TokenHPE: Learning Orientation Tokens for Efficient Head Pose Estimation via Transformers

Cheng Zhang[1] Hai Liu[1,*] Yongjian Deng[2,3] Bochen Xie[4] Youfu Li[4,*]

[1]National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China

[2]College of Computer Science, Beijing University of Technology, Beijing, China

[3]Engineering Research Center of Intelligence Perception and Autonomous Control, Ministry of Education, Beijing, China

[4]Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong, China

zc2021@mails.ccnu.edu.cn, hailiu0204@ccnu.edu.cn, yjdeng@bjut.edu.cn,

boxie4-c@my.cityu.edu.hk, meyfli@cityu.edu.hk

This supplementary material includes the region partitioning details in Sec. 1. In Sec. 2 the model details of the Transformer blocks and MLP head are elaborated. We present additional ablation study experiments in Sec 3. Finally, we illustrate abundant visualizations to validate the token learning capacity of our model in Sec. 4.

1. Region Partitioning Details

We illustrate here the partitioning details of two panoramic overviews. The panoramic overview is divided into several basic regions according to neighbor orientation similarities and spatial opposite symmetry. In strategy I, we divide the panoramic overview into nine basic orientation regions. In yaw direction, we set 60° and -60° as the division degree because of the appearance (or disappearance) of eyes. In pitch direction, we set 30° and -30° as the division degree because of the appearance (or disappearance) of the nostril and the overlapping of nose and mouth. As such, the nine basic orientation regions in strategy I are: (0) upper left, (1) top, (2) upper right, (3) middle left, (4) middle, (5) middle right, (6) bottom left, (7) bottom, and (8) Bottom right. As shown in Fig. 1, head poses in the same region are similar, and the opposite head poses are symmetric.

In strategy II, we divide the yaw direction in a finer-granularity because the significant facial part changes are complex when pitch angle is little. As shown in Fig. 2, in this partition strategy, the middle area of the panoramic overview is divided into five basic regions. The division degree is set as 60° because of the complete disappearance of eye. We set 20° as the other division degree for the start of the disappearance of eye. Therefore, when the pitch angle is within -30° and 30°, the basic orientation regions are as follows: (3) middle left 1, (4) middle left 2, (5) middle, (6) middle right 1, and (7) middle right 2.

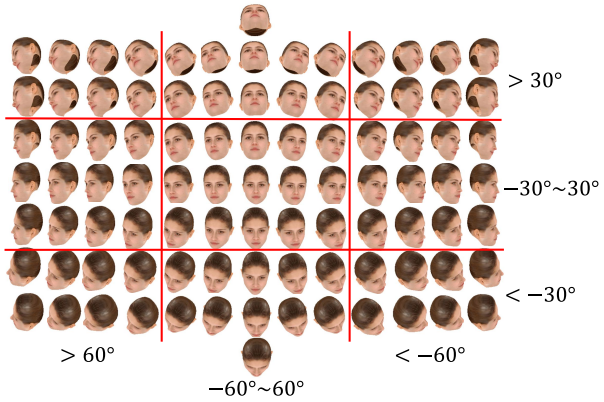

Figure 1. Partition strategy I with nine basic orientation regions.

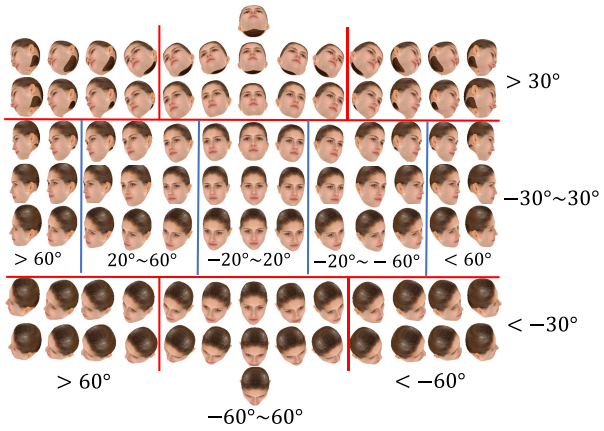

Figure 2. Partition strategy II with 11 basic orientation regions.

