JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

AIST

Deformable Mesh Transformer for 3D Human Mesh Recovery

Yusuke Yoshiyasu

National Institute of Advanced Industrial Science and Technology (AIST)

[THU-AM-050]

DeFormer

Deformable Mesh Transformer

Task: 3D human mesh recovery

Input: Single RGB image

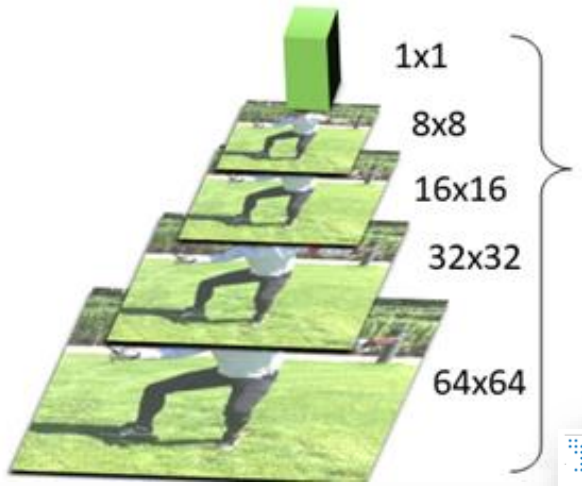
Output: 3D human body mesh

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

AIST

DeFormer

Deformable Mesh Transformer

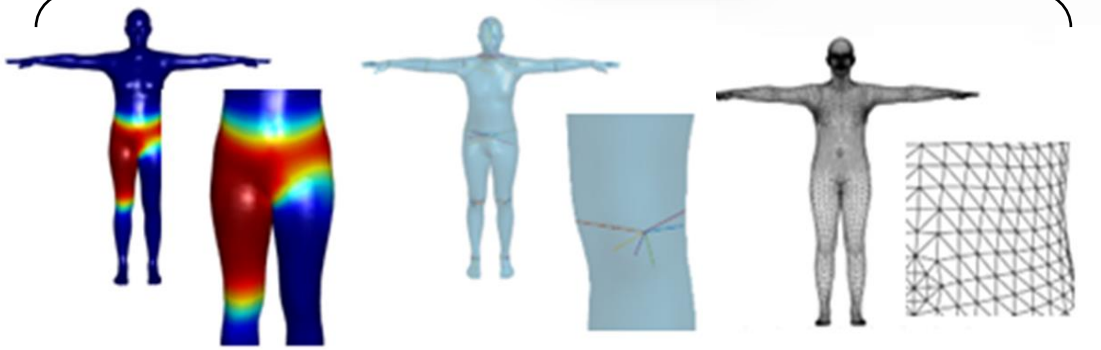


Multi-scale

7K vertex queries

Memory Efficient Transformer

Sparse Attention Pattern



Body prior knowledge

Contributions:

1. **Memory efficient** transformer decoder
2. **BodySparse Self-Attention**
   sparsifies self-attention based on prior knowledge extracted from body mesh/skeleton connectivity
3. **Deformable Mesh cross Attention**
   efficiently aggregates multi-scale image feature maps with deformation-driven attention

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

AIST

DeFormer

Deformable Mesh Transformer

- **DeFormer achieves SOTA performances**
- **Leverage multi-scale features & dense mesh queries**

Method	Human 3.6M	
	MPJPE↓	PA-MPJPE↓
METRO [CVPR 2021]	54.0	36.7
Graphormer [ICCV 2021]	51.2	34.5
PyMAF [ICCV 2021]	54.2	37.2
FastMETRO [ECCV 2022]	52.2	33.7
DeFormer (Ours)	**44.8**	**31.6**

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

AIST

Problem: 3D human mesh recovery

Input: Single RGB image

Output: 3D human body mesh

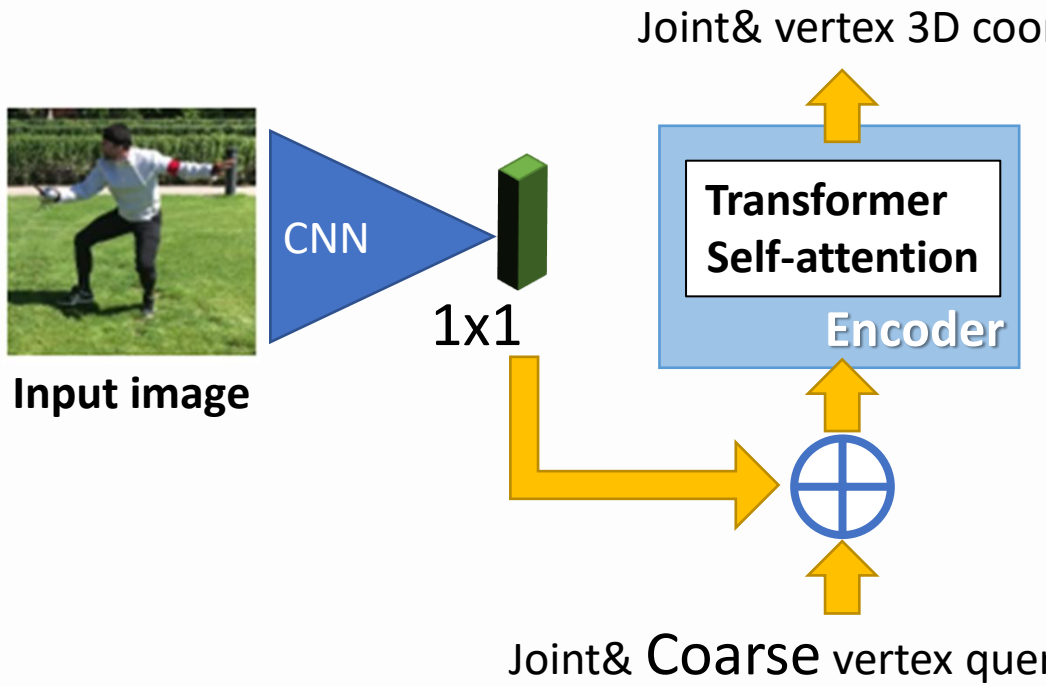
Previous approaches VS. DeFormer

- **Parametric-based:** Regress shape, pose and camera parameters
- **Vertex-based**: Regress 3D vertex coordinates

