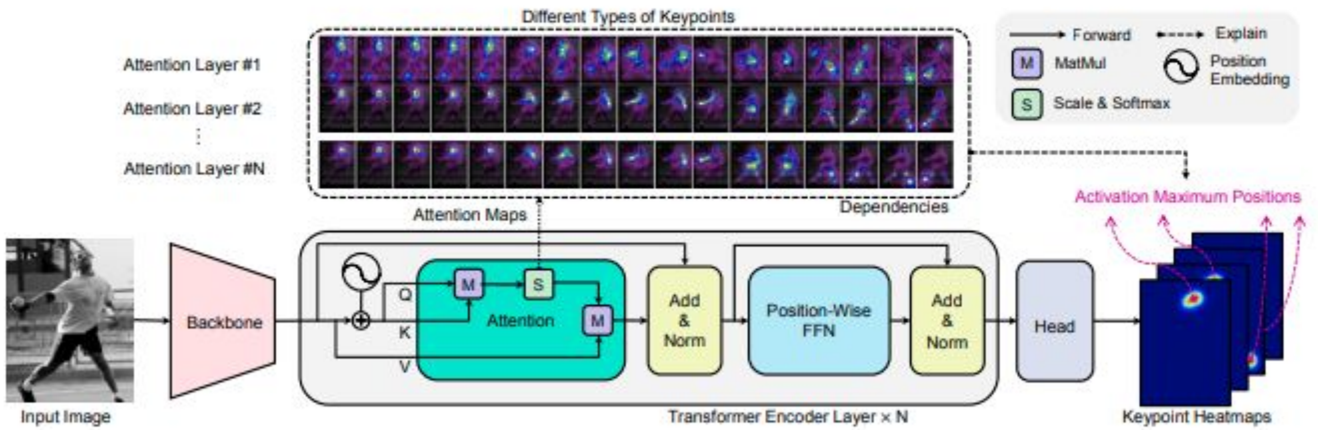
DARK

- Coordinate decoding method that gives us better 2D coordinates from heatmaps.
- Uses the fact that predicted heatmap must be gaussian centred around max.activation.
- Taylor series to approximate differential.



TRANSPOSE-H

- Common CNN backbone used for feature extraction.
- The transformer encoder part provides long-term spatial relationships.



TRAINING DETAILS

COMMON HYPERPARAMETERS:-

- Batch size = 8
- epochs = 10
- Input size = (192, 256)

MSPN:-

- SGD with a learning rate of 1e-2, momentum = 0.9, weight_decay = 1e-5.

OTHER NETWORKS:-

- Adam with a learning rate of 1e-3.

RESULTS

Joint	Training Accuracy(%)				Validation Accuracy(%)			
	MSPN	HRNet-W48	HRNet-W32 + DARK	TransPose	MSPN	HRNet-W48	HRNet-W32 + DARK	TransPose
Left Shoulder	**97.1**	95.27	95.08	91.44	**90.13**	86.13	86.92	85.83
Right Shoulder	**96.63**	95.23	95.07	91.53	**89.96**	86.23	86.98	84.95
Left Elbow	**94.86**	89.97	90.80	86.92	**87.12**	81.07	82.51	81.95
Right Elbow	**94.91**	88.33	89.52	85.53	**89.03**	81.69	85.96	84.00
Left Wrist	**93.34**	87.69	87.13	83.70	**86.13**	80.27	80.38	80.77
Right Wrist	**92.98**	86.11	84.58	81.16	**86.61**	81.18	82.26	82.04
Left Hip	**93.76**	90.52	89.07	84.78	**79.38**	75.69	75.04	74.80
Right Hip	**94.27**	90.72	88.88	84.4	**79.75**	75.63	78.45	79.50
Left Knee	**97.19**	94.38	94.70	92.59	**92.39**	89.54	89.25	88.27
Right Knee	**97.28**	94.67	94.78	92.15	**92.76**	89.35	90.66	80.81
Left Ankle	**97.21**	93.89	93.02	91.23	**92.01**	87.00	86.51	86.65
Right Ankle	**97.40**	94.18	93.92	91.4	**93.70**	87.48	86.19	86.93
Total accuracy	95.71	91.86	91.40	83.75	88.35	83.51	84.32	88.16

