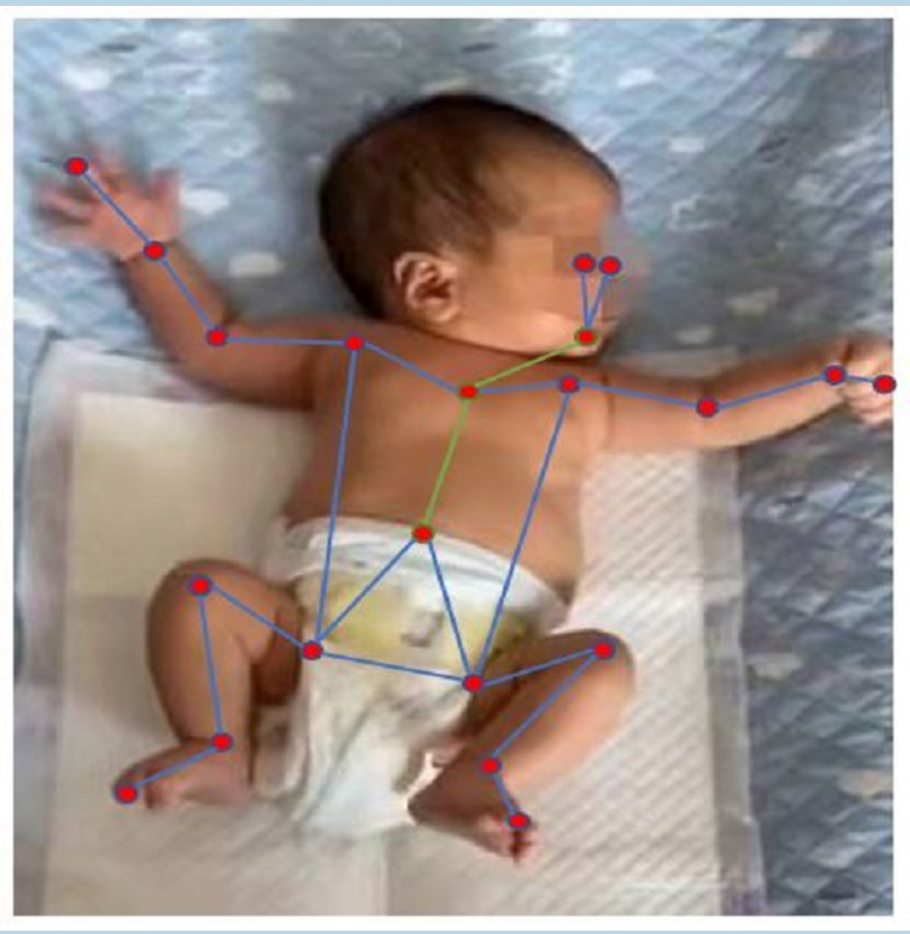
AggPose: Deep Aggregation Vision Transformer for Infant Pose Estimation

