Seminar/Project: Computer Vision and Deep Learning

S3: 3D Pose Estimation using Transformers

Ayush Chamoli
chamoli@rhrk.uni-kl.de
Supervisor: Ahmed Tawfik Aboukhadra

Do we really need to estimate pose?

- Kinect for Xbox 360: A revolution of its time.
- Paved a way for future of gaming and immersion.
- But the applications of pose estimation is far beyond…
- What’s the current trend though?

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/real-time-human-pose-recognition-in-parts-from-single-depth-images-1.png

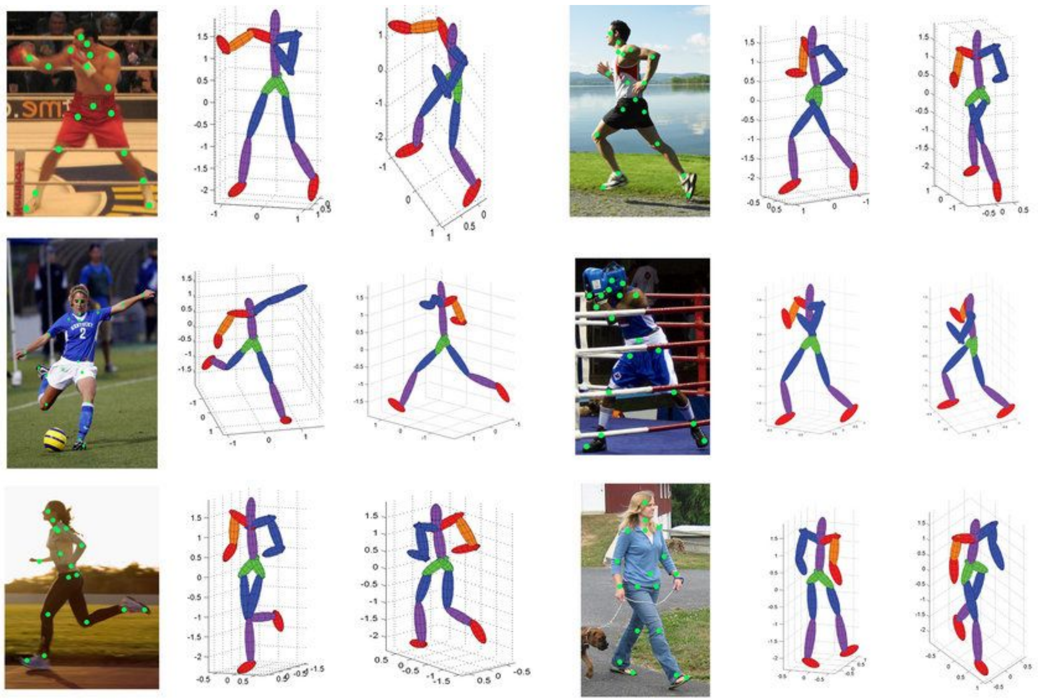
Do we really need to estimate pose?

- Currently used for action recognition, sport motion analysis and
- Also helps robots in providing medical assistance and rehabilitation.
- Hand pose detection is also really common now.

https://developer.nvidia.com/blog/training-optimizing-2d-pose-estimation-model-with-tao-toolkit-part-1/

3D Pose Estimation

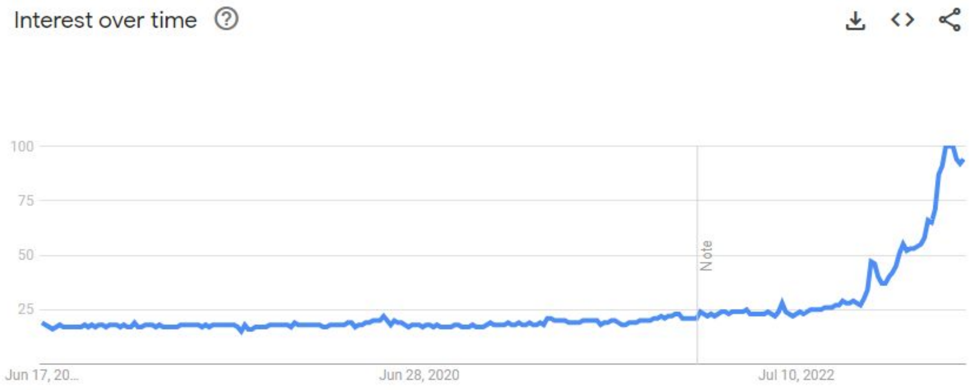
- Two common approaches to detect a 3D pose.
- Common methods exploit information in spatial domain or temporal domain or both.
- Convolution based and graph based architecture are commonly used for Pose Estimation.



https://link.springer.com/chapter/10.1007/978-3-319-10590-1_12

Transformers

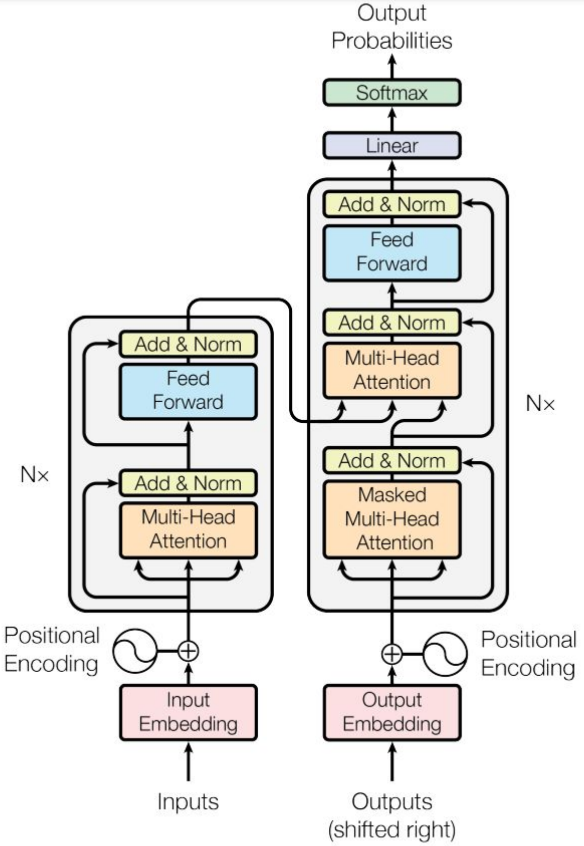
- The “buzzword” for ML Enthusiasts of today.
- “Attention is All You Need” walked so ChatGPT could run.
- Can be applied to things outside Natural Language.



https://trends.google.com/trends/explore?date=today%205-y&q=AI&hl=en-US
https://arxiv.org/pdf/1706.03762.pdf

