




Efficient 3D Human Pose and Shape Estimation Using Group-Mix Attention in Transformer Models

Yushan Wang¹^a, Shuhei Tarashima^{1,2}^b and Norio Tagawa¹^c

¹*Faculty of Systems Design, Tokyo Metropolitan University, Tokyo, Japan*

²*Innovation Center, NTT Communications Corporation, Tokyo, Japan*


Keywords: Transformer, 3D Human Pose and Shape Estimation, ViT, Attention Visualization.


Abstract: Fully-transformer frameworks have gradually replaced traditional convolutional neural networks (CNNs) in recent 3D human pose and shape estimation tasks, especially due to its attention mechanism that can capture long-range and complex relationships between input tokens, surpassing CNN's representation capabilities. Recent attention designs have reduced the computational complexity of transformers in core computer vision tasks like classification and segmentation, achieving extraordinary strong results. However, their potential for more complex, higher-level tasks remains unexplored. For the first time, we propose to integrate the group-mix attention mechanism to 3D human pose and shape estimation task. We combine token-to-token, token-to-group, and group-to-group correlations, enabling a broader capture of human body part relationships and making it promising for challenging scenarios like occlusion+blur. We believe this mix of tokens and groups is well suited to our task, where we need to learn the relevance of various parts of the human body, which are often not individual tokens, but larger in scope. We quantitatively and qualitatively validated our method successfully reduces the parameter count by 97.3% (from 620M to 17M) and the FLOPs count by 96.1% (from 242.1G to 9.5G), with a performance gap of less than 3%.


1 INTRODUCTION

Estimating 3D human pose and shape (HPS) from a monocular image is a process of inferring the three-dimensional positions of a person's joints (pose) as well as their overall body shape. This task is highly challenging considering frequently encountered problems such as the complexity of human articulation, environmental occlusion, human self-occlusion, and blurring in 2D images caused by rapid movement during photography. 3D human pose and shape estimation receives significant attention in the computer vision community and holds a crucial role in various applications, including motion capture for film and animation, virtual fashion shows and runway modeling, remote rehabilitation, augmented reality (AR) and virtual reality (VR). Existing approaches can be divided into two types: optimization-based methods and deep learning-based methods. Optimization-based approaches, such as Hybrik (Li et al., 2021),

and (Zanfir et al., 2018), typically involve iteratively fitting body models to 2D observations, e.g., silhouettes, segmentations and 2D keypoints. The parameters of statistical body model like SMPL (Bogo et al., 2016) are optimized to minimize the error between its 2D projection and these observations. However, these optimization-based methods are highly sensitive to the chosen initialization, leading to challenges in fine-tuning algorithms. Early deep learning-based approaches such as HMR (Kanazawa et al., 2018), PyMAF (Zhang et al., 2021) and PARE (Kocabas et al., 2021), leverage the nonlinear mapping capability of neural networks to directly predict model parameters from pixel-level information from the raw image, and these pixel-level features lead to more realistic and plausible predictions. More recently, researchers have developed fully-transformer frameworks such as METRO (Lin et al., 2021a), Adaptive Token (Xue et al., 2022) and transcends the spatial limitations of CNNs thanks to its attention mechanism, which is adept at capturing intricate long-distance relationships between input tokens. We observed that the majority of current 3D HPS estimation from images predominantly concentrates on feature fusion or struc-

^a <https://orcid.org/0009-0006-4317-7342>

^b <https://orcid.org/0009-0007-6022-2560>

^c <https://orcid.org/0000-0003-0212-9265>

tural design. Given the substantial focus on reducing model parameters and computational complexity in both Natural Language Processing (NLP) and foundational Computer Vision (CV) tasks (e.g., classification and segmentation), it is worth exploring the integration of state-of-the-art attention designs into our higher-level HPS task.