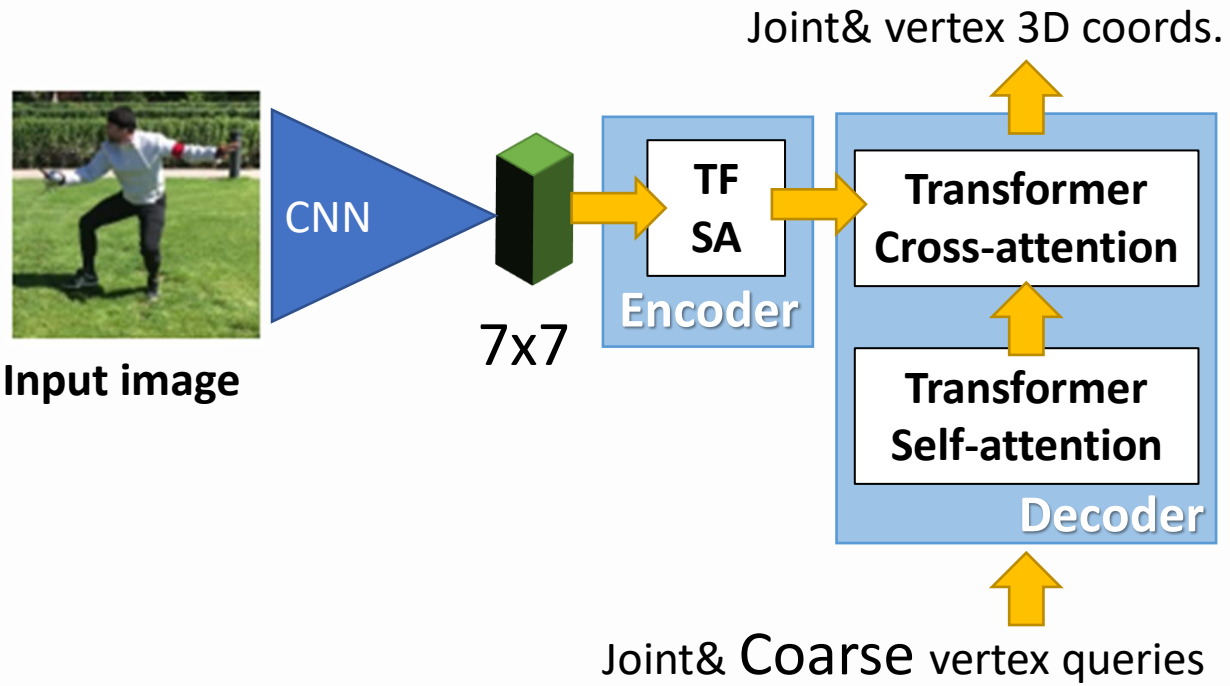
	Parametric-based	**Vertex-based**
Single-scale Feature maps	• HMR [CVPR 2018] • SPIN [ICCV 2019] MPJPE↓: ～60 PA-MPJPE ↓ : 41.1	• METRO [CVPR 2021] • Graphormer [ICCV 2021] • FastMETRO [ECCV 2022] MPJPE↓: 51.2 PA-MPJPE ↓ : 34.5
Multi-scale Feature maps	• PyMAF [ICCV 2021] • PyMAF-X [TPAMI 2023] MPJPE ↓ : 54.2 PA-MPJPE ↓ : 37.2	**Ours: DeFormer** **MPJPE ↓ : 44.8 PA-MPJPE ↓ : 31.6**

Previous Transformer Approaches

- **METRO [CVPR 2021]**



- **FastMETRO [ECCV 2022]**

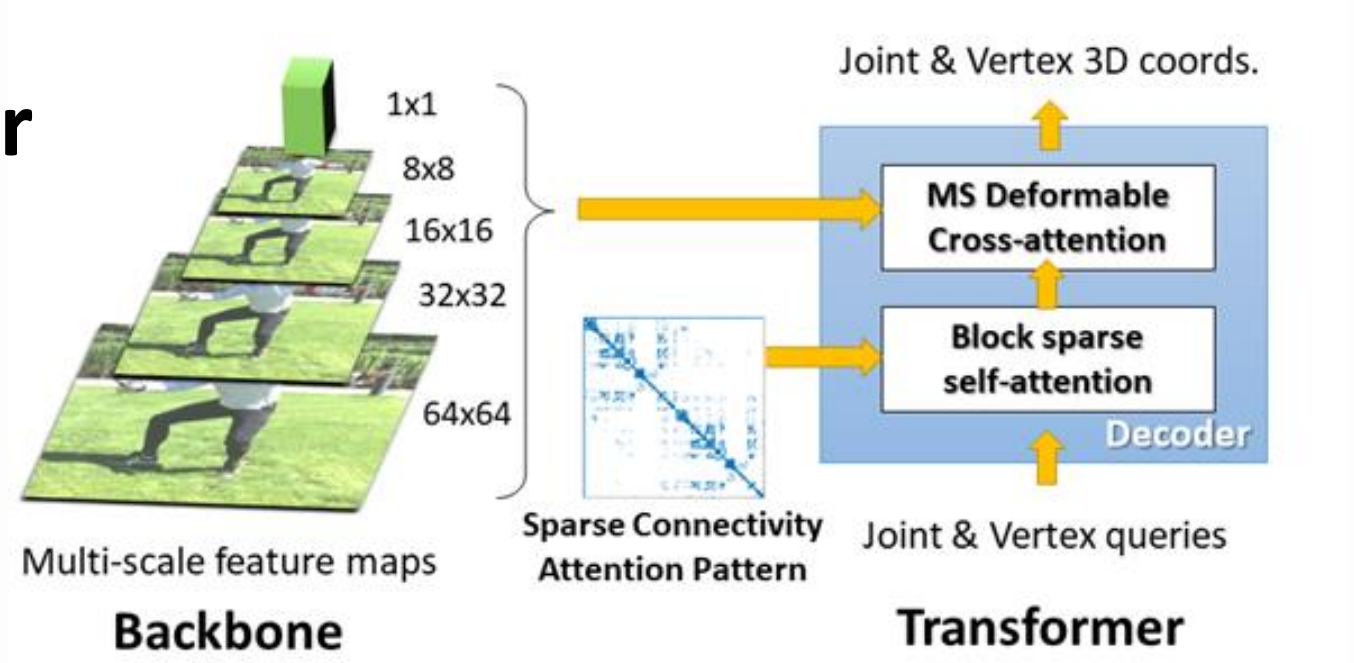


*Limited to **single-scale low res.** image features and **coarse** vertex queries*