Transformers

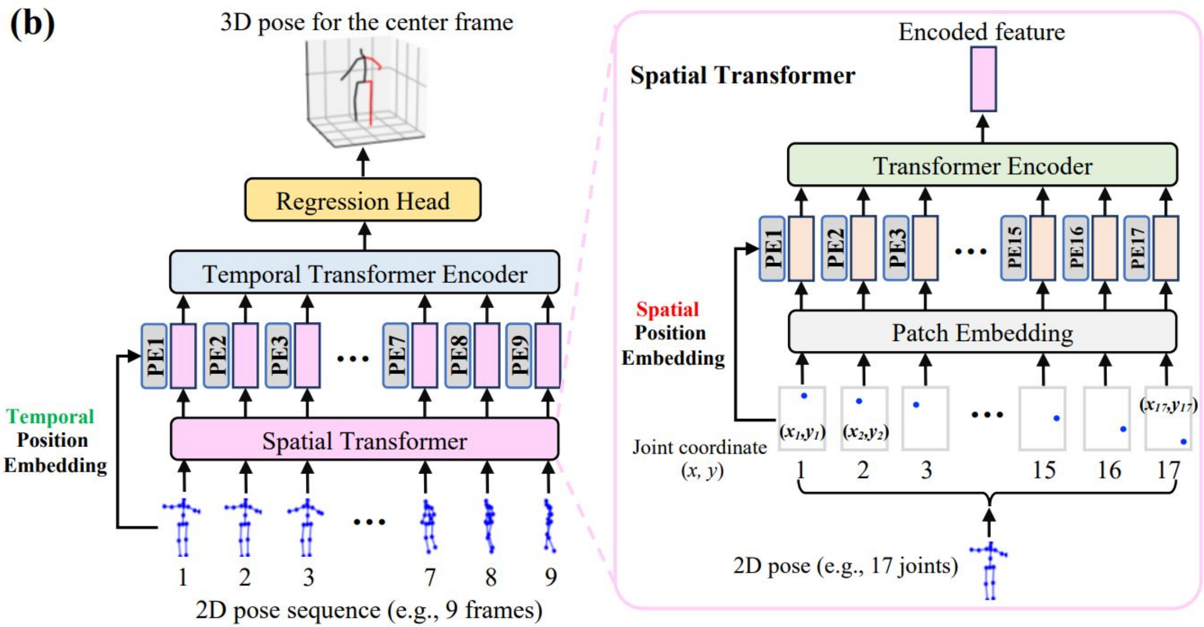
- Uses self attention mechanism to capture long-range dependencies.
- Can be parallelized with multi-head self attention.
- Gives state of the art performance in CV tasks



https://arxiv.org/pdf/1706.03762.pdf

PoseFormer

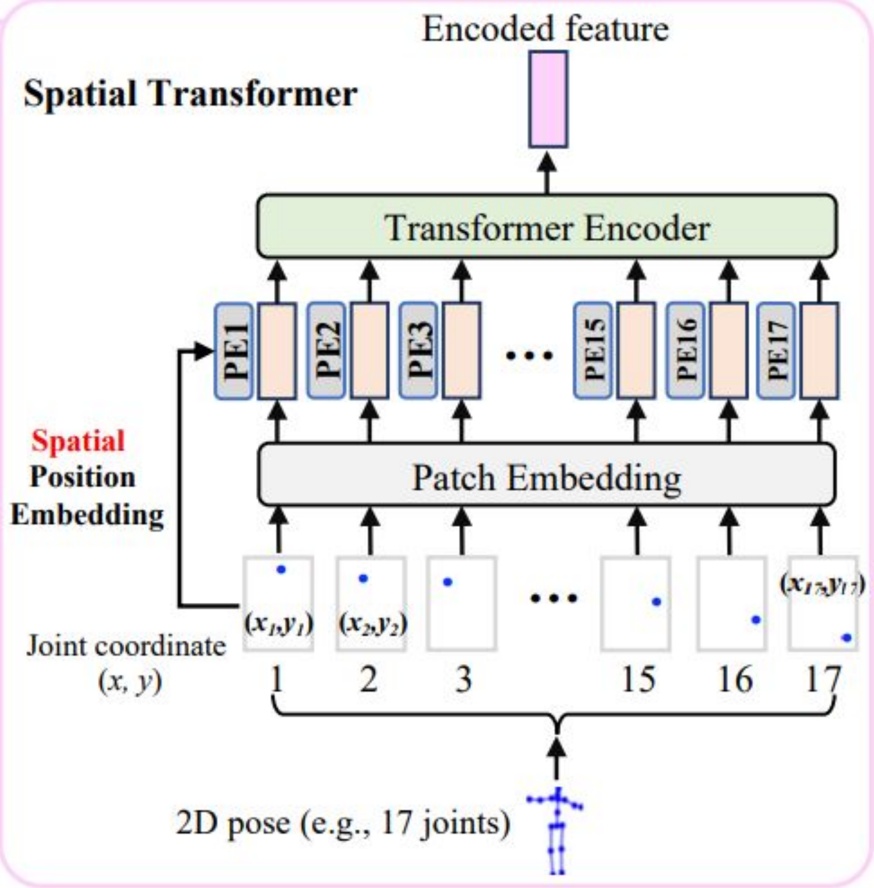
- Lifts human pose from 2D to 3D in videos.
- Models the spatial and temporal aspects with distinct transformer module.



https://openaccess.thecvf.com/content/ICCV2021/papers/Zheng_3D_Human_Pose_Estimation_With_Spatial_and_Temporal_Transformers_ICCV_2021_paper.pdf