In this work, for the first time, we propose to integrate group-mix attention (GMA) (Ge et al., 2023) to fully-transformer architecture in 3D HPS task. In GMA, the author argues that Conventional attention map, derived from queries and keys, merely can capture token-to-token correlations at a single granularity, whereas self-attention should embrace a broader mechanism to capture correlations between tokens and groups, enhancing its representational power. We strongly agree with this perspective, as in 3D human pose and shape estimation we need to learn the relevance of various parts of the human body, which are often not individual tokens, but larger in scope, like groups. By aggregating tokens into a group, we not only capture the relationships between body parts on a larger scope but also significantly reduce computational costs. This is because the grouped tokens serve as a single new proxy, and the number of proxies is much smaller than the original number of individual tokens. Consequently, during the attention computation, we operate on these proxies rather than the individual tokens, which reduces the token count in the attention operation and lowers the computational load. Traditional attention mechanisms also introduce attention redundancy, as neighboring tokens often contain similar or overlapping information. We address this problem by grouping multiple tokens into a single proxy, effectively reducing this redundancy, lowering computational costs, and enhancing the model’s ability to capture broader relationships among human body parts.

We validated our method on popular 3D HPS datasets following HMR2.0 (Goel et al., 2023), reducing the parameter count by 97.3% (from 620M to 17M) and the FLOPs count by 96.1% (from 242.1G to 9.5G), with a performance gap of less than 3%. Our model excels in occlusion+blur scenarios by capturing broader correlations among human body parts. For explainability, we import the attention visualization for each branch of our architecture, and these visualizations reveal interesting insights into how the model processes information. For instance, they can show whether the model attends to global structures or local details, providing clues about the model’s reasoning process.

2 RELATED WORKS

2.1 3D Human Pose and Shape Estimation from Images

We focus on deep learning-based methods, which have two main types of outputs: parametric outputs and non-parametric outputs. Most image-based methods such as I2LMeshNet (Moon and Lee, 2020), Pose2Mesh (Choi et al., 2020), ROMP (Song et al., 2020) opt to directly regress the parameters of a parametric model. Since they leverage the nonlinear mapping capability of neural networks to directly predict model parameters from pixel-level information from the raw image, the networks only need to produce a low-dimensional vector in the parametric model, which includes body pose θ , shape β , and camera parameters $\pi = (R, t)$ comprising global orientation R and translation t . For non-parametric method, instead of predicting template parameters, they directly output body shapes in the form of voxels (Varol et al., 2018) or positions of mesh vertices (Kolotouros et al., 2019) in three-dimensional space. Given the parametric model’s strong a priori knowledge and its capability to handle occlusion, blurring, and joint articulation issues effectively, we choose the parametric output approach.

2.2 Transformer Based Methods

The prevailing transformer-based methods for 3D human pose and shape estimation primarily focus on feature fusion and structural design, PMCE (You et al., 2023) proposes a symmetric transformer enabling joint-vertex interaction via cross-attention and adaptive layer normalization (AdaLN). Mesh graphormer (Lin et al., 2021b) combines graph convolutions and self-attentions in a transformer to model both local and global 3D vertex-pose interactions. Component aware transformer (Lin et al., 2023) introduces feature-level upsampling-crop to enhance resolution for small body parts like hands and face. All these methods have addressed specific challenges in 3D human pose and shape estimation tasks and achieved promising results. However, none have focused on the issue of reducing model parameters and computational effort, which motivates our work.

2.3 Attention Designs for Computational Reduction

The attention mechanism, as a key component of both transformer and ViT architectures, involves comput-

ing the attention scores between each pair of tokens, resulting in a quadratic increase in computation as the sequence length increases. This leads to significant computational costs when modeling global information. In foundational CV tasks, RMT (Fan et al., 2024) achieves linear complexity by decomposing attention through horizontal and vertical directions respectively. CSWin Transformer (Dong et al., 2022) achieves efficient attention by employing a cross-shaped window mechanism along horizontal and vertical stripes, balancing computational cost and interaction range. We strongly resonate with GMA (Ge et al., 2023) that the reduction in computation should be accompanied by a more sophisticated expressive capability to capture the relationship between tokens and groups, rather than solely a single token-to-token correlation. Thus, we aggregate adjacent tokens into one group to form a single proxy, enabling token-to-group and group-to-group relationships, while directly reducing the computational load of the attention mechanism by processing fewer resources (i.e., proxy).

3 METHOD

In the following subsections, we offer an overview of our methodology for estimating 3D human poses from an input image. Next, we delve into the details of group-mix attention.