VISUALIZATIONS

HRNet-W48

HRNet-W32 + DARK

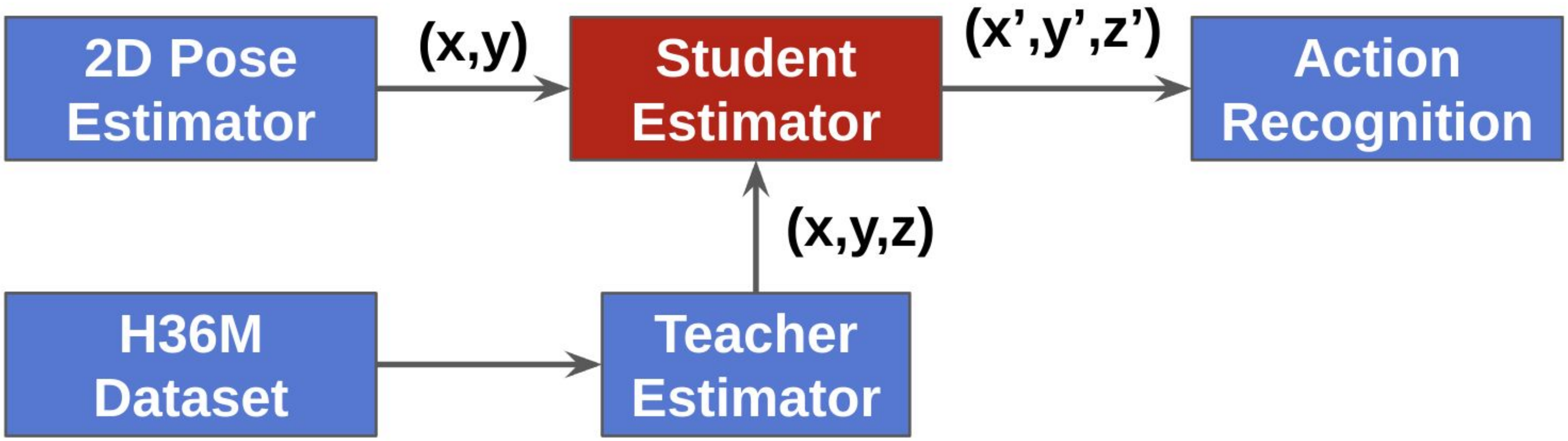
TransPose-H(W4 8)

MSPN

3D Pose Estimation

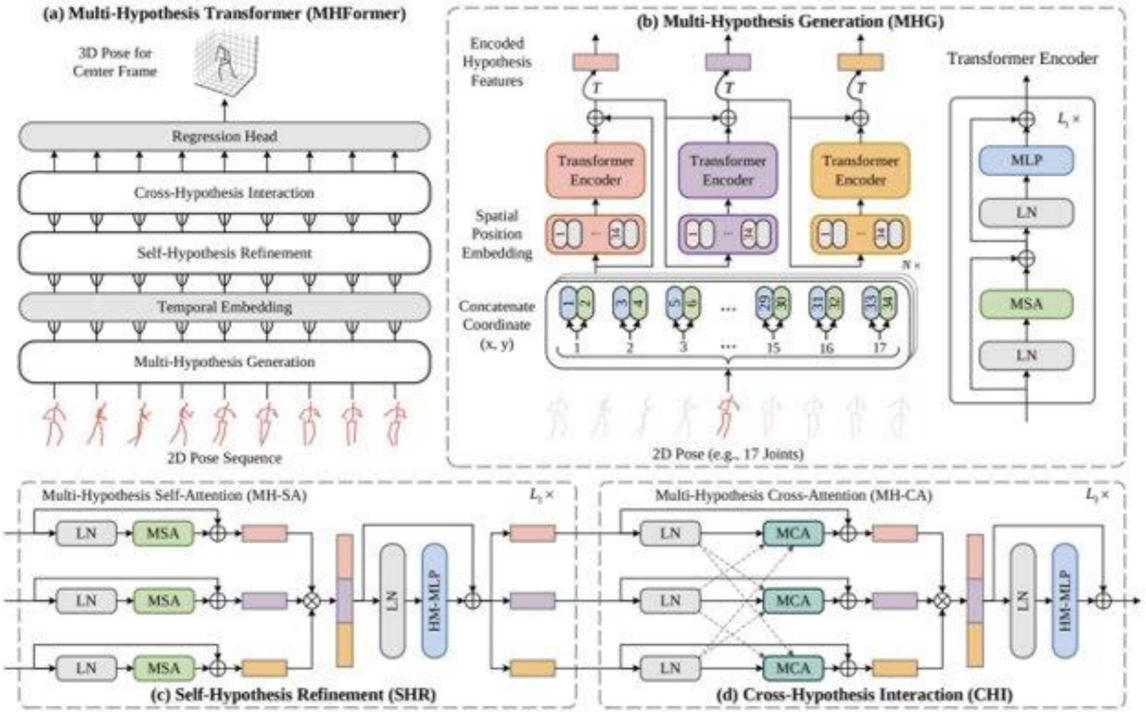
Idea

- Teacher Estimator – Trained on H36M dataset. Estimated 3D pose is feed as input to Student Estimator for training.
- Student Estimator – Trained using 2D pose from hockey data and 3D pose data from the teacher estimator.



