Machine Perception Report [Learners]

Adrià López Escoriza David Meyer Fatih Oezdemir

ABSTRACT

We present a method for human pose estimation based on SMPL parameters, addressing the challenges posed by occlusions and capturing long-range dependencies between body parts. Building upon the Part Attention REgressor (PARE) technique, which utilizes a part attention mechanism, we enhance part attention modeling by integrating a self-attention mechanism. By leveraging the Skinned Multi-Person Linear (SMPL) model, our approach accurately estimates the 3D human model and joint positions from RGB images. Through selective focus on informative body parts and their spatial relationships, our method effectively handles occlusions and captures holistic dependencies.

1 INTRODUCTION

Human pose estimation, the task of inferring the 3D body model and joint positions from RGB images, is a challenging problem in computer vision. Accurate pose estimation is crucial for various applications, but it is hindered by occlusions and the complexity of human body articulation. To address this, the Skinned Multi-Person Linear (SMPL) model has been widely used, providing a parametric representation of human body shape and pose. However, estimating SMPL parameters from RGB images remains a non-trivial task.

Part Attention REgressor (PARE) [4] is an effective method that tackles occlusion-related challenges in pose estimation. PARE incorporates a part attention mechanism to handle occlusions that commonly occur in RGB images. This mechanism selectively focuses on informative body parts, enabling robust pose estimation. However, PARE relies on deconvolution networks that struggle to capture long-range dependencies between different body parts, limiting its performance.

In [5], the authors propose the use of the Graphormer [7] to combine the long-term tracking capabilities of the transformer with the sensibility of a Graph Neural Network. In this work, we build on top of [4] and draw inspiration from [5] to propose an approach that enhances part attention modeling by leveraging a self-attention mechanism.

2 METHOD

Our proposed method for human pose estimation builds upon the Part Attention REgressor (PARE) technique and enhances its performance by incorporating a Multi-Head Self-Attention mechanism. In this section, we provide a detailed description of the implementation of PARE and the integration of the Multi-Head Self-Attention mechanism.

2.1 Datasets used

We use the 3D People in the Wild (3DPW) [6] and the MPII Human Pose dataset [2]. The 3DPW dataset consists of 60 video sequences with among others 2d and 3d pose annotations. The MPII Human Pose dataset features 25k images from online videos with one or more people and 40k 2d-pose annotated people.

2.2 Implementation of PARE

PARE (Part Attention REgressor) is a method specifically designed to address two key aspects in human pose estimation: attending to meaningful regions and modeling body part dependencies. The approach leverages a pixel-aligned structure, where each pixel in the image corresponds to a region and stores a pixel-level representation as a feature volume.

PARE incorporates two separate feature extraction branches within its architecture. The first branch, denoted as P, focuses on 2D part attention and estimates attention weights for each body part and a background mask. The feature volume P has dimensions $(H, W, J+1)$, where H and W represent the height and width of the feature volume, and J is the number of body parts. Each pixel (h, w) in P stores the likelihood of belonging to a specific body part, denoted as P_j.

The second branch, denoted as F, is responsible for 3D body parameter estimation. It has the same spatial dimensions as P but with a different number of channels, denoted as C. Each channel F_c in F contributes proportionally to the final feature tensor, denoted as F'. This contribution is determined by the corresponding elements in P_j after applying spatial softmax normalization (σ). The calculation of the element at location (j, c) in F' is given by:

$$F'_{j,c} = \sum_{h,w} \sigma(P_j)[h, w] \cdot F_c[h, w]$$

Here, $\sigma(P_j)$ serves as a soft attention mask that aggregates the features in F_c, with the element-wise multiplication denoted by the Hadamard product ($\odot$).