3.1 Group-Mix Attention for HPS

The Group Mix Attention (GMA) mechanism enhances Vision Transformers (ViTs) by overcoming the limitations of traditional self-attention: conventional self-attention focuses solely on pairwise token interactions at a single granularity, resulting in quadratic complexity and limited contextual understanding. GMA extends attention beyond token-to-token interactions to include token-to-group and group-to-group relationships, improving both efficiency and representational power. This is accomplished by dividing the qkv entries into token groups and creating proxies for these groups, which are then utilized in the attention calculation process. In particular, we divide the conventional Query, Key, and Value components into segments and employ group aggregation operation to establish these proxies.

As a core component, the aggregation operation is different for each segment, as illustrated in figure 1, we adopt depth-wise convolutions with various kernel sizes to implement aggregation. In order to present the aggregation operation more clearly,

we can roughly divide the aggregation structure into two parts: attention branch and no-attention branch. We divide Q , K , and V into five parts represented as $x[0, 1, 2, 3, 4]$ and employ aggregators with kernel sizes 1, 3, 5, 7 to create group proxies for three of them. Note that the branch $x[0]$ is an identity mapping equivalent to a traditional token-to-token mechanism. i.e. attention branch is represented as $x[0, 1, 2, 3]$. We obtain different group proxies from four branches, which will be used for calculating attention later. This allows us to perform attention calculations on combinations of individual tokens and group proxies at different levels of detail. For non-attention branch, incorporating a non-attention branch can introduce a form of architectural diversity that potentially increases the robustness of the model.

Compared to traditional attention, by using aggregation, we can get (1) Efficient group proxies: we aggregate adjacent tokens into proxies using depth-wise convolution, significantly reducing the number of token pairs for attention calculation, and thus lower computational costs. (2) Multi-level attention: By managing attention across token-to-token, token-to-group, and group-to-group levels, we can capture relationships at multiple granularities without redundant recalculations, streamlining the attention process. Additionally, we visualize this aggregation process during attention mechanism analysis, as it serves as one of our motivations.

3.2 Overview

The overall architecture of our approach is illustrated in figure 1 (A). Given an input image $I \in \mathbb{R}^{C \times H \times W}$, where C denotes the 3 channels for the input image initially, and H and W represent height and width, respectively. We apply a slicing operation to transform the input into a shape of $[3, 256, 192]$. This is followed by a convolutional stem block (figure 1 (B)) consisting of two convolution operations, which changes the shape from $[3, 256, 192]$ to $[200, 64, 48]$. Subsequently, through the patch embedding block (figure 1 (C)) we get the input $I_p \in \mathbb{R}^{N \times C}$, where $N = H \times W = 64 \times 48 = 3072$, denoting the number of tokens. This I_p serves as input to stage 0 in the network structure when p equals 0. In contrast to general attention mechanism in transformer (Vaswani et al., 2017) or HMR2.0 (Goel et al., 2023) we adopt GMA for higher representational capacity, where we set group sizes to 1, 3, 5, and 7. A group size of 1 signifies the aggregation of individual tokens as general attention. Unlike the previous GMFormer (Ge et al., 2023), which has 4 stages, original ViT in HMR2.0 (Goel et al., 2023) maintains a consistent feature map

