Technical Report for EgoExo4D Body Pose Challenge from Team SJTU-SEIEE

Congsheng Xu[1], Jinfan Liu[1], Yifan Liu[1], Shuwen Wu[1], Yichao Yan[1], Weiming Zhao[1]
[1]Shanghai Jiao Tong University
acondaway@sjtu.edu.cn

Abstract

We propose a new solution in EgoExo4D Body Pose Challenge. We utilize the Level-wise Cross Attention ViT as our model to deal with the body pose estimation task. Our method achieves 18.09 MPJPE and 0.62 MPJVE on the Challenge set.

1. Overview

In this challenge, we combine the strengths of each of the two structures of the ViT into one model. In detail, we use the baseline ViT method as a coarse-grained estimation block and the Huge-ViT method as a fine-grained estimation block. And to better gain both local and global information, we especially design the MLP of each ViT block. And we finally combine the estimation result of each block by leveraging a cross attention weight block to acquire the estimation results. We will describe the details of our method and detailed parameter setting in the following sections.

2. Method

As shown in Figure 1, after forming the embeddings, we input the embeddings into the coarse-grained and fine-grained block in parallel. In coarse-grained design, we basically utilize the ViT with nhead-8 and layer-3 according to the official baseline instance of Ego-Exo4D Body-pose Baseline Model [2]. In fine-grained design, we choose Huge-ViT structure with nhead-16 and layer-32 [1]. With both ViTs, the global information can be captured by the coarse-grained block and the local information can be mined by the fine-grained block. To perform better in fine-grained process, we especially augment the MLP structure of the fine-grained block. And with both coarse- and fine-grained estimation results computed, we set a weight on both results. In our final submission, we choose the 0.1 on coarse result and 0.9 on fine result.

Before we work on the structure above, we have tried, 1) Only instance ViT. 2) Only Huge-ViT. We find that the former performs well in the scenarios that require little fine-grained detection such as Cooking, Health and Music, but really awfully in the scenarios that require meticulous detection such as Bike-repair, Dance and Soccer. The Basketball series are just in the middle. Aber the latter can partly focus on the meticulous detection task while find it unstable in other scenarios' estimation. So we spontaneously ponder the combination of each ViT to both boost the precision of the coarse-grained model and alleviate the unstableness of the fine-grained model.

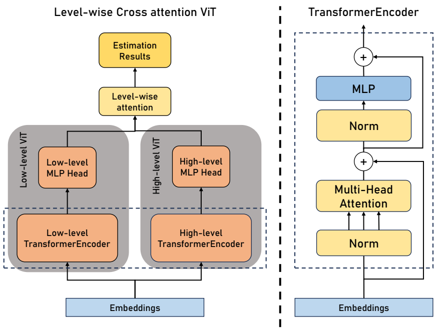

Figure 1. Pipeline of our method.

3. Result

Results are shown in Table 1. Baseline(online) represents the baseline metrics provided by Ego-Exo4D Body-pose [2]. Baseline(10000G) represents the coarse-grained model's single perfomance and the Huge-ViT Only is clearly explained. The Level-wise ViT(local 10000G) is the exact model we propose. Empirically, the model we propose performs better than the baseline model in all metrics. But as a matter of fact, the MPJPE on the Soccer scenario is really awful. There should be a method to especially train these takes to integrate a better model. And for the reason of time, we are unable to train our model for enough epoches and iterations, the performance seems to be better if trained sufficiently with better convergence.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[2] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives, 2024. 1

Result Model	MPJPE(↓)	MPJPE Physical(↓)	MPJPE Procedural(↓)	MPJVE(↓)
Baseline(online)	18.51	19.97	12.67	0.64
Baseline(local 10000G)	18.56	19.89	113.16	0.62
Huge-ViT Only(local 10000G)	18.28	19.79	12.15	0.61
Level-wise ViT(local 10000G)	**18.09**	**19.59**	**11.92**	**0.62**

Table 1. Experiment results