Multi-Hypothesis Former [1]

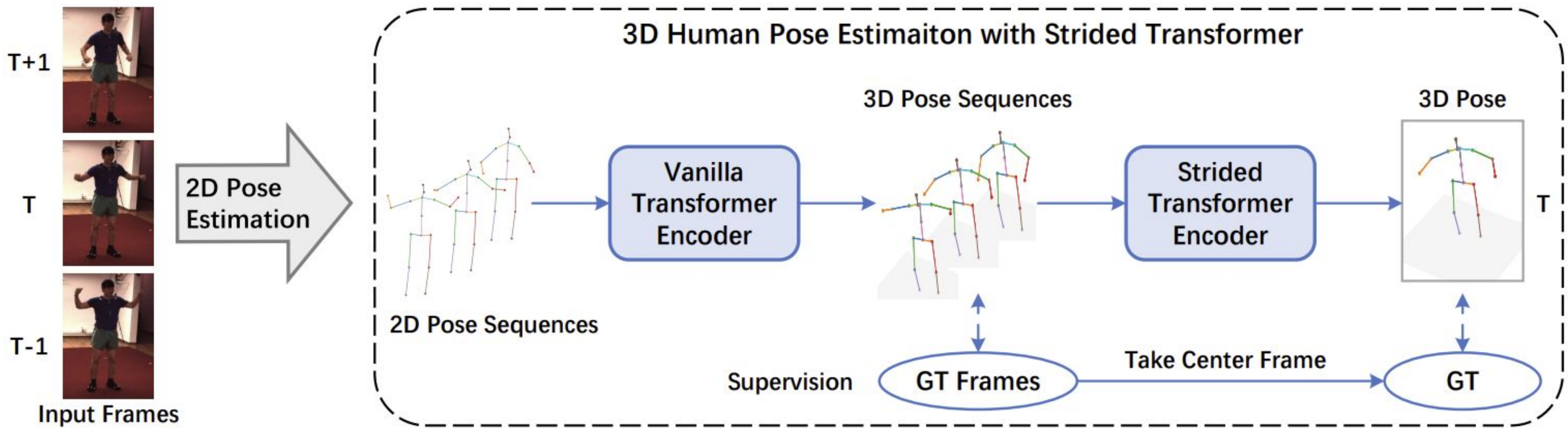
- MHG – Generate Hypotheses
- Temporal Embedding
- SHR – MSA+MLP
- CHI – MCA+MLP



[1] W. Li, H. Liu, H. Tang, P. Wang, and L. V. Gool, "Mhformer: Multi-hypothesis transformer for 3d human pose estimation," CoRR, vol. abs/2111.12707, 2021. [Online].

Strided Transformer Encoder [1]

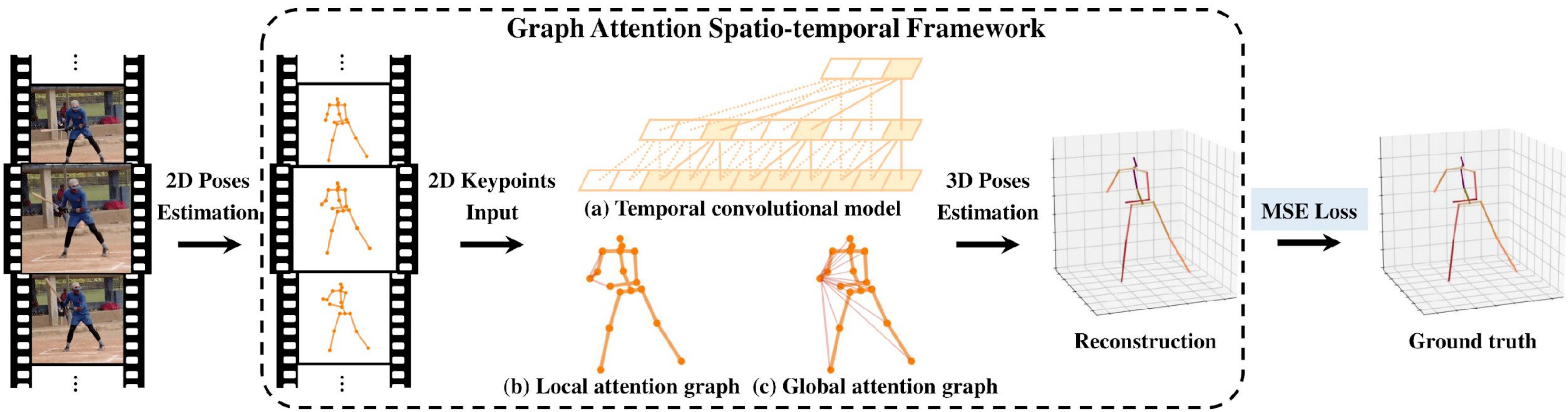
- VTE – Captures global contexts
- STE – Captures local contexts
- Pose Refinement



[1] Li, Wenhao, et al. "Exploiting temporal contexts with strided transformer for 3d human pose estimation." *IEEE Transactions on Multimedia* (2022).