The operation of aggregating features using soft attention masks can be efficiently implemented as a dot product, similar to existing attention mechanisms. This can be represented as:

$$F' = \sigma(\tilde{P}) \odot \tilde{F}$$

where $\tilde{P}$ and $\tilde{F}$ denote the reshaped feature volumes P and F, respectively, omitting the background mask. The reshaped P and F have dimensions $(H \times W \times J)$ and $(H \times W \times C)$, respectively. By employing spatial softmax normalization and soft attention masks, PARE effectively attends to meaningful regions and models the dependencies between body parts, contributing to improved accuracy in human pose estimation.

2.3 Integration of Multi-Head Self-Attention

After the part estimation branch, the proposed method incorporates a self-attention mechanism to capture dependencies across the different part heatmaps and correct them. This self-attention mechanism operates at the image level and facilitates the modeling of long-range relationships between body parts.

Given the part heatmaps obtained from the previous stage, denoted as $H \in \mathbb{R}^{H \times W \times J}$, where H represents the height and width of the heatmaps and J is the number of body parts, the self-attention mechanism computes attention weights across the part heatmaps

to identify the importance of each body part in the context of the entire image.

To calculate the attention weights, the part heatmaps H are first flattened into a 2D tensor of shape $(HW) \times J$, denoted as $H_f \in \mathbb{R}^{(HW)\times J}$. Then, two learnable linear transformation matrices, $W_q \in \mathbb{R}^{J\times D_k}$ and $W_k \in \mathbb{R}^{J\times D_k}$, are applied to H_f to obtain the query matrix $Q \in \mathbb{R}^{(HW)\times D_k}$ and the key matrix $K \in \mathbb{R}^{(HW)\times D_k}$, respectively. The attention weights are computed by taking the dot product between the query and key matrices, scaled by the square root of the key dimension D_k, followed by applying the softmax function. The attention matrix $A \in \mathbb{R}^{(HW)\times(HW)}$ is given by:

$$A = \text{softmax}(\frac{QK^T}{\sqrt{D_k}})$$

where A_{ij} represents the attention weight between the i-th and j-th pixel in the part heatmaps. Next, the attention matrix A is used to correct the part heatmaps H by attending to informative regions and emphasizing the relevant body parts. The corrected part heatmaps, denoted as $\tilde{H} \in \mathbb{R}^{H\times W\times J}$, are obtained by applying the attention matrix A to the original part heatmaps H as follows:

$$\tilde{H} = A \odot H$$

The resulting corrected part heatmaps $\tilde{H}$ capture the long-range dependencies between different body parts and provide a refined representation that integrates global context information. This enables the model to have a more comprehensive understanding of the spatial relationships among body parts and facilitates accurate human pose estimation.

2.4 Final SMPL parameters derivation

To obtain the final outputs, the corrected heatmaps $\tilde{H}$ are combined with the SMPL 3D parameters, following the same approach as described in the PARE paper. The combination is performed through a linear regression process that maps the corrected heatmaps to the SMPL parameters.

$$F'_{j,c} = \sum_{h,w} \sigma(\tilde{H}_j)[h,w] \cdot F_c[h,w]$$

Then, each F_j parameter is passed through independent MLPs that extract each one of the pose and shape parameters (θ, β) that serve as inputs to the SMPL model $\mathcal{M}(\theta, \beta)$. To better understand the SMPL parameters and its impact on human body poses, we additionally experienced with the open source SMPL body-visualizer implemented in [1].

2.5 Loss function

The loss is computed as in the PARE paper by taking a weighted sum of the 3D joint regression loss, the 2D joint regression loss and the SMPL parameters loss (only available in the 3DPW dataset).

Overall, our total loss is:

$$L = \lambda_{3D}L_{3D} + \lambda_{2D}L_{2D} + \lambda_{SMPL}L_{SMPL}$$

where each term is calculated as: $L_{3D} = \frac{1}{F}\sum_{j=1}^{J}(J_{3D} - \hat{J}_{3D})^2$, $L_{2D} = \frac{1}{F}\sum_{j=1}^{J}(J_{2D} - \hat{J}_{2D})^2$, $L_{SMPL} = \frac{1}{2}|\Theta - \hat{\Theta}|_2^2$.