PoseFormer

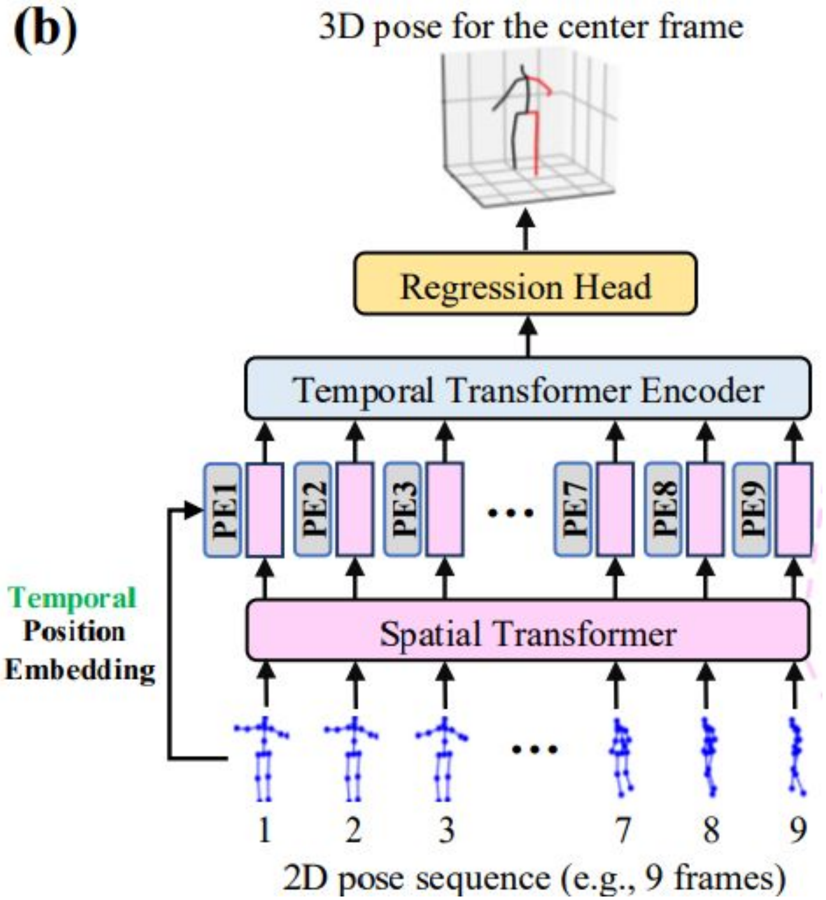
- Spatial Module extracts the high dimensional feature embedding of each frame.
- The coordinates of the joints are projected to a higher dimension and spatial embeddings are added.
- These features are put through a spatial transformer with self attention over all joints.



https://openaccess.thecvf.com/content/ICCV2021/papers/Zheng_3D_Human_Pose_Estimation_With_Spatial_and_Temporal_Transformers_ICCV_2021_paper.pdf

PoseFormer

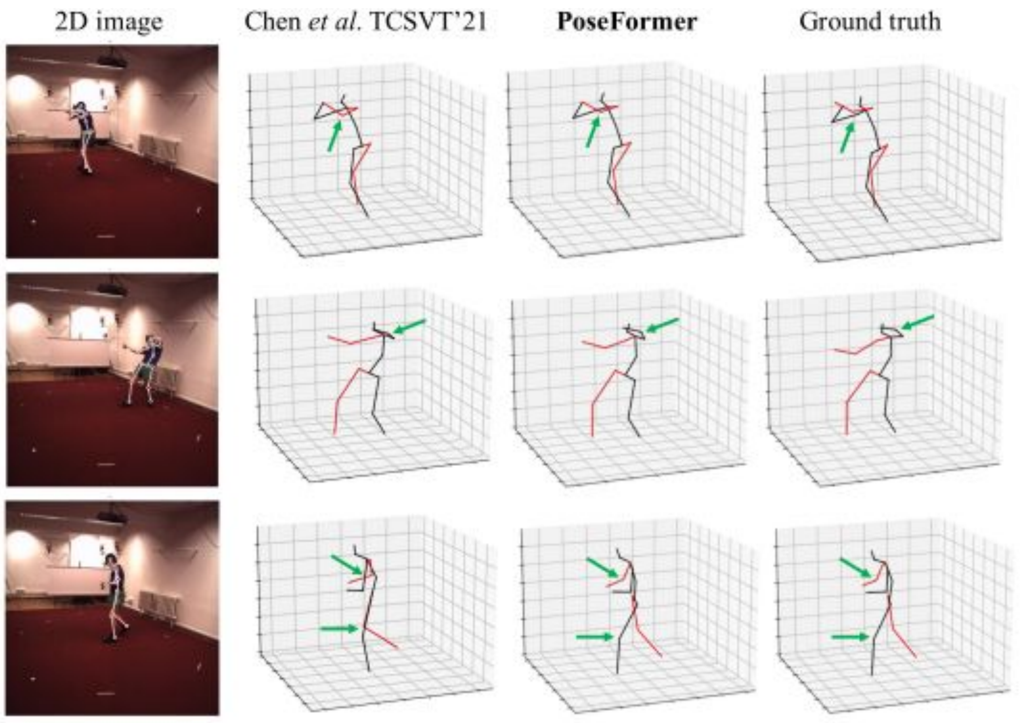
- Temporal Module extracts dependencies across frames.
- Temporal positional embeddings are added.
- Passed through temporal transformer with multihead self attention.



https://openaccess.thecvf.com/content/ICCV2021/papers/Zheng_3D_Human_Pose_Estimation_With_Spatial_and_Temporal_Transformers_ICCV_2021_paper.pdf

PoseFormer

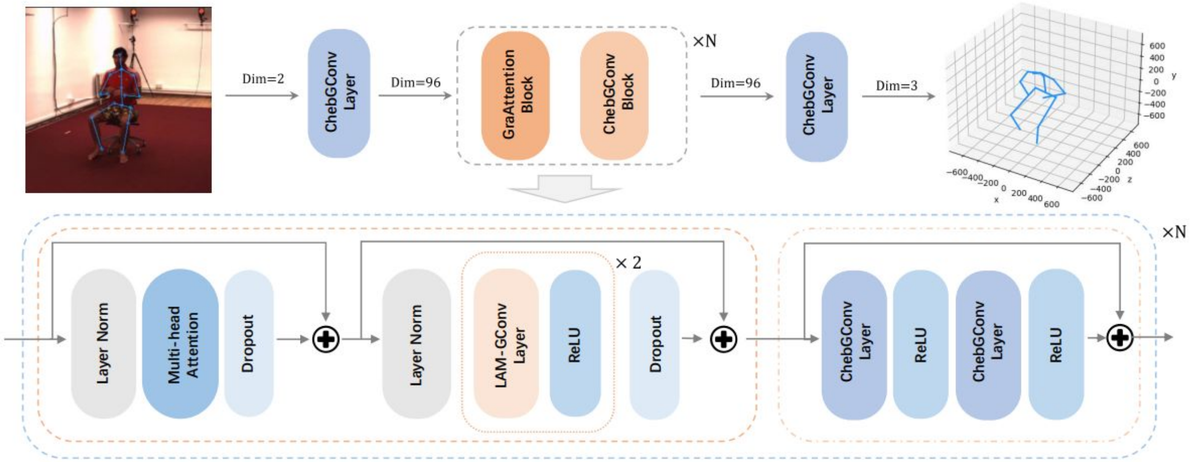
- Regression head is used on the output of temporal transformer.
- Weighted mean operation with learnable parameters is used.
- MPJPE is used as an error metric.



https://openaccess.thecvf.com/content/ICCV2021/papers/Zheng_3D_Human_Pose_Estimation_With_Spatial_and_Temporal_Transformers_ICCV_2021_paper.pdf