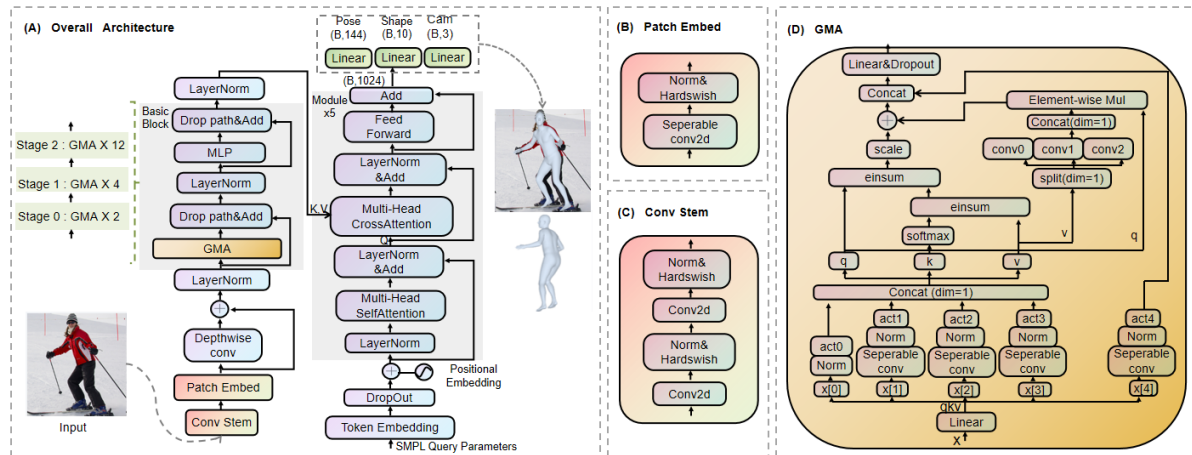


Figure 1: Panel (A) presents an overview of our method, where we employ a fully-transformer framework: taking monocular images as input, extracting image features using a 4-stage ViT backbone integrated with Group-Mix Attention, and then passing the extracted features through a transformer decoder to regress human shape and pose parameters. Panels (B) and (C) depict patch embedding process, representing the same operations in our backbone, as indicated by their matching colors. Panel (D) (in orange) illustrates the detailed structure of the group-mix attention (GMA) mechanism.

resolution of $H/14 \times W/14$ throughout. In GMFormer, the resolution varies: it is $H/14 \times W/14$ in stage 3 and downsampled to $H/28 \times W/28$ in Stage 4. To align the backbone’s feature map resolution with the decoder’s input, we adopt a 3-stage architecture for the GMA-based ViT, with a serial depth of [2, 4, 12], following the first three stages of GMFormer (Goel et al., 2023). The final output of the GMA-based ViT, referred to as context, is shaped as [192,320] and further transformed into [192,512]. We employ an 8-head attention mechanism, deriving the key and value from the context with a shape of [8,192,64]. The decoder initializes with a zero input token of shape [1,1], which is subsequently passed through a token embedding layer to obtain the decoder input of shape [1,1024]. This decoder input is further transformed into query of shape [8,1,64]. We process this query by cross-attending to the key and value derived from context. A stack of 6 decoder layers is employed, concluding with a linear regression step to predict human pose, shape, and camera parameters, respectively.

4 EXPERIMENTS

4.1 Datasets and Implementation Details

4.1.1 Datasets

For the training, Human3.6M (Ionescu et al., 2014), MPI-INF-3DHP (Mehta et al., 2017), COCO (Lin

et al., 2014), MPII (Andriluka et al., 2014), InstaVariety (Kanazawa et al., 2019), AVA (Gu et al., 2018), AI Challenger (Wu et al., 2017) are used, these datasets include controlled indoor scenes and dynamic outdoor settings, catering to a variety of needs for our human pose and shape estimation task. We use 3DPW (Von Marcard et al., 2018), Human3.6M val split for 3D evaluation and COCO validation set, Posetrack validation set (Andriluka et al., 2018) for 2D evaluation following previous work (Goel et al., 2023).

4.1.2 Implementation Details

Our GMA-based ViT network adopts the pretraining paradigm proposed by HMR2.0 (Goel et al., 2023). Specifically, we adopt a training procedure similar to ViTPose (Xu et al., 2022). First, our proposed model is pretrained on an ImageNet-based classification task to learn global feature priors. Subsequently, we fine-tune the model on the COCO 2D keypoint dataset for keypoint prediction. This fine-tuning process allows the model to acquire prior knowledge of human pose estimation. Notably, our method achieves pose estimation performance that is slightly superior to ViTPose, as shown in table 1. In 3D HPS task, the input image is resized to 256×192 before being fed into the GMA-based ViT encoder. In our experiments, the main model is trained using 8 A100 GPUs with a batch size of 512 (8 GPUs \times 64). We employ the AdamW optimizer with a learning rate of $5e^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.1. Our implementation is in PyTorch.

Table 1: We compare the pose estimation performance of our GMA-based ViT with ViTPose in terms of AP (Average Precision) and AR (Average Recall).