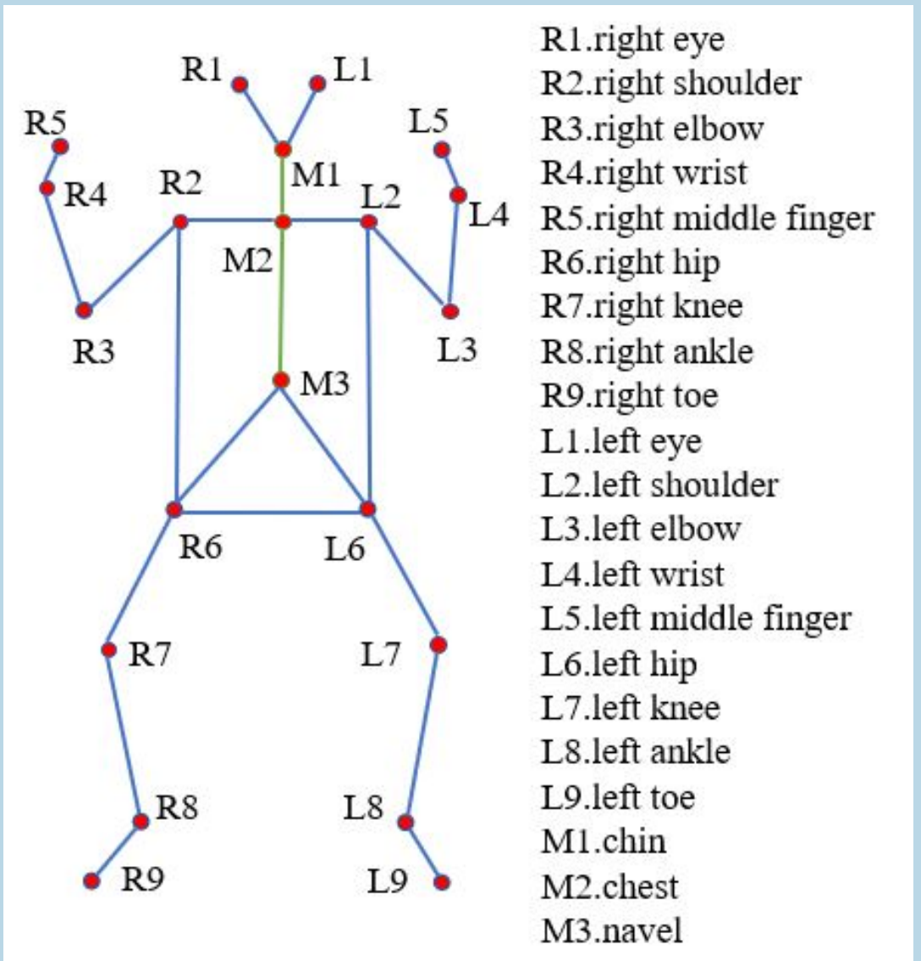
Xu Cao, Xiaoye Li, Liya Ma, Yi Huang, Xuan Feng, Zening Chen, Hongwu Zeng, Jianguo Cao

Shenzhen Automatic Rehabilitation Laboratory
xc2057@nyu.edu, caojgsz@126.com (Corresponding author)

Introduction

Movement and pose assessment of infants lets experienced medical doctors predict neurodevelopmental disorders, allowing early intervention for related diseases. However, most new AI approaches for human pose estimation focus on adults, lacking a public benchmark for infant pose. In this paper, we fill this gap by proposing an infant pose dataset and a new Transformer-based model without using convolution operations to extract features in the early stages. Our new framework generalizes Transformer with MLP to high-resolution deep layer aggregation, thus enabling information fusion between different vision levels. We pre-train our model on COCO pose dataset and apply it to our newly released large-scale infant pose estimation dataset. The results show that our model could effectively learn the multi-scale features among different resolutions and improve the performance of infant pose estimation. We show that our model outperforms the hybrid models such as HRFormer and TokenPose in COCO and our task.





Performance

We compare AggPose with several state-of-art methods, including transformer-based method HRFormer, TokenPose and TransPose in both COCO Key point dataset and our Infant Pose dataset. For input size of 256×192, AggPose is best among all methods. The model parameter has released to the github.