GraFormer

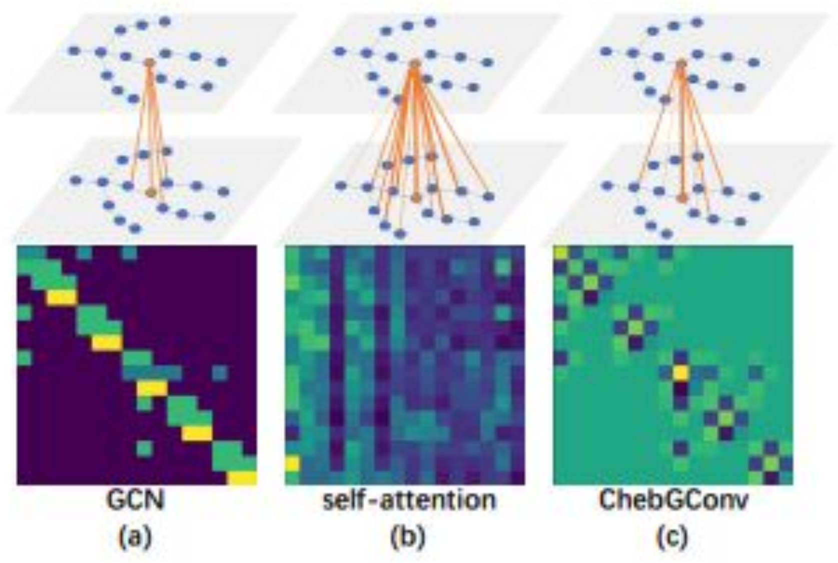
- Proposed a new architecture by stacking GraAttention and ChebGConv block.
- Model the best part of implicit and explicit relationship between nodes.
- Improves performance by enlarging receptive field of node information transmission.



https://openaccess.thecvf.com/content/CVPR2022/papers/Zhao_GraFormer_Graph-Oriented_Transformer_for_3D_Pose_Estimation_CVPR_2022_paper.pdf

GraFormer

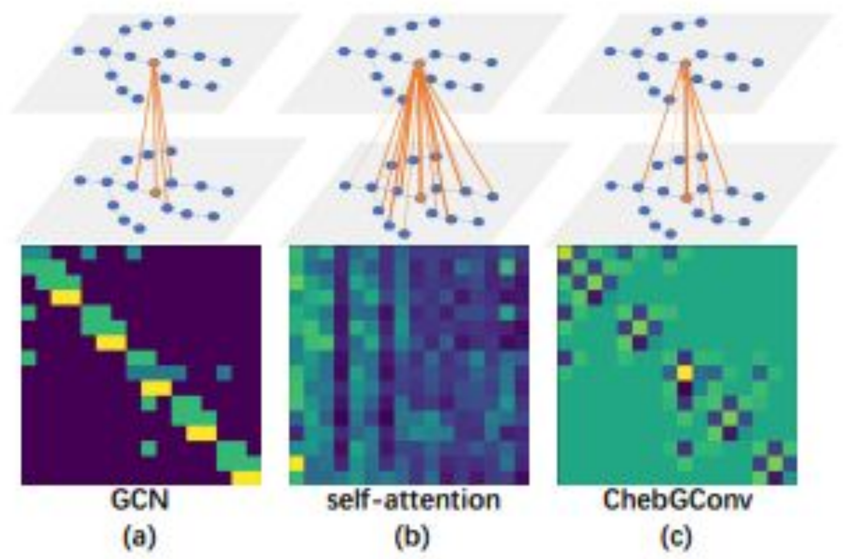
- GraAttention is a combination of multihead self-attention block and GCN layer.
- Includes a dropout layer for regularization of self-attention output.
- Each element of output of multihead attention block contains information of all 2D joints.



https://openaccess.thecvf.com/content/CVPR2022/papers/Zhao_GraFormer_Graph-Oriented_Transformer_for_3D_Pose_Estimation_CVPR_2022_paper.pdf

GraFormer

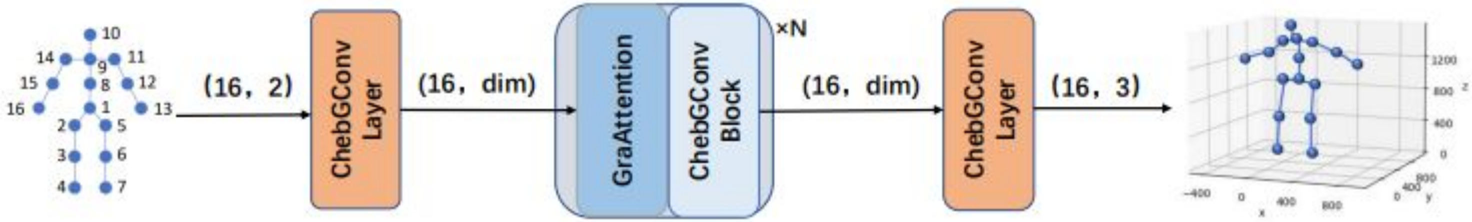
- Chebyshev graph convolution (aka ChebGConv) is used for the graph convolution operation.
- ChebGConv is more powerful compared with traditional GCN layers.
- Boosts the performance by fusing the information among the top K neighbors of a joint.



https://openaccess.thecvf.com/content/CVPR2022/papers/Zhao_GraFormer_Graph-Oriented_Transformer_for_3D_Pose_Estimation_CVPR_2022_paper.pdf

GraFormer

- 2D joint coordinates are preprocessed by ChebGConv layer.
- They are passed through a stack of GraAttention and ChebGConv blocks.
- GraAttention block exploits the global dependencies of joints whereas ChebGConv block exploits the local dependencies among the joints.
- MPJPE is used as ae error metric.



https://openaccess.thecvf.com/content/CVPR2022/papers/Zhao_GraFormer_Graph-Oriented_Transformer_for_3D_Pose_Estimation_CVPR_2022_paper.pdf