Challenge: Mesh transformer approaches are ***memory intensive***

Our Approach

- **DeFormer**

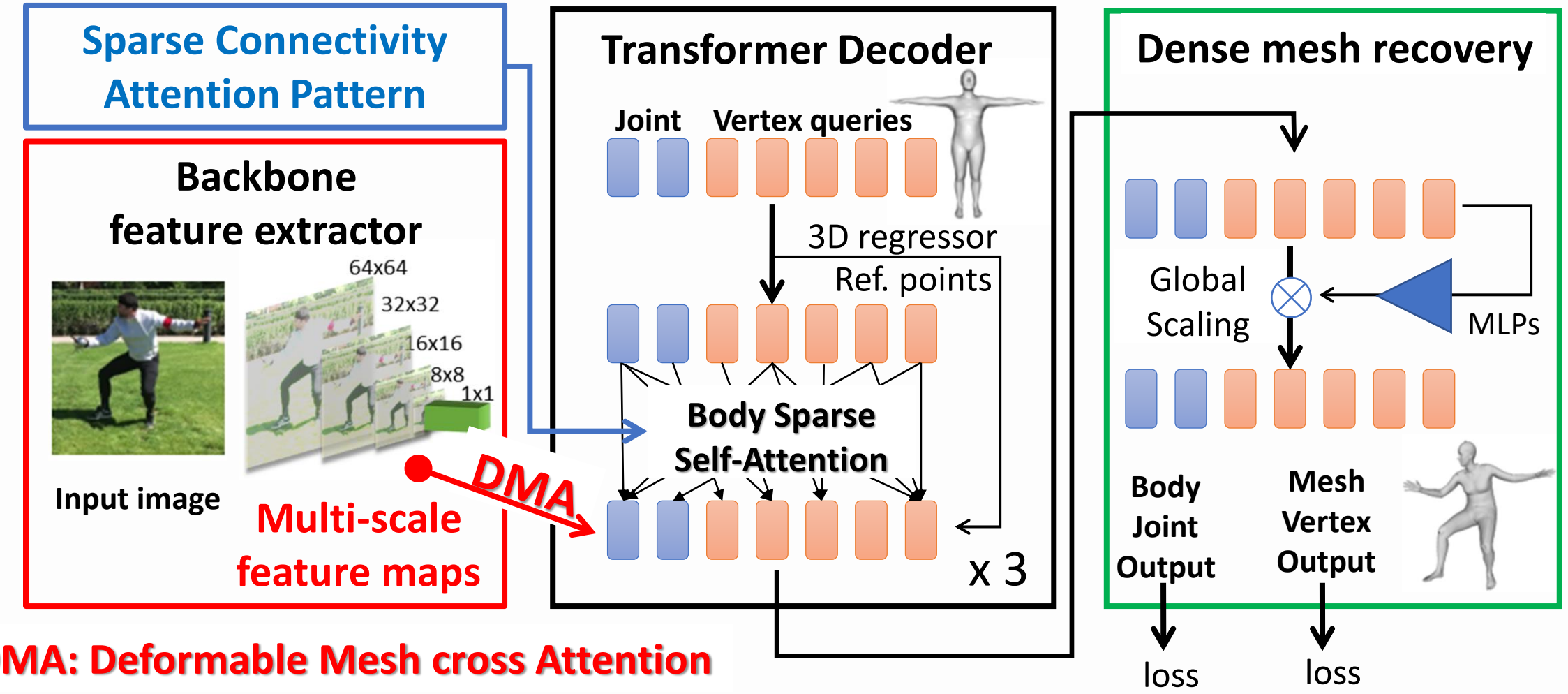


Challenge: Mesh transformer approaches are ***memory intensive***

Goal: Mesh transformer approach to human mesh recovery that ***efficiently*** leverages multi-scale visual feature maps and dense mesh queries

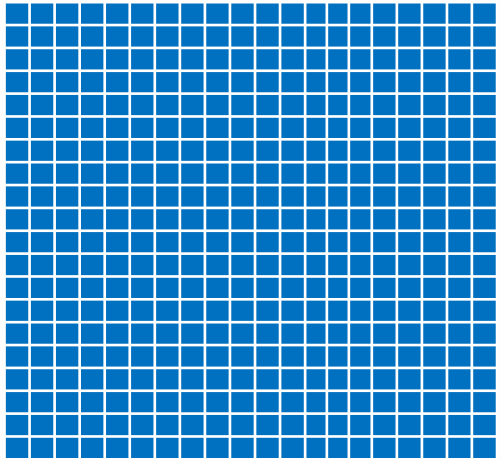
DeFormer: A **decoder-only** mesh transformer with memory-efficient attention modules