Graph Attention Spatio-Temporal Network [1]

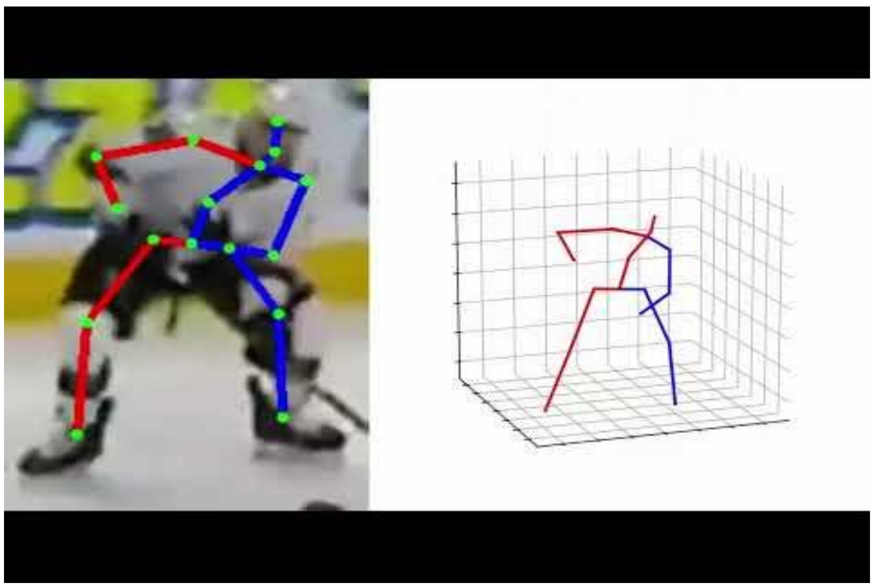
- Dilated Temporal Convolutional Networks
- Local Spatial Attention Graph
- Global Spatial Attention Graph



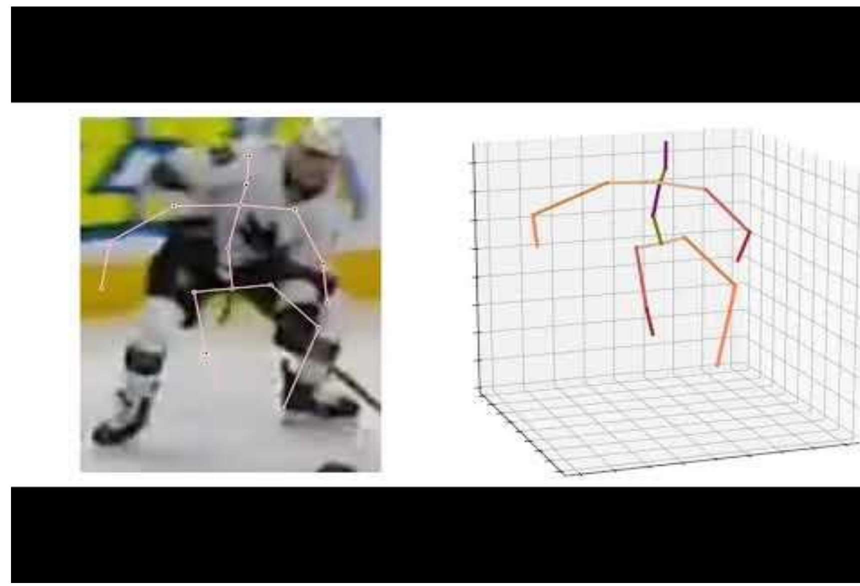
[1] Liu, Junfa, et al. "A graph attention spatio-temporal convolutional network for 3D human pose estimation in video." *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

Results

- MHFormer



- GASTNet



3D Pose Estimation

- Based on our literature survey and experimentations, we observed that transformers (PoseFormer, STE, MHFormer) have an inherent advantage in processing time (FLOPs), while CNNs (VideoPose3D, GAST-Net) exhibit a better accuracy under similar constraints.
- This can be attributed to the ‘<u>Inductive Bias</u>’ in CNNs – extensive correlations between feature spaces that help our target function learn a set of assumptions associated with the input and unseen output, intrinsically.
- Transformers lack this => Requires extensive feature engineering;

  CNNs => Generalizes well (relatively) w/o it.

“ Why not induce ‘**inductive bias**’ into transformer? ”

<u>Base Paper:</u> Published in IEEE Transactions on Multimedia, in 2022.