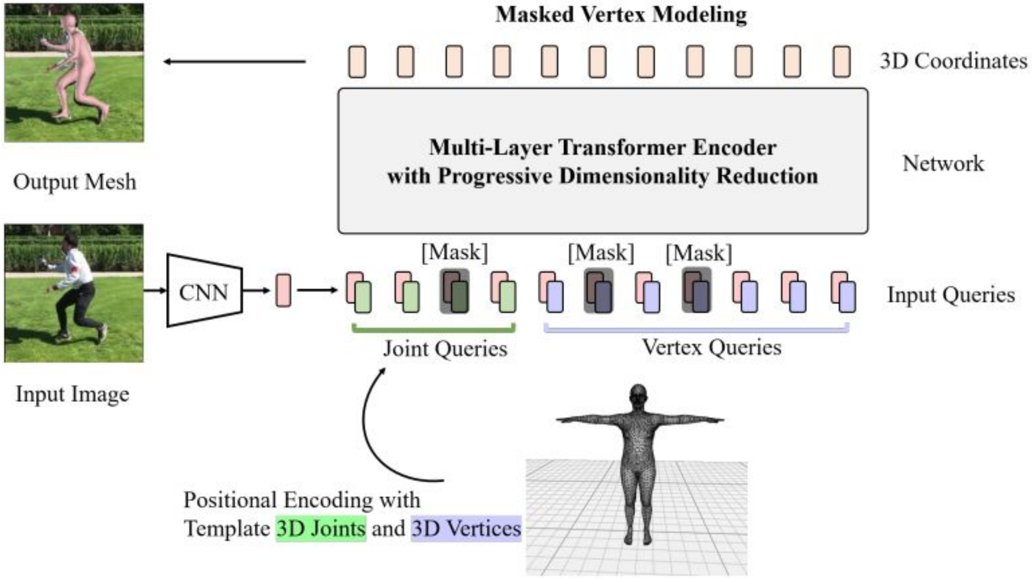
MEsh TRansfOrmer (METRO)

- Proposed a new transformer-based method aka METRO.
- Reconstructs 3D human pose and mesh vertices from a single image.
- Is versatile and can be used to predict different type of 3D mesh (3D Hand etc).

https://openaccess.thecvf.com/content/CVPR2021/papers/Lin_End-to-End_Human_Pose_and_Mesh_Reconstruction_with_Transformers_CVPR_2021_paper.pdf

MEsh TRansfOrmer (METRO)

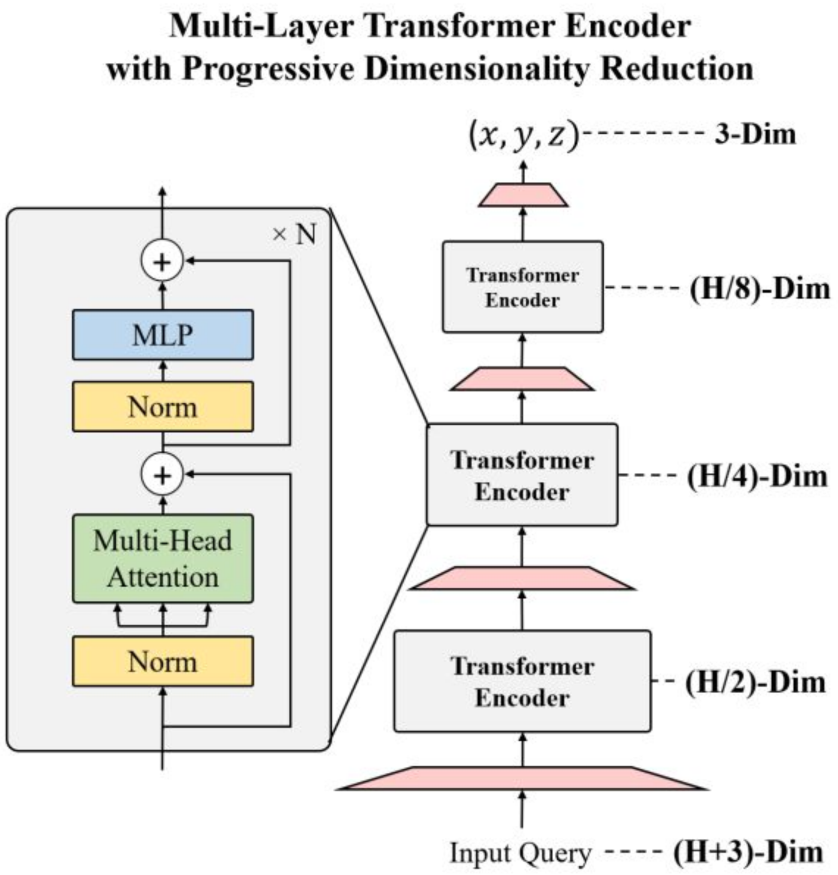
- Convolutional neural network block extracts features.
- It is pretrained on ImageNet classification task.
- Feature vector from the last hidden layer is taken as input for transformer block.



https://openaccess.thecvf.com/content/CVPR2021/papers/Lin_End-to-End_Human_Pose_and_Mesh_Reconstruction_with_Transformers_CVPR_2021_paper.pdf

MEsh TRansfOrmer (METRO)

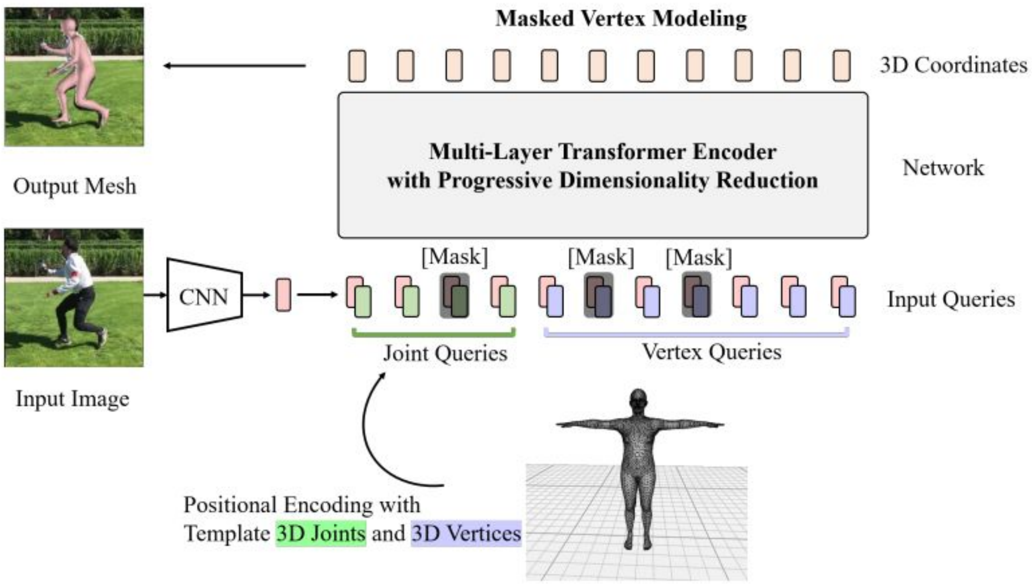
- Multi-Layer Transformer Encoder is a transformer encoder with several modifications.
- It performs gradual dimensionality reduction followed by transformer encoder block.
- Dimensionality reduction is necessary in order to obtain 3D coordinates output.



https://openaccess.thecvf.com/content/CVPR2021/papers/Lin_End-to-End_Human_Pose_and_Mesh_Reconstruction_with_Transformers_CVPR_2021_paper.pdf

MEsh TRansfOrmer (METRO)

- Masked Vertex Modeling (MVM) is used for regression task.
- It regress 3D coordinates by taking other vertices and joints in consideration.
- Helps in learning local and global interaction of each joint.
- MPJPE is used as an error metric.



https://openaccess.thecvf.com/content/CVPR2021/papers/Lin_End-to-End_Human_Pose_and_Mesh_Reconstruction_with_Transformers_CVPR_2021_paper.pdf

HandOccNet

- Proposed a framework for occlusion-robust 3D Hand Mesh estimation.
- Uses two transformer based modules, FIT and SET.