Backbone	Resolution	AP	AR
ViTPose	256 × 192	73.8	79.2
GMA-based (Ours)	256 × 192	74.9	80.0

4.2 Quantitative Comparison

In this section, we report quantitative comparison. The comparison is conducted on 3DPW, Human3.6M val split for 3D evaluation and COCO validation set, Posetrack validation set for 2D evaluation, focusing on 3D metrics such as Mean Per Joint Position Error (MPJPE), Procrustes Aligned Mean Per Joint Position Error (PA-MPJPE), 2D metrics such as Percentage of Correct Keypoints (PCK), and Computational Complexity.

Table 2: We report reconstructions evaluated in 3D: Reconstruction errors (in mm) on the 3DPW dataset. Lower ↓ is better. The top three colors range from dark to light.

Method	MPJPE↓	PA-MPJPE↓
(Kanazawa et al., 2019)	116.5	72.6
VIBE (Kocabas et al., 2020)	93.5	56.5
TCMR (Choi et al., 2021)	95.0	55.8
HMR (Kanazawa et al., 2018)	130.0	76.7
I2L-MeshNet (Moon and Lee, 2020)	100.0	60.0
PyMAF (Zhang et al., 2021)	92.8	58.9
Pose2Mesh (Choi et al., 2020)	89.5	56.3
ROPM (Song et al., 2020)	91.3	54.9
PIXIE (Feng et al., 2021)	91.0	61.3
Hand4Whole(Moon et al., 2022)	86.6	54.4
ProHMR (Kolotouros et al., 2021)	-	59.8
OCHMR(Khirodkar et al., 2022)	89.7	58.3
HMR2.0 (Goel et al., 2023)	81.3	54.3
GMA-based (Ours)	80.8	54.6

Our GMA-based model significantly reduces parameter count by 97.3% (from 620M to 17M) and FLOPs count by 96.1% (from 242.1G to 9.5G) compared to previous baseline methods (Goel et al., 2023), and achieving superior efficiency in both parameters and FLOPs over existing CNN-based and Transformer-based approaches, see table 6.

While achieving the highest efficiency, our model also delivers competitive results on the benchmark datasets. On the 3DPW dataset (shown in table 2), compared to the baseline model, we reduced MPJPE (lower ↓ is better) by 0.5mm, with PA-MPJPE (lower ↓ is better) misses baseline model by only 0.3mm, demonstrating highly competitive performance. On the Human3.6M dataset (shown in table 3), our MPJPE matches the baseline, while PA-MPJPE misses baseline by 2mm. However, as shown

Table 3: We report reconstructions evaluated in 3D: Reconstruction errors (in mm) on the Human3.6M val split dataset. Lower ↓ is better.

Method	MPJPE↓	PA-MPJPE↓
(Kanazawa et al., 2019)	-	56.9
VIBE (Kocabas et al., 2020)	65.9	41.5
TCMR (Choi et al., 2021)	62.3	41.1
HMR (Kanazawa et al., 2018)	88.0	56.8
I2L-MeshNet (Moon and Lee, 2020)	55.7	41.1
PyMAF (Zhang et al., 2021)	57.7	40.5
Pose2Mesh (Choi et al., 2020)	64.9	46.3
ROPM (Song et al., 2020)	-	-
PARE (Kocabas et al., 2021)	76.8	50.6
ProHMR (Kolotouros et al., 2021)	-	41.2
THUNDR (Zanfir et al., 2021)	55.0	39.8
Mesh Graphormer(Lin et al., 2021b)	51.2	34.5
METRO (Lin et al., 2021a)	54.0	36.7
PyMAF-X (Zhang et al., 2022)	54.2	37.2
VisDB (Yao et al., 2022)	51.0	34.5
VirtualMarker (Ma et al., 2023)	-	32.0
HMR2.0 (Goel et al., 2023)	50.0	32.4
GMA-based (Ours)	50.0	34.4

Table 4: We report reconstructions evaluated in 2D: PCK scores of projected keypoints at different thresholds on the COCO validation set. Higher ↑ is better.