	AP	AR
TokenPose-L/D	75.8	80.9
HRFormer-B	75.6	80.8
AggPose-L	**76.4**	**81.3**

Table 1: Performance on COCO key point Dataset

	AP	AR
TokenPose-L/D	93.0	93.9
HRFormer-B	93.8	95.0
AggPose-L	**95.0**	**95.7**

Table 2: Performance on Infant Pose Dataset

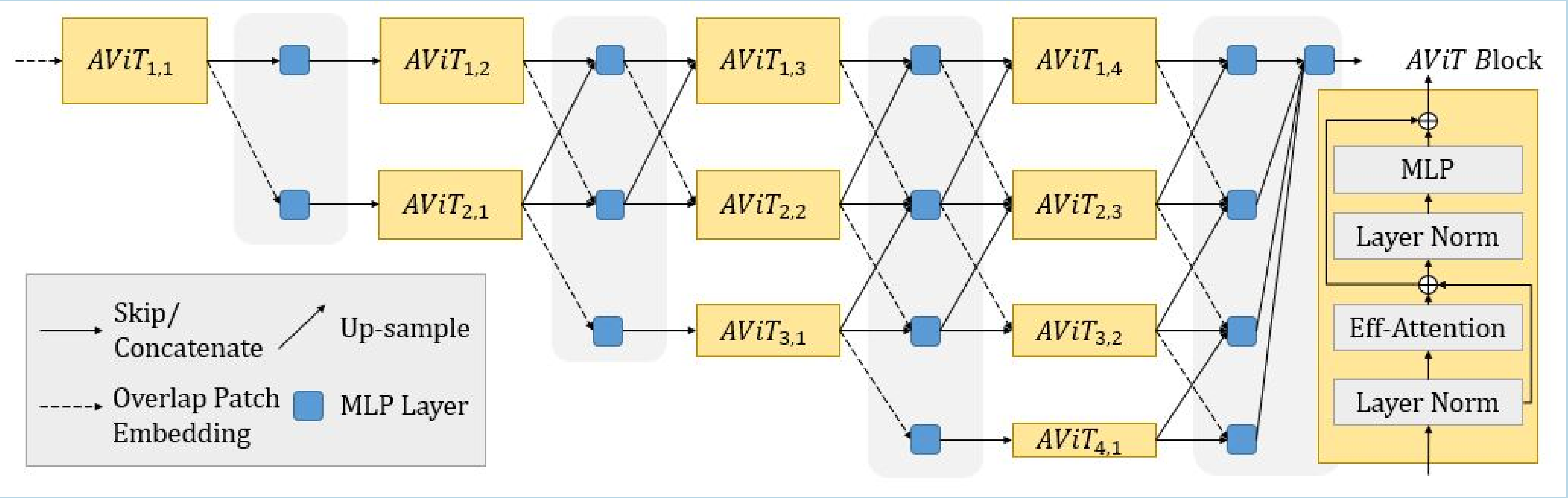
Associations

深圳市儿童医院
Shenzhen Children's Hospital
厚德 精医 团结 奉献

NEW YORK UNIVERSITY

Our Solution

InfantPose dataset We proposed a new 2D infant pose dataset. Dr. Xiaoye Li collects the GMA infant image data in Shenzhen Baoan Women's and Children's Hospital. To collect data, we adopt GMA devices to record infant movement videos. The age of the infant is between 0 to 1. More than 216 hours of videos were collected. Our dataset's size and scalability are much better than the MINI-RGBD dataset. We randomly sampled over 20,000 frames from the videos and let professional clinicians annotate infant key points. Then, we divided the dataset into 11,756 for the training and validation sets. Experienced pediatricians propose the 21 keypoint format for infant pose. Our dataset also reduces key points on infants' heads and comprises more refined body key points. The final version has added the face mosaic suggested by reviewers. The first version of the dataset is visible at https://szar-lab.github.io/AggPose/.



AggPose Architecture

Model Structure Our DL framework is a new Transformer-based model for human pose estimation. This figure shows the pipeline of the model. Inspired by SegFormer, we adopt a full Transformer with Overlapped Patch Embedding to replace HRNet's CNN feature extractor and down-sampling block of each stage. Compared with the early convolutions in hybrid models such as HRFormer, and TokenPose, Overlapped Patch Embedding can obtain better low-level features, enhancing the high-resolution Transformer's feature representation and reducing computation complexity.

About the self-attention module in each transformer block, we use the sequence reduction process referred to SegFormer, which significantly reduces the amount of calculation inside the Transformer and accelerates the convergence process during model training. We also adopt depth-wise convolution into the feed-forward network (FFN) to expand the receptive field and reduce the harmful effect caused by positional embedding.

$$K = Linear(\gamma C, C)(K.Reshape(\frac{N}{\gamma}, \gamma C)) \tag{1}$$

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_{head}}})V \tag{2}$$

K is the token representation with initial shape $N \times C$. γ is the reduction ratio that decrease the dimension of K from $N \times C$ to $N/\gamma \times C$.

In our model, we expand the usage of Mix-FFN into the deep aggregation approach across different resolution layers. In the connection part of different resolutions, the proposed cross-layer aggregation module consists of two main steps for each level. First, multi-level features from different resolutions go through a mixed feed-forward network with 3×3 depth-wise convolution to unify the channel dimension and upsample or downsample the feature map to the same shape. Then, we concatenate the feature vector from adjacent levels and adopt an additional FFN layer to fuse the information.

$$x_{i,j} = \begin{cases} OverlappedPE_{i,j}(FFN(x_i)) & i < j \\ x_i & i = j \\ Upsample_{i,j}(FFN(x_i)) & i > j \end{cases} \tag{3}$$

$$x_j = MixFFN(Concat(x_{j-1}, x_j, x_{j+1})) + x_j \tag{4}$$

$x_{i,j}$ is the input of aggregation MLP layer. x_i denotes the feature map from adjacent resolution. The cross-layer aggregation module apply Mix-FFN to merge adjacent resolution features.

Qualitative Result

The figure below shows the output feature map visualization for AggPose.