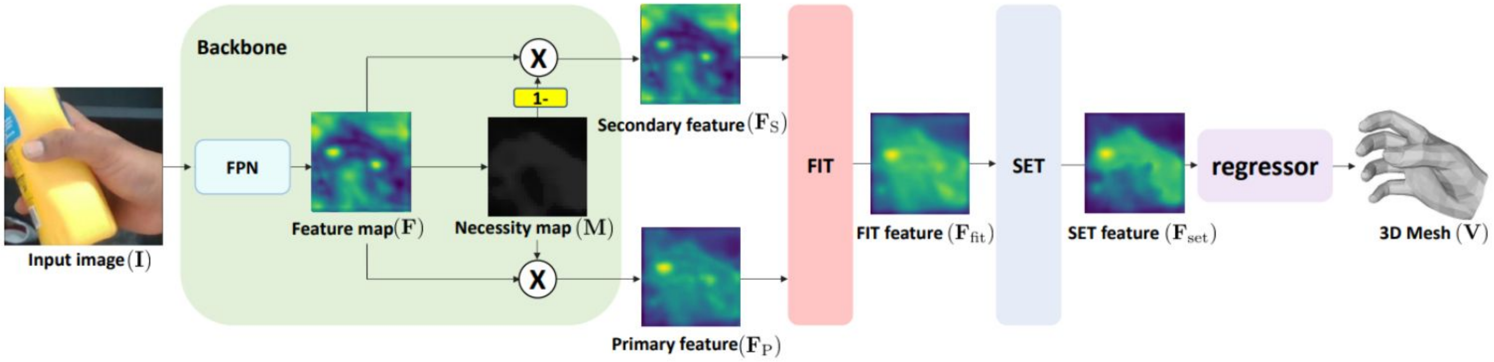
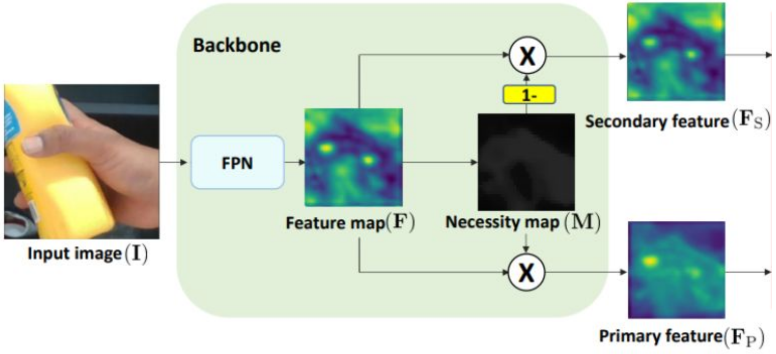


Fig. 1: Architecture of HandOccNet

https://openaccess.thecvf.com/content/CVPR2022/papers/Park_HandOccNet_Occlusion-Robust_3D_Hand_Mesh_Estimation_Network_CVPR_2022_paper.pdf

HandOccNet

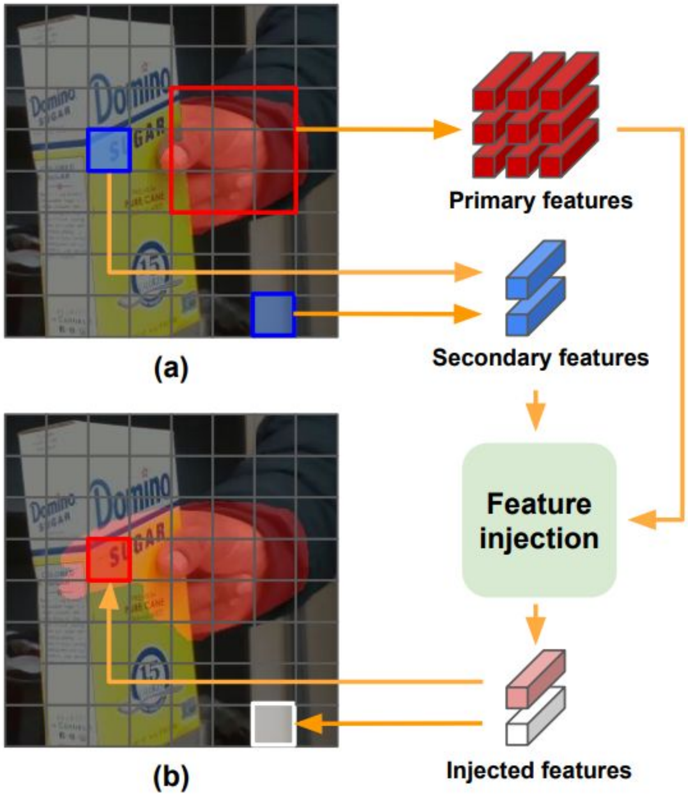
- Hand image is passed through a ResNet50 based Feature Pyramid Network.
- A Feature Map and a Necessity map is obtained.
- These are used to extract Primary and Secondary features.



https://openaccess.thecvf.com/content/CVPR2022/papers/Park_HandOccNet_Occlusion-Robust_3D_Hand_Mesh_Estimation_Network_CVPR_2022_paper.pdf

HandOccNet

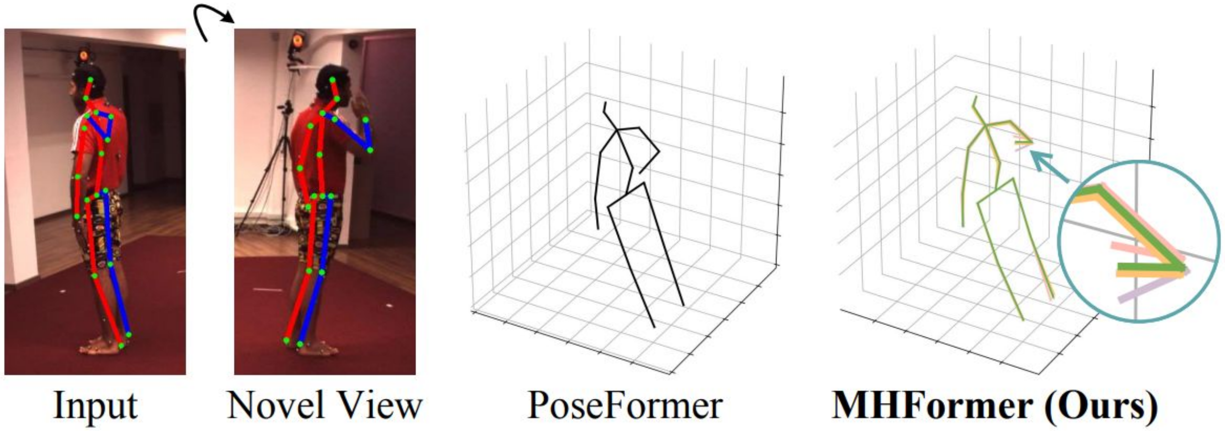
- Primary features describe information of region of hand.
- Secondary feature describe information of occluded region.
- FIT injects information into occluded region.
- SET refines the output of FIT.
- Regressor is used to estimate 3D pose.