Existing Network Architecture

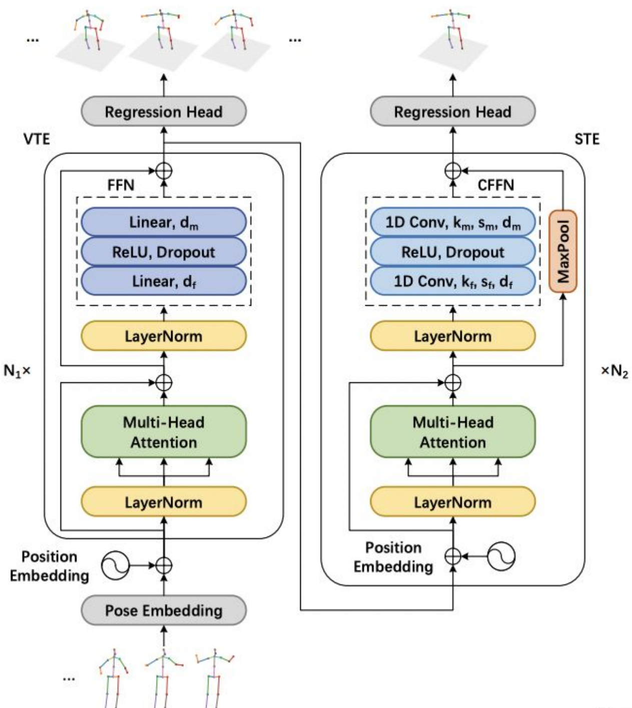


Fig 1. Existing VTE (left) and STE (right) architectures. Each block is x3 times, followed by a regression head at each stage to capture long-range information + aggregate local contexts.

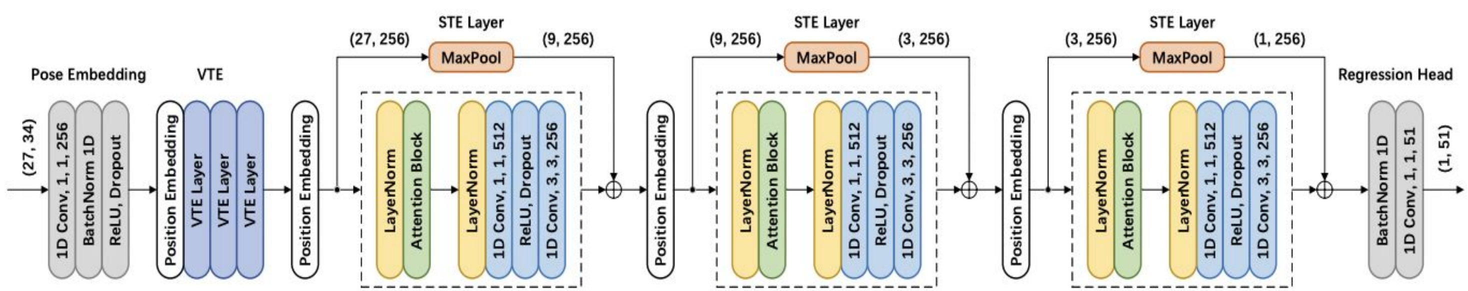
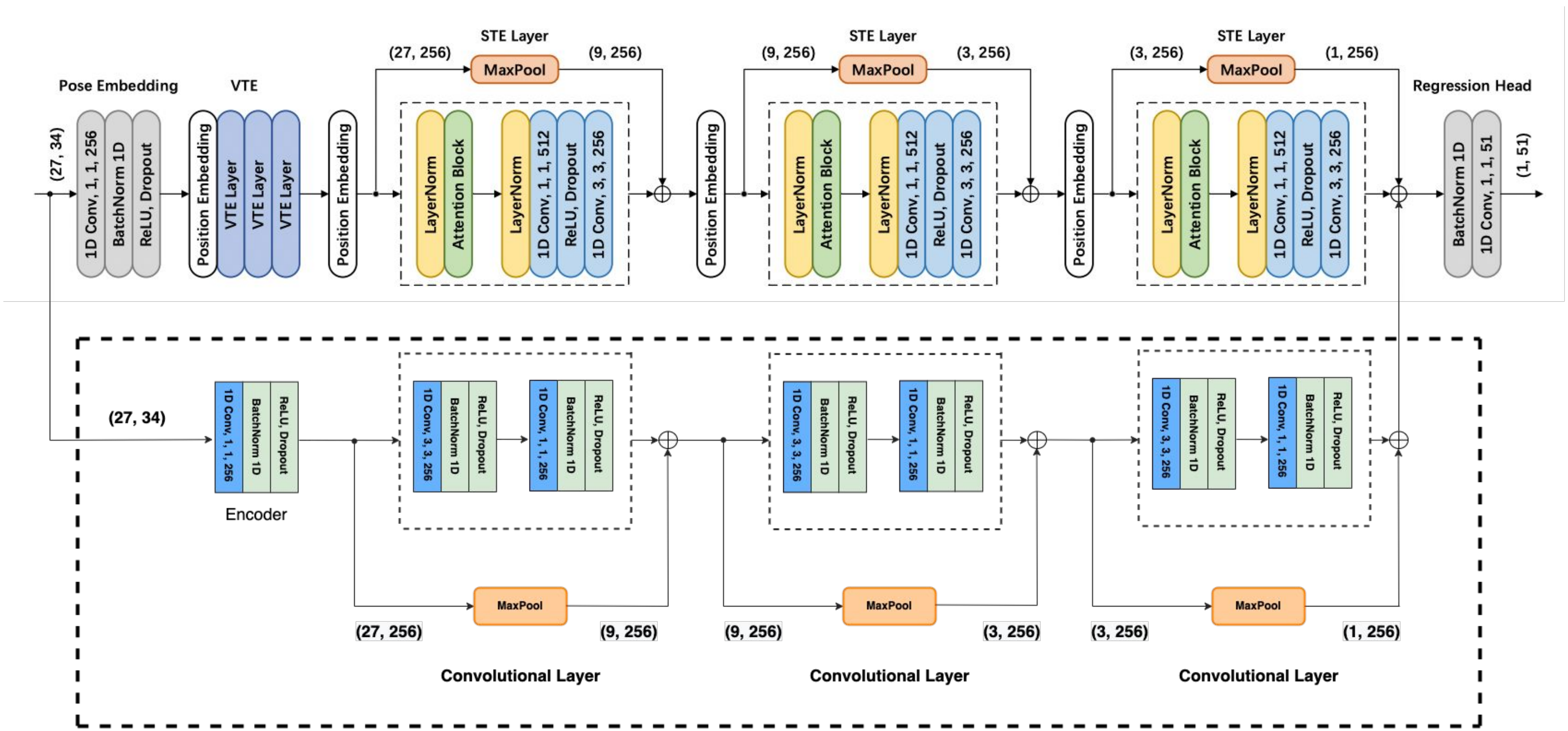