Overview of DeFormer architecture

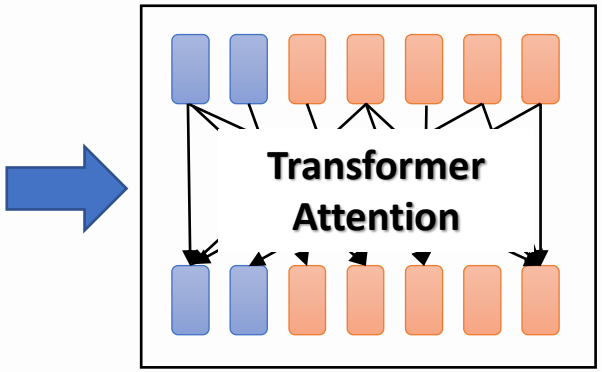


DMA: Deformable Mesh cross Attention

Background: Transformer Attention



Standard Transformer Attention Pattern



Train/inference

500 x 500 queries OK

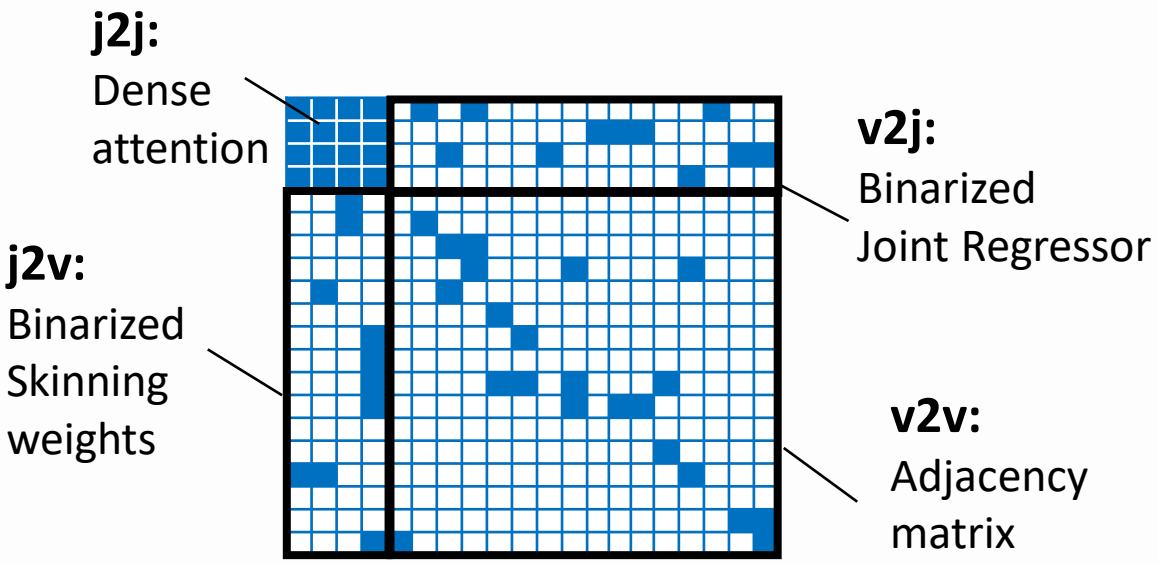
7000x7000 queries ✗

Multi-head attention (MHA)

- Long-range dependencies
- Quadratic complexity w.r.t seq length

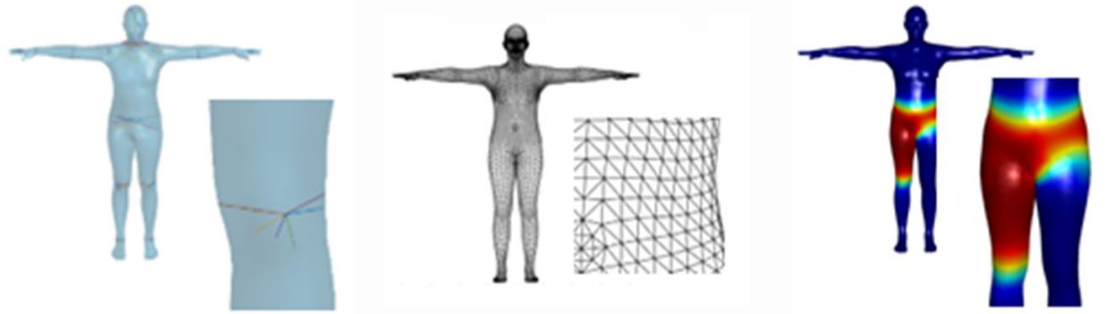
$$\mathrm{MHA}(\mathbf{z}_q, \mathbf{x}) = \sum_{m=1}^{M} \mathbf{w}_m [\sum_{k=\Omega_k} A_{mqk} \cdot \mathbf{w'}_m \mathbf{x}_k]$$

Body Sparse Self-Attention



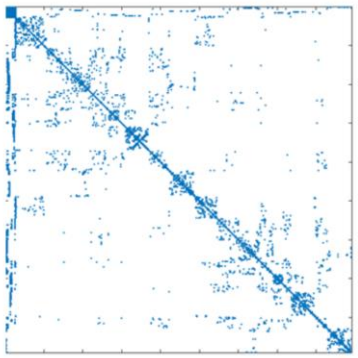
Sparse Connectivity Attention Pattern (SCAP)