Method	PCK@0.05↑	PCK@0.1↑
PyMAF (Zhang et al., 2021)	0.68	0.86
PARE (Kocabas et al., 2021)	0.72	0.91
CLIFF (Li et al., 2022)	0.63	0.88
PyMAF-X (Zhang et al., 2022)	0.79	0.93
HMR2.0 (Goel et al., 2023)	0.86	0.96
GMA-based (Ours)	0.83	0.95

Table 5: Performance comparison on 2D dataset: Posetrack validation set. Higher ↑ is better.

Method	PCK@0.05↑	PCK@0.1↑
PyMAF (Zhang et al., 2021)	0.77	0.92
PARE (Kocabas et al., 2021)	0.79	0.93
CLIFF (Li et al., 2022)	0.75	0.92
PyMAF-X (Zhang et al., 2022)	0.85	0.95
HMR2.0 (Goel et al., 2023)	0.90	0.98
GMA-based(Ours)	0.87	0.96

in table 6, our model achieves these results with the fewest parameters and FLOPs, highlighting its exceptional efficiency. We also validated our model on 2D datasets: Both on the COCO validation set and PoseTrack, we achieved the second-best results (see table 4 and table 5). While the top performance belongs to the baseline SOTA model, our approach achieves a remarkable reduction in parameter count (97.3%) and FLOPs (96.1%), with only a slight performance gap of less than 3%.

Table 6: We compare the computational complexity with parameters and FLOPs with both CNN-based and Transformer-based methods. Input image size is 256×192 . We focus exclusively on the ViT-based backbones in HMR2.0 (Goel et al., 2023), Expose (Choutas et al., 2020) and OSX (Lin et al., 2023), excluding the influence from decoders. Similarly, for CNN-based methods, we compare the CNN-based backbones, ensuring a fair and focused evaluation of the backbone architectures.

Method	Parameter (M)	FLOPs (G)
Hand4Whole(Moon et al., 2022)	77.9	16.7
PIXIE (Feng et al., 2021)	192.9	34.3
ExPose (Choutas et al., 2020)	135.8	28.5
OXS (Lin et al., 2023)	102.9	25.3
PyMAF-X (Zhang et al., 2022)	205.9	35.5
HMR2.0 (Goel et al., 2023)	630	242.1
GMA-based (Ours)	17.0	9.5

4.3 Attention Mechanism Analysis

Given the input image, we then show the attention maps of the outputs from the attention branches $x[0]$ to $x[3]$ and non-attention branch in $x[4]$, shown in figure 2. From $x[1]$ 3×3 attention branch to $x[3]$ 7×7 attention branch in early stage[0], our model gradually expands the focus to slightly larger areas, showing more contiguous body parts, useful for capturing

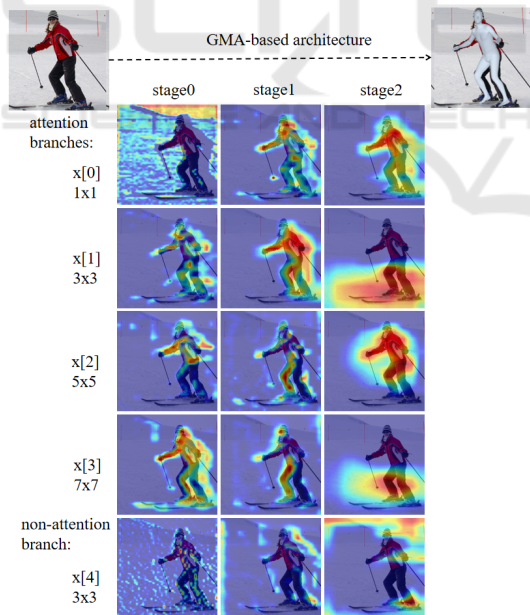


Figure 2: Attention visualization. Visualization shows that which part of the input data a model is focusing on. The model focuses more on the pixels highlighted in red compared to the other pixels. From $x[0]$ to $x[3]$, we show the attention maps with varying group sizes in the attention branches, specifically branches with group sizes of 3, 5, and 7, as well as the identity mapping branch in $x[0]$. In $x[4]$, we show the maps in the non-attention branch.