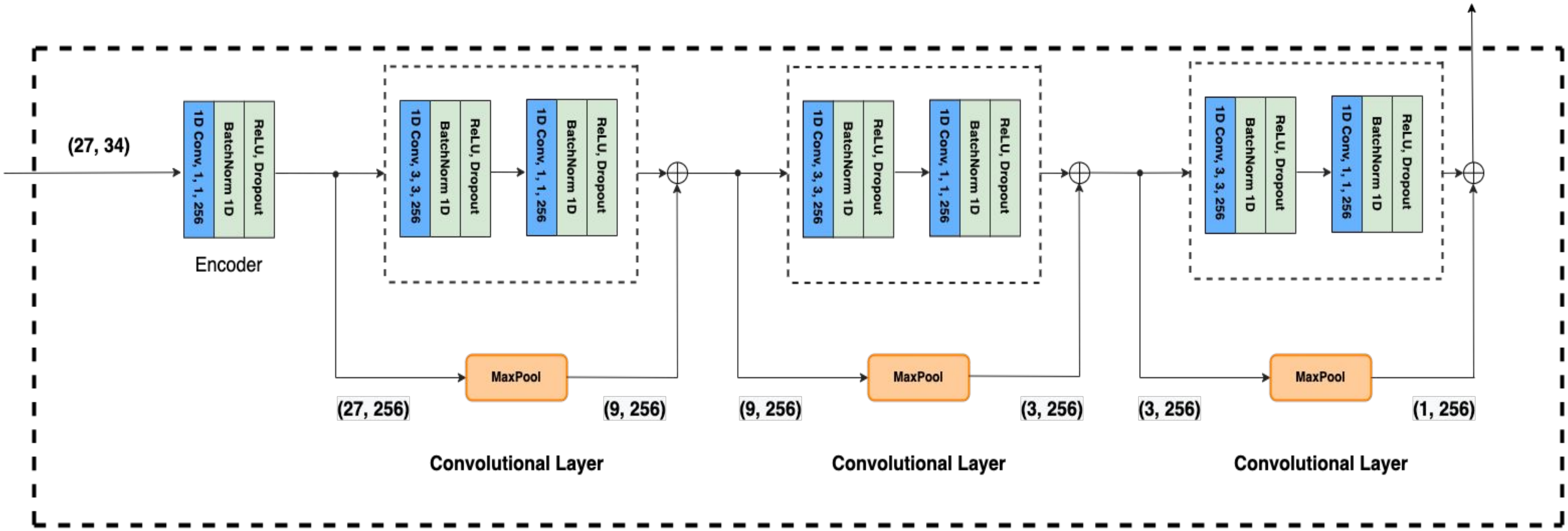
Fig 2. Detailed view of STE architecture.