Prior knowledge

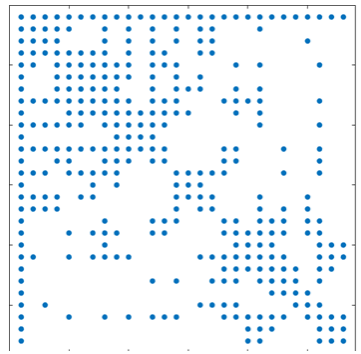


vertex to joint (v2j) | **verts to verts (v2v)** | **joint to vertex (j2v)**

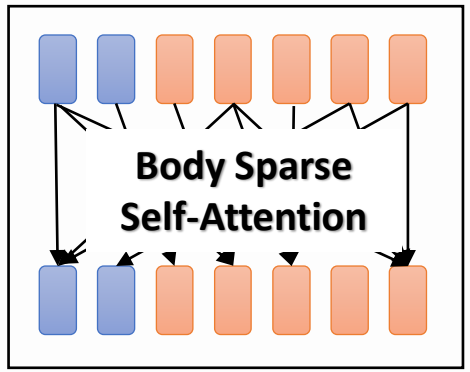
Body mesh/skeleton connectivity



Tile 16x16 blocks



To GPU

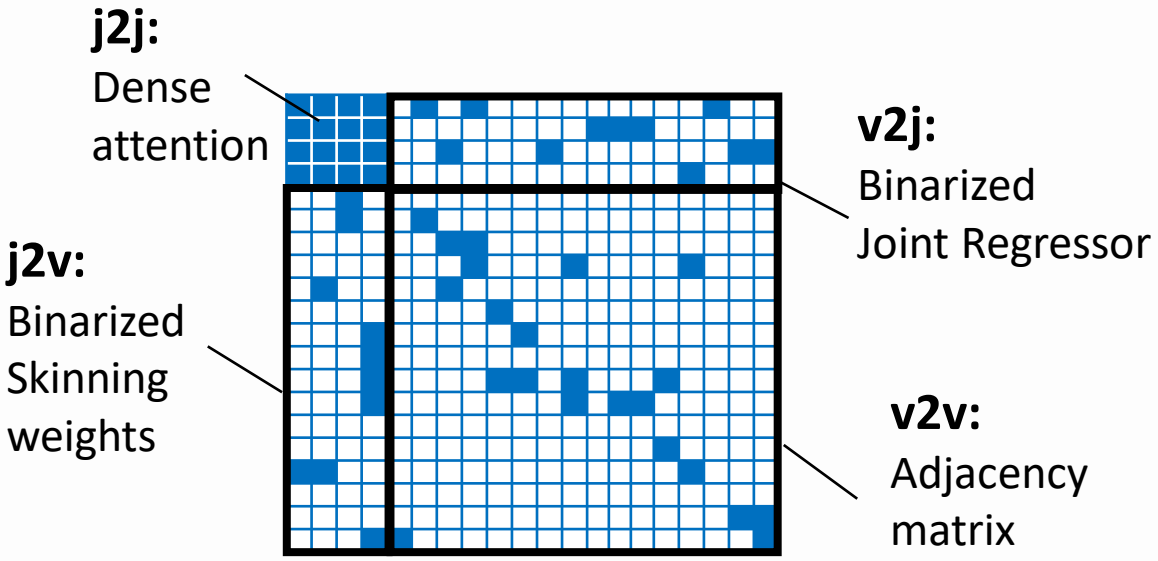


1. Construct SCAP

Step 2. Blockify SCAP

3. Train/inference

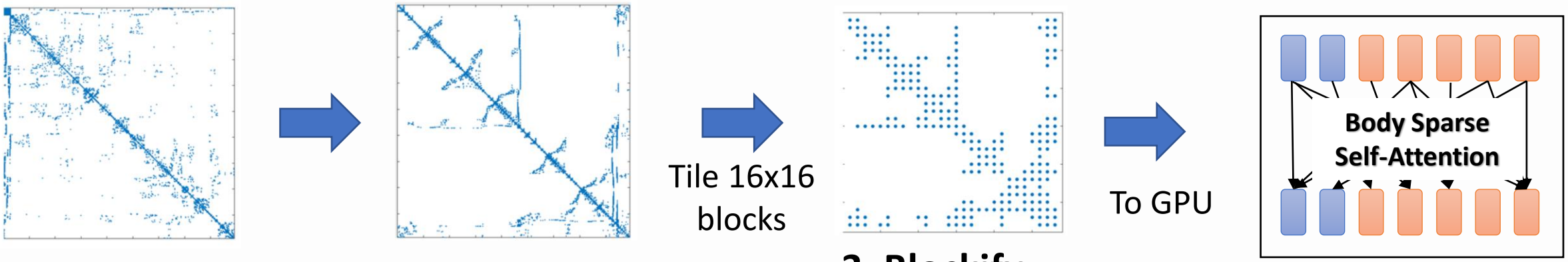
Body Sparse Self-Attention



500 x 500 queries OK

7000x7000 queries OK

Sparse Connectivity Attention Pattern (SCAP)



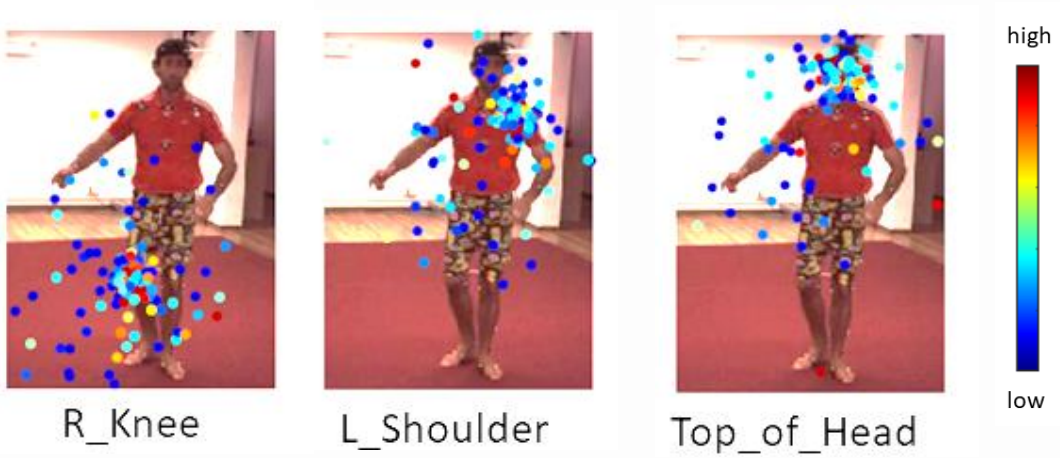
1. Construct SCAP **2. Perform AMD** **3. Blockify permuted SCAP** **4. Train/inference**

Deformable Mesh cross Attention

- **Deformation driven cross attention**
- **Mesh alignment feedback loop in transformer decoder**

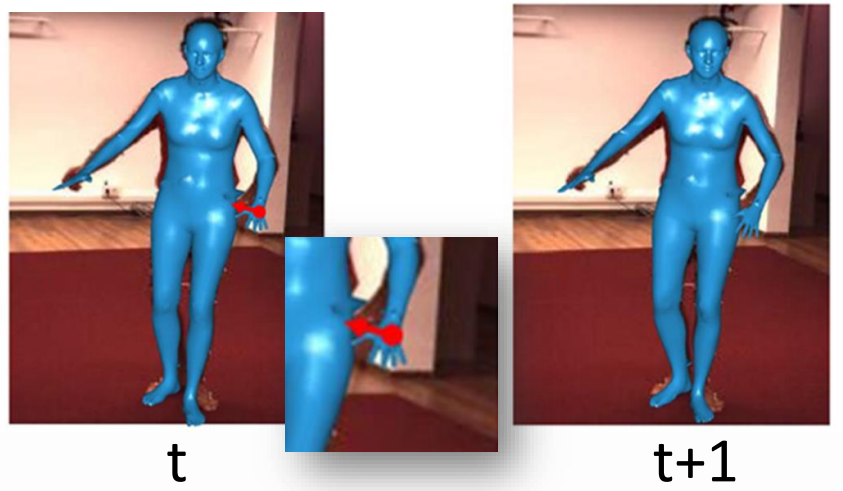
1. Get reference points from the current 3D joints/mesh