https://openaccess.thecvf.com/content/CVPR2022/papers/Park_HandOccNet_Occlusion-Robust_3D_Hand_Mesh_Estimation_Network_CVPR_2022_paper.pdf

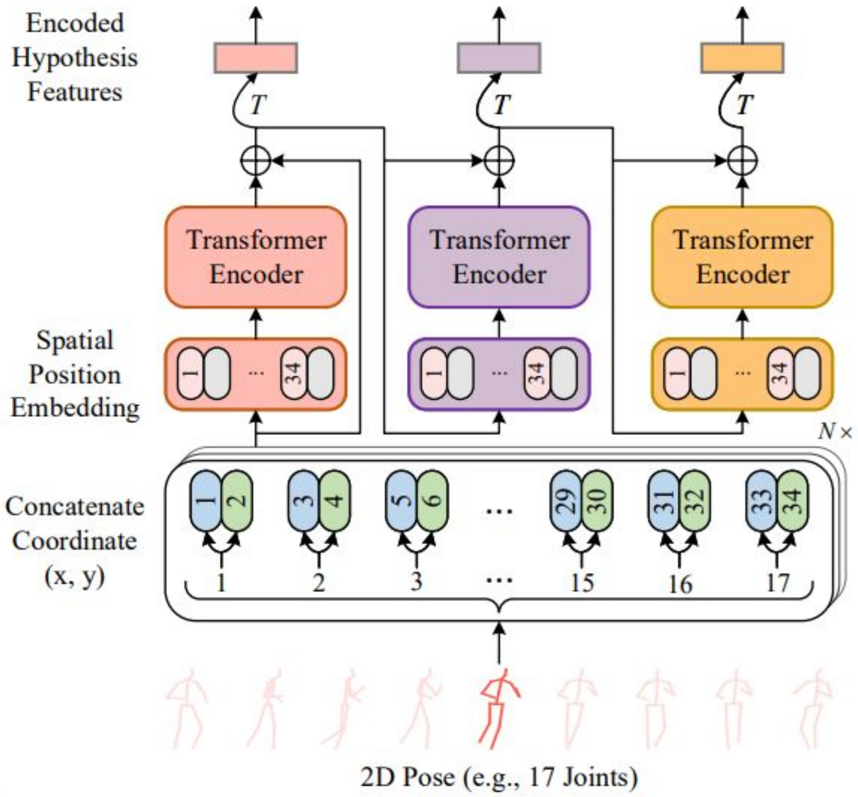
MHFormer

- Exploits the idea where a human pose in a video can have depth ambiguity and self-occlusion.
- Makes use of three major modules: Multi-hypothesis generation (MHG), self-hypothesis refinement (SHR) and cross-hypothesis interaction (CHI).



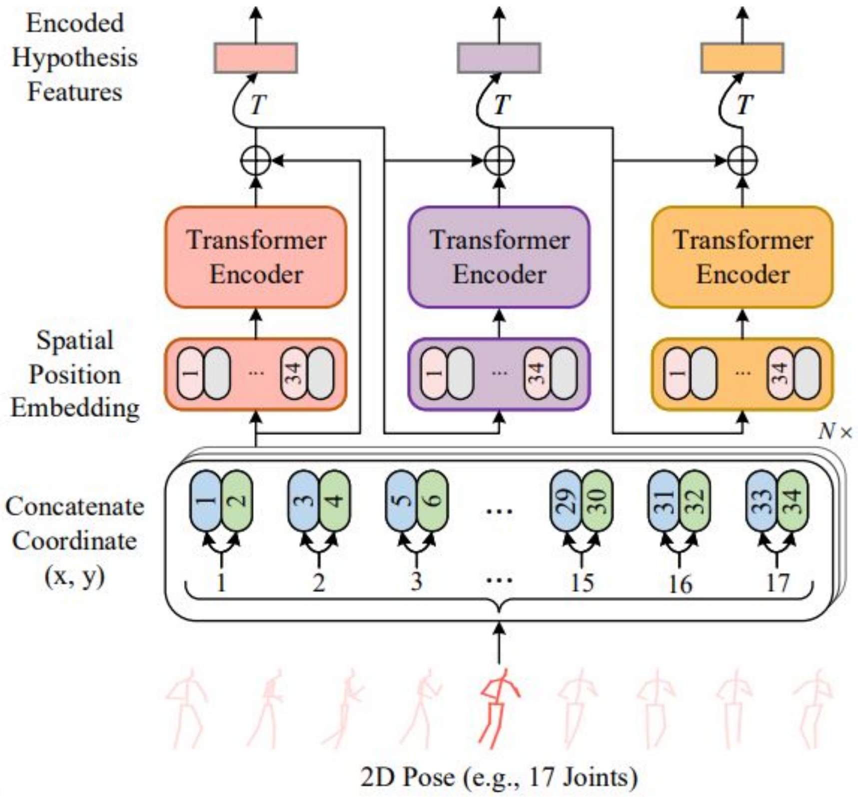
MHFormer

- Multi-hypothesis Generation takes a sequence of poses.
- Concatenates the (x, y) coordinates of joint for each frame and retain their spatial information.
- Output contains diverse information assuming different depth and occlusion.