the relationship between adjacent keypoints and helps in understanding the spatial configuration. $x[4]$ 3×3 non-attention branch serves as a control, showing how the model behaves without specific attention mechanisms, which appears more dispersed and less focused. In stage[1], almost every attention branch focuses on the human body. While we observe that the $x[1]$ 1×1 branch emphasizes small regions, such as the hands and face. However, the $x[2]$ 5×5 and $x[3]$ 7×7 branches focus more on the features of larger areas, such as the entire leg region. In stage[2], the $x[0]$ 1×1 branch continues to focus on small regions, such as features around the shoulder area, while the $x[1]$ 3×3 , $x[2]$ 5×5 , and $x[3]$ 7×7 branches capture features from increasingly larger areas: below the knees for $x[1]$, above the knees for $x[2]$, and specifically around the knee region for $x[3]$.

We observe that applying a self-attention mechanism on pairwise tokens sometimes fails to focus on the human in early stage, but rather attends to the background, as shown in stage[0] of $x[0]$. In this scenario, when computing correlations between groups, using aggregations with kernel sizes of 3, 5, and 7 proves effective in centering on the human and attends to larger regions of the body. Our findings indicate that accurately estimating human pose and shape requires considering all elements together to capture broader characteristics. Emphasizing larger body regions contributes to a better understanding of human spatial configuration, which is essential for our needs.

4.4 Qualitative Result

We present a qualitative comparison between our GMA-based model and the baseline HMR 2.0 as shown in figure 3. Our model demonstrates superior robustness in handling blur+occlusion scenarios, accurately reconstructing arm and leg poses where the baseline often struggles. The occasional slight misalignment in highly complex poses may stem from joint ambiguity or overlapping limbs. In general blur or occlusion scenarios, our model achieves comparable performance to the baseline, which is why no dashed boxes are highlighted in those cases. Overall, the reduction in computational complexity does not compromise the visual quality of the results.

5 CONCLUSION

In this paper, we introduce Group-Mix Attention (GMA) into a fully-transformer framework for 3D human pose and shape estimation, capturing multi-level relationships—token-to-token, token-to-group,



Figure 3: Qualitative comparison with baseline. For each example we show comparison in complex pose, blur, occlusion, and blur+occlusion. We have highlighted the best reconstruction results and the failure cases. Our model generally outperforms the baseline (Goel et al., 2023) in scenarios with occlusion+blur, accurately reconstructing arm and leg poses where the baseline often struggles, particularly with elbows and knees. However, we observed that in highly complex poses, our model occasionally shows slightly lower alignment compared to the baseline.

and group-to-group—enabling a detailed spatial representation of body parts. Our method achieves high efficiency while delivering competitive performance on key metrics. We believe our model offers an optimal balance between computational efficiency and model accuracy, making it effective and interpretable for complex 3D pose-related tasks. Future work could enhance attention visualization and token grouping analysis to improve interpretability and reveal deeper insights into spatial relationship modeling.

REFERENCES

- Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., and Schiele, B. (2018). PoseTrack: A Benchmark for Human Pose Estimation and Tracking. In *CVPR*.
- Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *CVPR*.
- Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., and Black, M. J. (2016). Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *ECCV*.
- Choi, H., Moon, G., Chang, J. Y., and Lee, K. M. (2021). Beyond Static Features for Temporally Consistent 3D Human Pose and Shape from a Video. In *CVPR*.
- Choi, H., Moon, G., and Lee, K. M. (2020). Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose. In *ECCV*.
- Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., and Black, M. J. (2020). Monocular Expressive Body Regression Through Body-Driven Attention. In *ECCV*.
- Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., and Guo, B. (2022). CSwin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In *CVPR*.
- Fan, Q., Huang, H., Chen, M., Liu, H., and He, R. (2024). RMT: Retentive Networks Meet Vision Transformers. In *CVPR*.
- Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., and Black, M. J. (2021). Collaborative Regression of Expressive Bodies Using Moderation. In *3DV*.
- Ge, C., Ding, X., Tong, Z., Yuan, L., Wang, J., Song, Y., and Luo, P. (2023). Advancing Vision Transformers with Group-Mix Attention. *ArXiv*.

- Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., and Malik, J. (2023). Humans in 4D: Reconstructing and Tracking Humans with Transformers. In *ICCV*.
- Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al. (2018). AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *CVPR*.
- Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. (2018). End-to-End Recovery of Human Shape and Pose. In *CVPR*.
- Kanazawa, A., Zhang, J. Y., Felsen, P., and Malik, J. (2019). Learning 3D Human Dynamics from Video. In *CVPR*.
- Khirodkar, R., Tripathi, S., and Kitani, K. (2022). Occluded human mesh recovery. In *CVPR*.
- Kocabas, M., Athanasiou, N., and Black, M. J. (2020). VIBE: Video Inference for Human Body Pose and Shape Estimation. In *CVPR*.
- Kocabas, M., Huang, C.-H. P., Hilliges, O., and Black, M. J. (2021). PARE: Part Attention Regressor for 3D Human Body Estimation. In *ICCV*.
- Kolotouros, N., Pavlakos, G., and Daniilidis, K. (2019). Convolutional Mesh Regression for Single-Image Human Shape Reconstruction. In *CVPR*.
- Kolotouros, N., Pavlakos, G., Jayaraman, D., and Daniilidis, K. (2021). Probabilistic Modeling for Human Mesh Recovery. In *ICCV*.
- Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., and Lu, C. (2021). Hybrik: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation. In *CVPR*.
- Li, Z., Liu, J., Zhang, Z., Xu, S., and Yan, Y. (2022). CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation. In *ECCV*.
- Lin, J., Zeng, A., Wang, H., Zhang, L., and Li, Y. (2023). One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer. In *CVPR*.
- Lin, K., Wang, L., and Liu, Z. (2021a). End-to-End Human Pose and Mesh Reconstruction with Transformers. In *CVPR*.
- Lin, K., Wang, L., and Liu, Z. (2021b). Mesh Graphormer. In *ICCV*.
- Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In *ECCV*.
- Ma, X., Su, J., Wang, C., Zhu, W., and Wang, Y. (2023). 3D Human Mesh Estimation from Virtual Markers. In *CVPR*.
- Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., and Theobalt, C. (2017). Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. In *3DV*.
- Moon, G., Choi, H., and Lee, K. M. (2022). Accurate 3D Hand Pose Estimation for Whole-Body 3D Human Mesh Estimation. In *CVPR*.
- Moon, G. and Lee, K. M. (2020). I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image. In *ECCV*.
- Song, J., Chen, X., and Hilliges, O. (2020). Human Body Model Fitting by Learned Gradient Descent. In *ECCV*.
- Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., and Schmid, C. (2018). BodyNet: Volumetric Inference of 3D Human Body Shapes. In *ECCV*.
- Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. In *NeurIPS*.
- Von Marcard, T., Henschel, R., Black, M. J., Rosenhahn, B., and Pons-Moll, G. (2018). Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. In *ECCV*.
- Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., Wang, Y., and Wang, Y. (2017). AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding. *ArXiv*.
- Xu, Y., Zhang, J., Zhang, Q., and Tao, D. (2022). Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584.
- Xue, Y., Chen, J., Zhang, Y., Yu, C., Ma, H., and Ma, H. (2022). 3D Human Mesh Reconstruction by Learning to Sample Joint Adaptive Tokens for Transformers. In *ACM*.
- Yao, C., Yang, J., Ceylan, D., Zhou, Y., Zhou, Y., and Yang, M.-H. (2022). Learning Visibility for Robust Dense Human Body Estimation. *ArXiv*.
- You, Y., Liu, H., Wang, T., Li, W., Ding, R., and Li, X. (2023). Co-Evolution of Pose and Mesh for 3D Human Body Estimation from Video. In *ICCV*.
- Zanfir, A., Marinoiu, E., and Sminchisescu, C. (2018). Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes: The Importance of Multiple Scene Constraints. In *CVPR*.
- Zanfir, M., Zanfir, A., Bazavan, E. G., Freeman, W. T., Sukthankar, R., and Sminchisescu, C. (2021). THUNDR: Transformer-Based 3D Human Reconstruction with Markers. In *ICCV*.
- Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., and Liu, Y. (2022). PyMAF-X: Towards Well-Aligned Full-Body Model Regression From Monocular Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., and Sun, Z. (2021). PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop. In *ICCV*.