2. Perform multi-scale deformable attention



Deformable Mesh cross Attention

- **Deformation driven cross attention**
- **Mesh alignment feedback loop in transformer decoder**



-> 1. Get reference points from the current 3D joints/mesh

2. Perform multi-scale deformable attention

3. Regress corrective displacements

– – – 4. Refine 3D joint/mesh vertex positions

x3

$$\Delta \mathbf{X}_{\mathrm{Im}}^{t} = \Psi(\Phi(\mathbf{X}_{\mathrm{Im}}^{t}) \oplus \mathbf{x}_{\mathrm{q}}^{t})$$

MLPs, Concat, Current reconstruction, Features

$$\mathbf{X}_{\mathrm{Im}}^{t+1} = \mathbf{X}_{\mathrm{Im}}^{t} + \Delta \mathbf{X}_{\mathrm{Im}}^{t}, \text{ for } t > 0.$$

Experiments

- Implementation: Pytorch, Deepspeed, MMPose
- Backbone feature extractor
  - CNNs: ResNet50, HRNet-w48
  - Visual transformer: AggPose, HRFormer
  - Pre-trained on 2D human pose dataset: MPII or COCO
- Training
  - Multiple dataset: Human3.6M, MuCo-3DHP, UP3D, COCO, MPII
  - AdamW optimizer
  - 50 epochs
  - About 1 day. Finishes within 2 days
  - 8 x NVIDIA A100 GPUs (40GB memory)
- Evaluation metrics: MPJPE, PA-MPJPE, MVE

Quantitative comparison

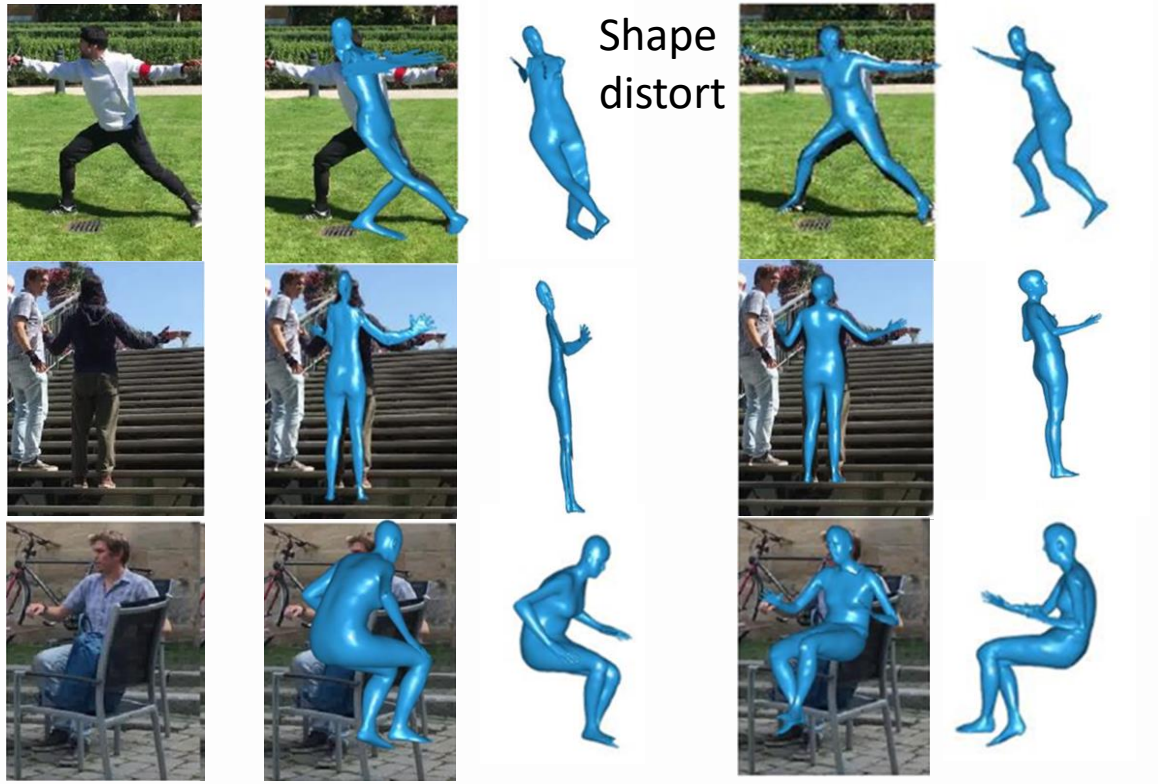
- **DeFormer achieve SOTA performances**
- **Larger input image, more queries, visual transformer backbone improves performance**

Method	Human 3.6M		3DPW		
	MPJPE↓	PA-MPJPE↓	MVE↓	MPJPE↓	PA-MPJPE↓
SPIN	—	41.1	116.4	—	59.2
METRO	54.0	36.7	88.2	77.1	47.9
Graphormer	51.2	34.5	87.7	74.7	45.6
PyMAF	54.2	37.2	87.0	74.2	45.3
FastMETRO	52.2	33.7	84.1	73.5	44.6
DeFormer (HRNet-W48)	**44.8**	**31.6**	**82.6**	**72.9**	**44.3**
DeFormer (384x288 img)	**43.7**	**30.7**	—	—	—
DeFormer (6904 queries)	**43.9**	**30.7**	—	—	—
DeFormer (HRFormer)	**43.3**	**30.0**	—	—	—