For further loss function design, we aim to build up on the insight obtained when visualizing the SMPL parameters. Analogously to Bogo et al. [3], for future work we plan to add a loss that further penalizes illegal joint angles.

Figure 1: "Sample training images generated."

3 EVALUATION

To compare the performance of our method with the vanilla PARE method, we evaluated them using the Mean Per Joint Position Error (MPJPE) and the Procrustes-aligned MPJPE (PA-MPJPE) metrics. These metrics provide a measure of the accuracy in estimating 3D human pose. We trained both a Vanilla PARE approach as described in [4] and our self-attention-enhanced mechanism using the same hyperparameters and datasets. During training, we monitored the validation error, calculated as the MPJPE and PA-MPJPE, over multiple epochs. As shown in figure 2, the convergence of both approaches regarding the validation loss showed similar behavior with the final loss value converging to a valiadion loss of around 80. We can demonstrate a better on PA-MPJPE when comparing our results with the validation values obtained in PARE as reported by the authors. This is however not true for the PA-MPJE. To provide a subjective impression of the achieved results, figure 2 shows a sample training output obtained after 75k iterations of training.

In conclusion, we were able to demonstrate the effectiveness of our proposed method and its potential in reducing pose estimation errors compared to the vanilla PARE method. The reason for the difference in achieved MPJE score requires further investigation. As pointed out by the PARE authors this might indicate errors regarding global orientations [4].

Validation performance		
	MPJE	PA-MPJPE
Our Approach	138.5	**81.5**
PARE (2021) (3DPW, ResNet-50)	**93.4**	93.9

4 CONCLUSION

Our preliminary results suggest that our proposed method shows promising performance improvements compared to the vanilla PARE method for 3D human pose estimation. However, further evaluations and experiments are necessary to validate these findings on a larger scale and across diverse datasets. It is important to conduct more comprehensive benchmarking and comparative studies to gain a deeper understanding of the strengths and limitations of our method. Future work could explore several directions. Firstly, it would be valuable to investigate the generalizability of our method by evaluating its performance on additional datasets and challenging scenarios. This would provide a more comprehensive assessment of its robustness and applicability in real-world

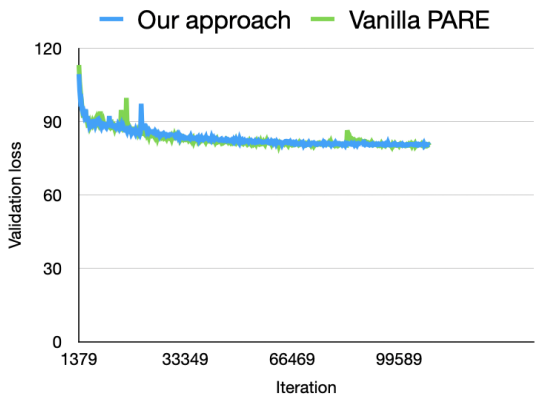

Figure 2: "Validation loss achieved during training runs vs. Vanilla PARE"

settings. Additionally, refining the architecture and exploring different hyperparameter settings could potentially further enhance the performance of our method. Finally, additional loss functions such as the already mentioned illegal joint angle penalization could bring further performance increases.

REFERENCES

[1] 2022. Body Model Visualizer. https://github.com/mkocabas/body-model-visualizer

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*. 3686–3693.

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 561–578.

[4] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. 2021. PARE: Part Attention Regressor for 3D Human Body Estimation. In *Proc. International Conference on Computer Vision (ICCV)*. 11127–11137.

[5] Kevin Lin, Lijuan Wang, and Zicheng Liu. 2021. Mesh Graphormer. In *ICCV*.

[6] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*. 601–617.

[7] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do Transformers Really Perform Badly for Graph Representation?. In *Thirty-Fifth Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=OeWooOxFwDa