Proposed Network Architecture



Introducing IB into Transformer

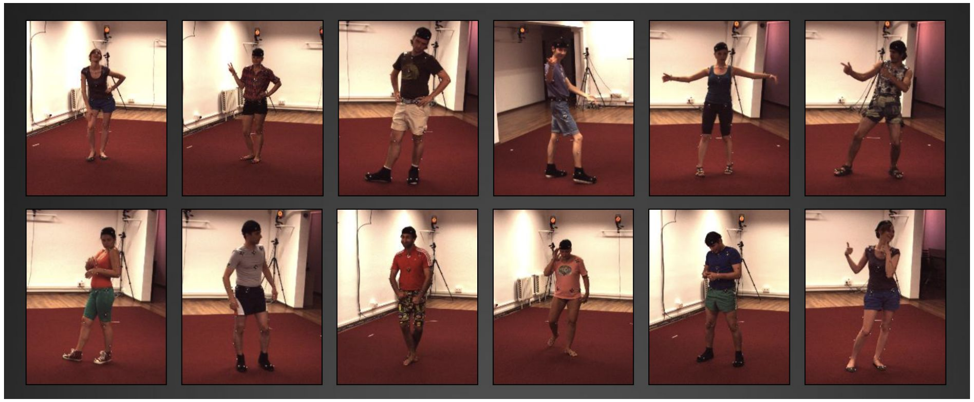
Proposed Network Architecture



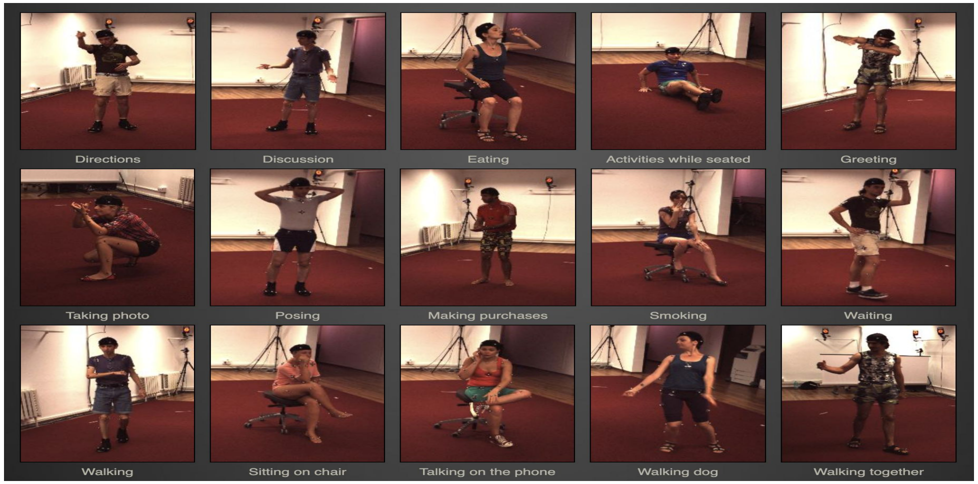
An instantiation of our proposed Temporal-convolution 3D Pose Estimation block. The input consists of 2D Keypoints with a receptive field of 27 frames with $J = 17$ Joints. The tensor (27, 34) denotes 27 frames and 34 channels. Each block has two 1D convolutional layers, where (3, 3, 256) denotes kernel size = 3, stride factor = 3 and output channels = 256, respectively. Every block also has a skip connection, which contains a MaxPooling1D + BatchNorm1D layer, to match the output tensors of Strided Convolutions.

Dataset

- We trained our proposed network on the Human3.6M Mocap dataset, with the same training scheme as STE, but with hyperparameter tuning. We choose this dataset since it's the de-facto for 3D PE tasks + useful for our model's relative performance measure.



It consists of 11 Subjects (S1-S11), whose actions are recorded using 4 calibrated cameras in an indoor, constrained setting.