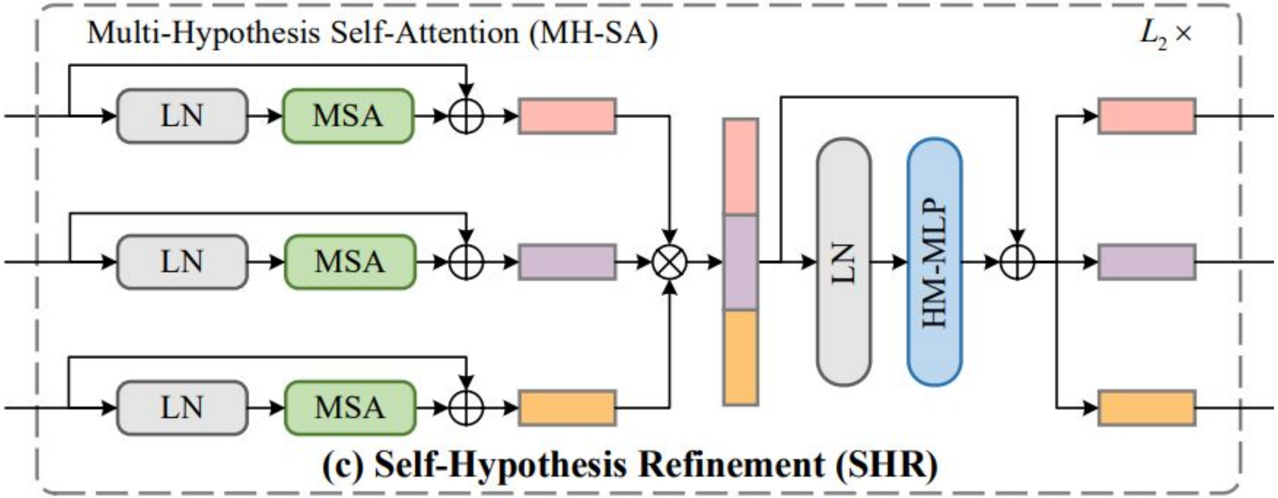
MHFormer

- Each output of MHG is embedded to a higher dimension feature.
- Learnable temporal positional embedding are added to it.



MHFormer

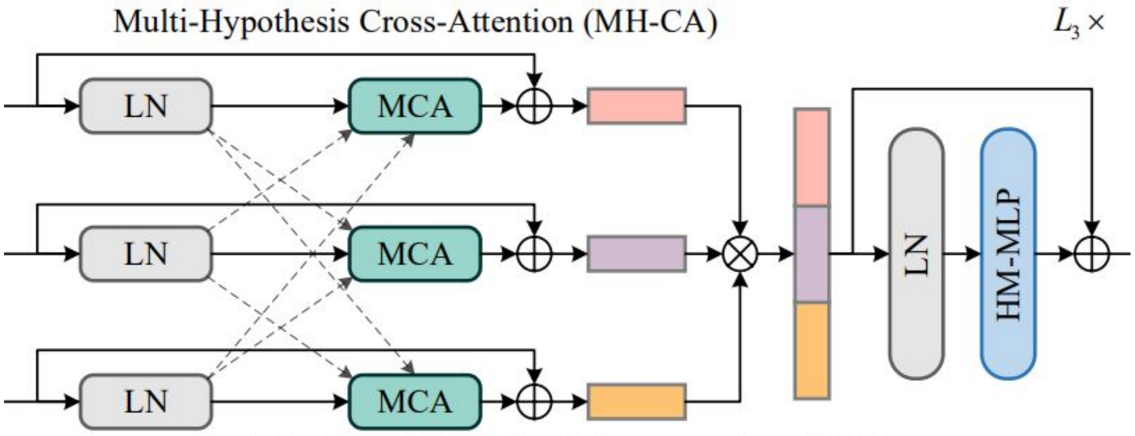
- Self Hypothesis refines each hypothesis in temporal domain with the use of multi-hypothesis self attention and hypothesis mixing MLP.
- MHSA leans from each hypothesis independently.
- Mixing MLP is used to exchange information among hypothesis.



https://arxiv.org/pdf/2111.12707.pdf

MHFormer

- Cross hypothesis interaction block boosts the performance.
- It includes multi hypothesis cross attention and hypothesis mixing MLP.
- MHCA computes cross correlation among cross hypothesis.
- Finally a regression head is used to produce 3D pose.



(d) Cross-Hypothesis Interaction (CHI)

Evaluation metric

- Protocol 1: Mean per joint projection error (MPJPE)

$$L = \frac{1}{J} \sum_{k=1}^{J} ||y_k - \hat{y}_k||_2$$

- Protocol 2: P-MPJPE: MPJPE after rigid alignment of the 3D pose using pose processing

Results

- Protocol 1(MPJPE)
- MHFormer performs the best in most of the categories.
- Poseformer performs better at Eat and Sit pose.

Method	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
PoseFormer	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
GraFormer	49.2	53.9	54.1	55.0	63.0	69.8	51.1	53.3	69.4	90.0	58.0	55.2	60.3	47.4	50.6	58.7
METRO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	54.0
MHFormer	**39.2**	**43.1**	40.1	**40.9**	**44.9**	**51.2**	**40.6**	**41.3**	54.5	**60.3**	**43.7**	**41.1**	**43.8**	**29.8**	**30.6**	**43.0**

Table 1: Protocol 1: MPJPE metrics for Poseformer, GraFormer, METRO and MHFormer on Human3.6M Dataset

Results

- Protocol 2(P-MPJPE)
- MHFormer still performs the best in most of the categories.
- GraFormer performs best at Phone.
- Poseformer performs best at Pose.
- MHFormer performs 12% better than Poseformer and 15% better than Graformer.

Method	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
PoseFormer	32.5	34.8	32.6	34.6	35.2	39.3	**32.1**	32.0	42.8	48.5	34.8	32.4	35.3	24.5	26.0	34.6
GraFormer	32.0	38.0	30.4	34.4	**34.7**	43.3	35.2	31.4	38.0	46.2	34.2	35.7	36.1	37.4	30.6	35.2
MHFormer	**27.7**	**32.1**	**29.1**	**28.9**	39.9	**33.9**	33.0	**31.2**	**37.0**	**39.3**	**30.0**	**31.0**	**29.4**	**22.2**	**23.0**	**30.5**

Table 2: Protocol 2: P-MPJPE metrics for Poseformer, GraFormer and MHFormer on Human3.6M Dataset

Results

- Hand Pose Evaluation
- Uses Mean Joint Error and Mean Mesh Error.
- HandOccNet performs better than METRO.
- An improvement of 12% and 27% is observed in joint and mesh error respectively.

Method	Joint	Mesh	F@5	F@15
METRO	10.4	11.1	48.4	94.6
HandOccNet	**9.1**	**8.1**	56.4	96.3

Table 3: Comparison metrics for METRO and HandOccNet using PA-MPJPE using HO-3D Dataset

Related Works

- Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers, 2021.
- Weixi Zhao, Yunjie Tian, Qixiang Ye, Jianbin Jiao, and Weiqiang Wang. Graformer: Graph convolution transformer for 3d pose estimation, 2021.
- Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers, 2021.

Related Works

- JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network, 2022.
- Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation, 2022.

Thank you for your attention!

DFKI GmbH
Department Augmented Vision
Trippstadterstr. 122
D-67663 Kaiserslautern

Ayush Chamoli

chamoli@rptu.de

https://av.dfki.de/