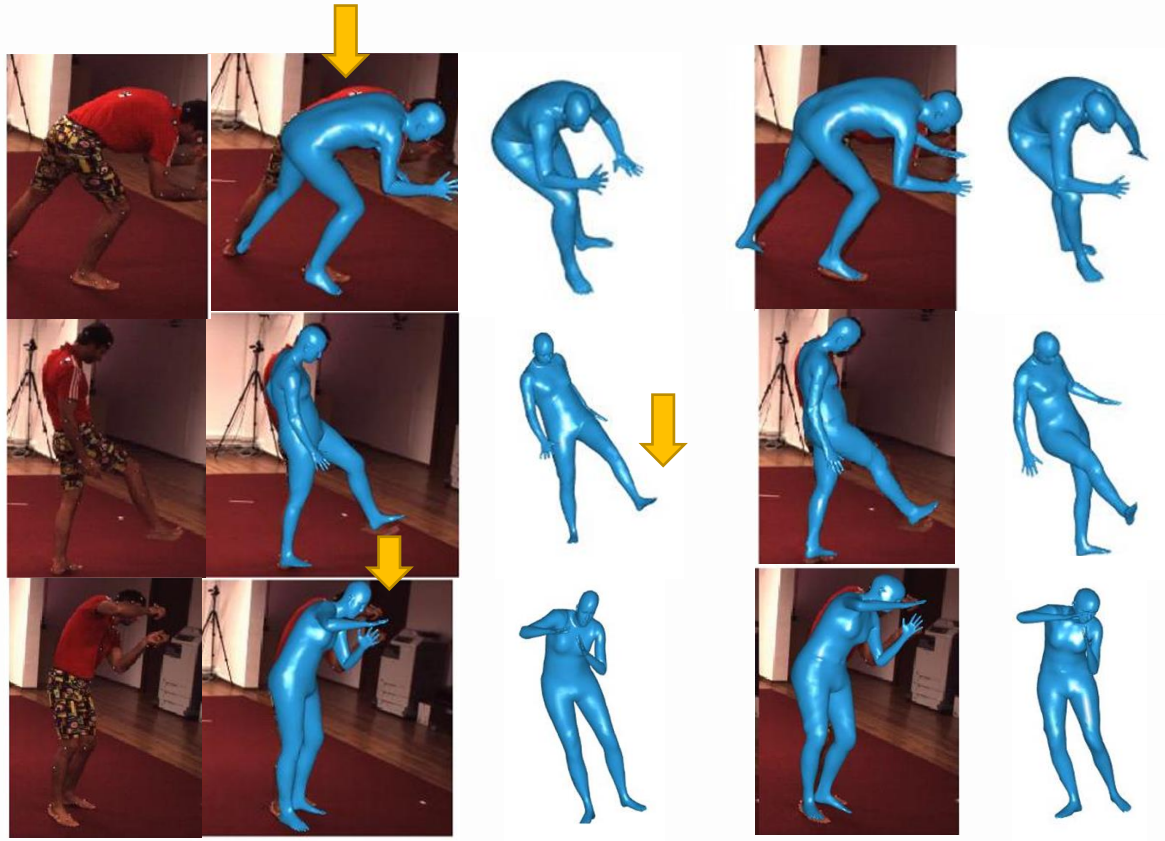
Qualitative comparison to METRO

- **Better image-mesh alignment**
- **Leveraging global/local spatial contexts from multi-scale features**