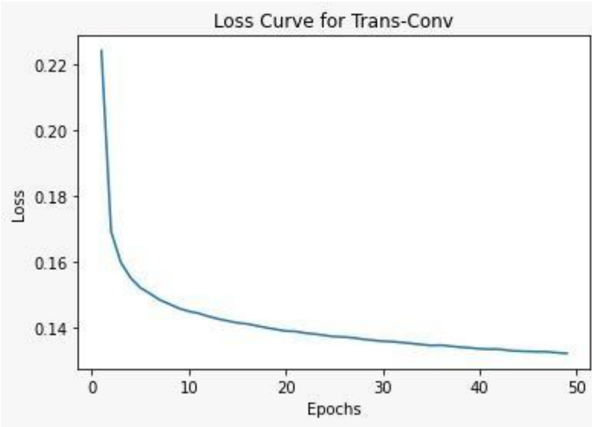


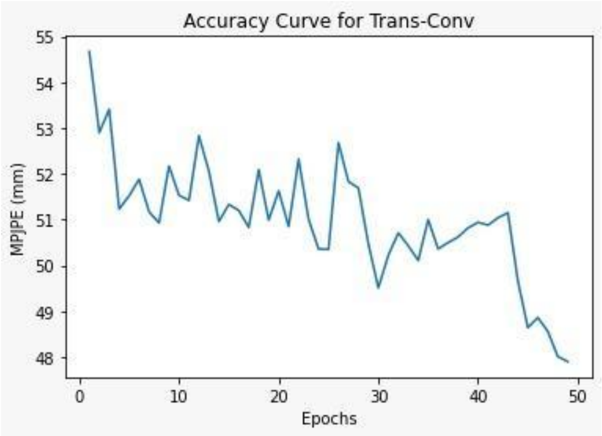
For all 11 subjects, there are sequences of 15 actions captured, recorded at 50Hz.

Quantitative Results

- We present results based on "Mean Per Joint Position Error (MPJPE)", widely called Protocol#1 in Pose Estimation Literature.
- Training: 3,119,616 frames;
- Validation/Testing: 543,488 frames;
- Optimizer: AdamW with Amsgrad;
- Loss: Average MPJPE (VTE + STE + TCN)

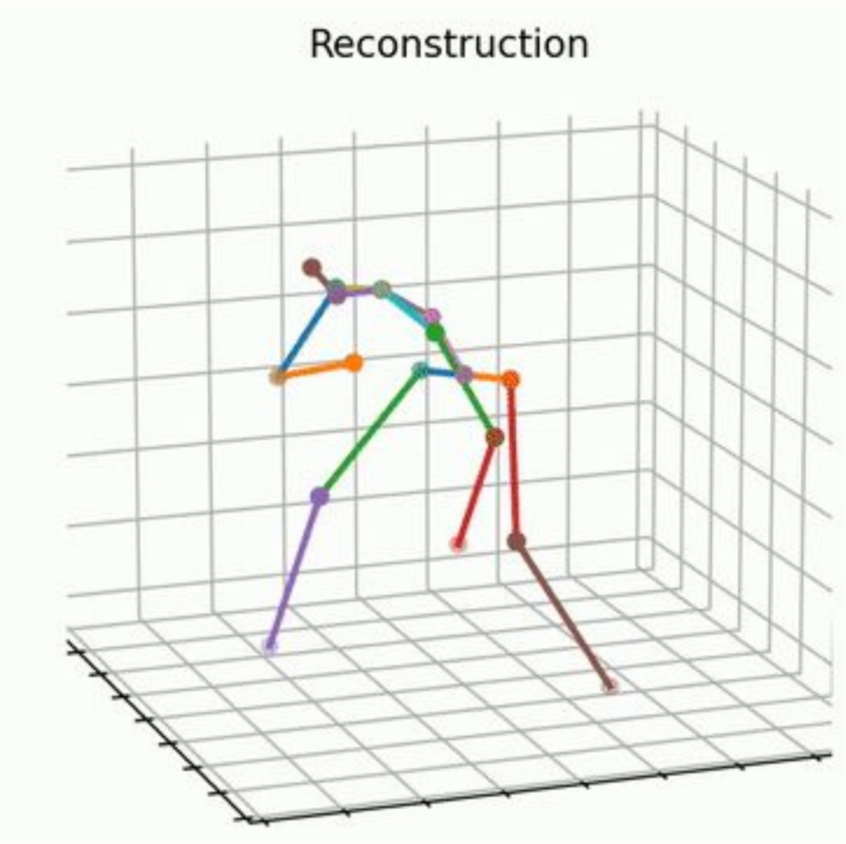
Model	**Receptive Field**	**MPJPE (mm)**
Hossain et al. [2]	27 frames	58.3
Cai et al. [3]	27 frames	48.8
Pavllo et al. [4]	27 frames	48.6
Chen et al. [5]	27 frames	48.3
Wenhao et al. [1]	27 frames	46.9
Ours (Trans-Conv)	27 frames	**47.9**





Qualitative Results

Inference on an Ice-hockey Sequence



References

[1] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. *Exploiting temporal contexts with strided transformer for 3D human pose estimation*. IEEE Transactions on Multimedia, 2022.

[2] M. Rayat Imtiaz Hossain and J. J. Little. *Exploiting temporal information for 3d human pose estimation.* Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 68–84.

[3] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann. *Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks.* Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2272–228

[4] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. *3D human pose estimation in video with temporal convolutions and semi-supervised training.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7753–7762.

[5] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. *Cascaded pyramid network for multi-person pose estimation.* Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7103–71