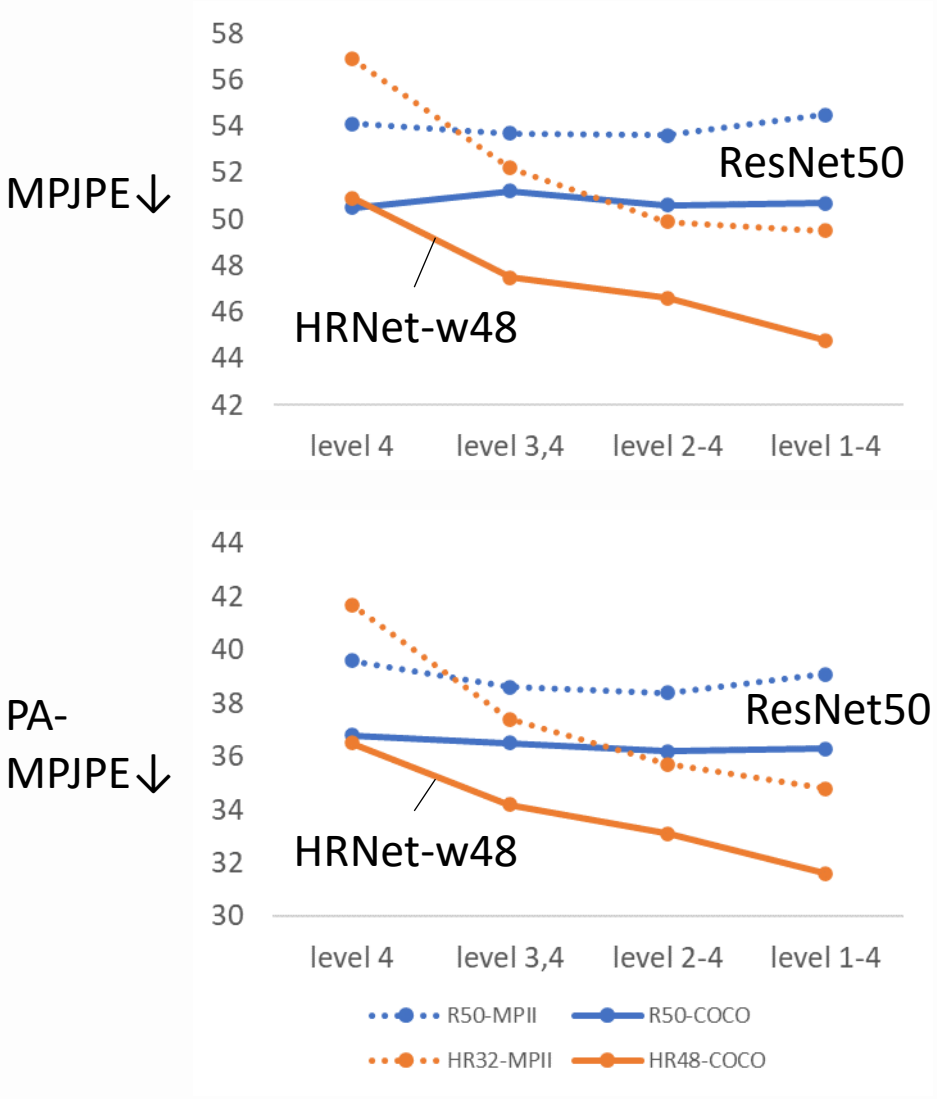
METRO DeFormer

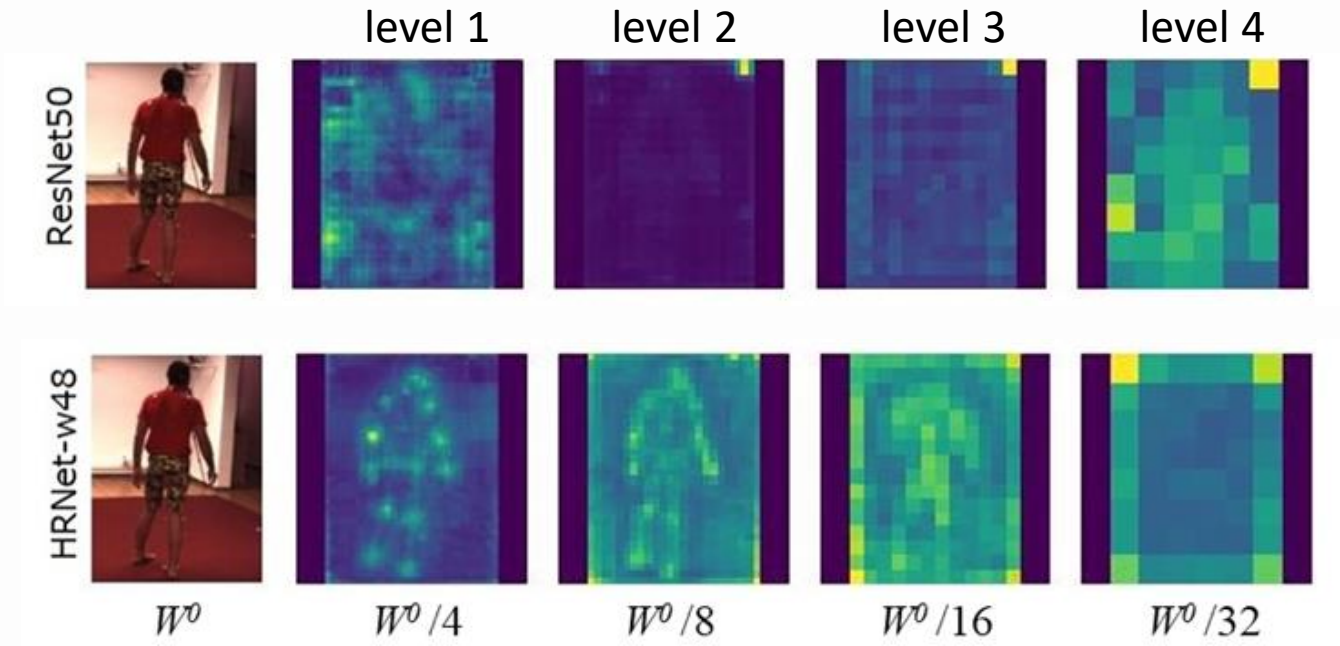


METRO DeFormer



Multi-scale feature maps



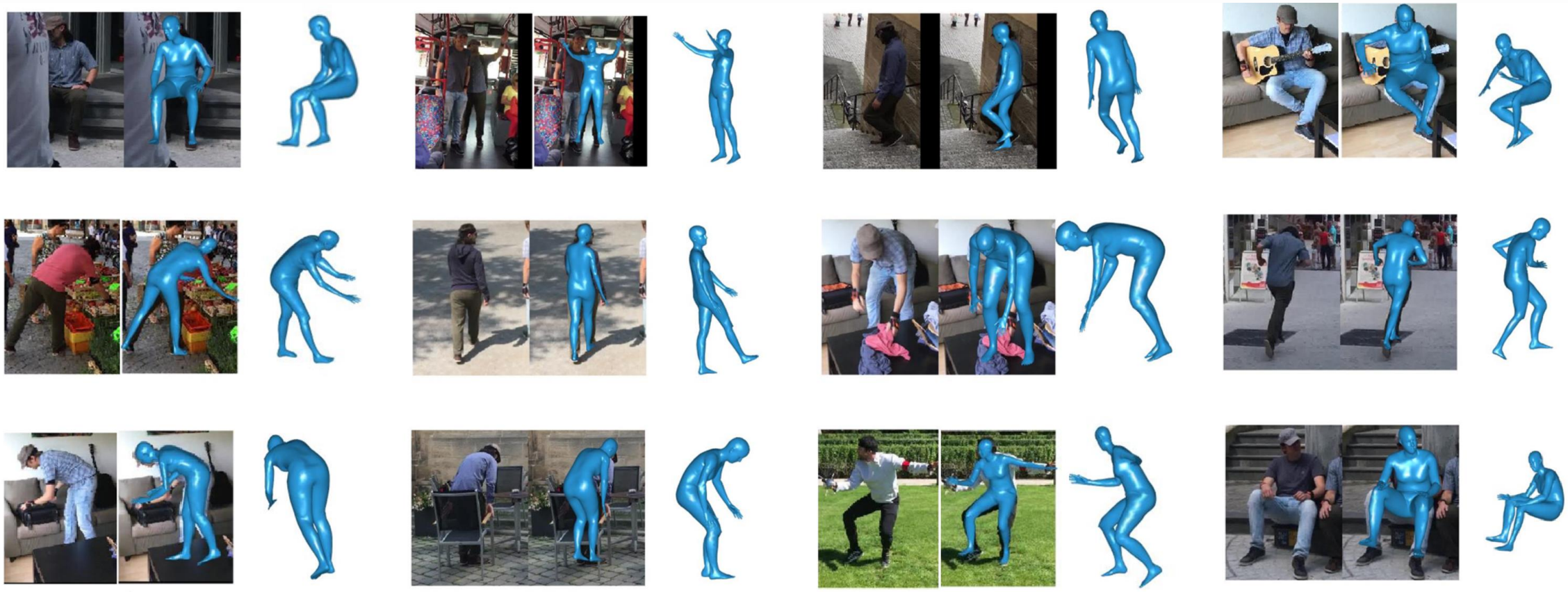


- DeFormer leverages multi-scale feature maps
- Works good with HRNet variants which produce meaningful high-res feature maps

Summary

- **DeFormer**: A **decoder-only** mesh transformer with memory-efficient attention modules
- **Body Sparse Self-Attention** sparsifies self-attention access patterns according to body mesh/skeleton connectivity
- **Deformable Mesh cross Attention (DMA) efficiently** aggregates multi-scale image feature maps with a deformation-driven attention mechanism
- **SOTA performances** leveraging multi-scale visual feature maps and dense meshes

Thank you



Poster [THU-AM-050]