2. Transformer block and MLP Head Details

Transformer blocks. The Transformer module is constructed by stacking identical M blocks. The Transformer

*: Corresponding author (Hai Liu, Youfu Li)

blocks take the the [visual] and [dir] tokens as the input. Each block comprises a multi-head self-attention (MSA) module and a multi-layer perception (MLP) module, with a layer norm (LN) operation and skip connection added between the two modules. We modify the MLP module such that it is constructed by two linear projections, with a $Tanh(\cdot)$ activation function and dropout operations in between. Self-attention (SA) is defined as:

$$SA(X^l) = softmax\left(\frac{(X^l W_Q (X^l W_K)^T}{\sqrt{s}}\right)(X^l W_V), \tag{1}$$

where W_Q, W_K, and $W_V \in \mathbb{R}^{d\times d}$ represent the query matrix, the key matrix, and the value matrix. X^l is the output of the l-th Transformer layer. s is part of the scaling factor $\frac{1}{\sqrt{s}}$. In SA, s equals the dimension d of the tokens. MSA is an extension of SA with h self-attention operations, which are called heads. In MSA, s is typically set to d/h. Therefore, MSA can be formulated as:

$$MSA(X^l) = [SA_1(X^l); SA_2(X^l); ...; SA_h(X^l)]W_P, \tag{2}$$

where $W_P \in \mathbb{R}^{(h\cdot s)\times d}$. After defining MSA, the operations of a Transformer block can be expressed as:

$$\begin{cases} \tilde{X}^{l-1} = MSA[LN(X^{l-1})] + X^{l-1}, \\ X^l = MLP[LN(\tilde{X}^{l-1})] + \tilde{X}^{l-1}. \end{cases} \tag{3}$$

After the last Transformer layer, the [dir] tokens are selected as the output of Transformer, whereas the [visual] tokens are not used in the following steps. Therefore, the output of M Transformer blocks is denoted as $\mathcal{S} = \{X_1^M, X_2^M, ..., X_k^M\}$, where k is the number of [dir] tokens.

Gram–Schmidt process. The orientation tokens $\mathcal{S}$ need to be transformed to rotation matrices for training and prediction. To achieve this, each [dir] token is firstly applied with a linear projection to obtain a 6D representation of head pose. Next, the Gram–Schmidt process is applied to generate the 9D rotation matrix. This transformation is formulated as:

$$\hat{R}_i = F_{GS}(W X_i^M), \tag{4}$$

where W is the projection matrix, and $\hat{R}_i$ is the predicted rotation matrix of the i-th basic orientation region. $F_{GS}(\cdot)$ denotes the Gram–Schmidt process that can be expressed as:

$$F_{GS}(a_1, a_2) = [b_1, b_2, b_3], \tag{5}$$

where $a_1, a_2 \in \mathbb{R}^3$ are 3D column vectors of a rotation matrix. b_i is 3D column vector of the rotation matrix defined as:

Table 1. Ablation study on Transformer block hyperparameters, including token dimension, activation function in MLP module, and the number of heads in multi-head self-attention. The models are trained on 300W-LP dataset and tested on AFLW2000 dataset.

	Pitch	Yaw	Roll	MAE	MAEV
Token dimension					
64	5.70	4.54	4.29	4.85	6.04
128	**5.54**	**4.36**	**4.08**	**4.66**	**5.98**
386	5.64	4.45	4.22	4.77	6.00
Activation function in MLP module					
GELU	5.63	**4.35**	4.19	4.72	**5.98**
ReLU	5.71	4.43	4.28	4.80	6.01
Tanh	**5.54**	4.36	**4.08**	**4.66**	5.99
No. of heads in MSA					
8	5.77	4.46	4.32	4.85	6.05
12	**5.54**	**4.36**	**4.08**	**4.66**	**5.98**
16	5.62	4.42	4.25	4.76	5.99

$$\begin{cases} b_1 = \dfrac{a_1}{\|a_1\|}, \\ u_2 = a_2 - (b_1 \cdot a_2)b_1, \\ b_2 = \dfrac{u_2}{\|u_2\|}, \\ b_3 = b_1 \times b_2. \end{cases} \tag{6}$$

A set of rotation matrices $\mathcal{C} = \{\hat{R}_1, \hat{R}_2, ..., \hat{R}_k\}$ can be generated by the transformation above, where k is the number of orientation tokens.

3. Additional Ablation Studies

Transformer block parameters. In [1], Transformer parameters had significant effect on model performance. Therefore, we investigate different options of Transformer block parameters. The parameter being studied is varied, and other parameters are kept constant to token dimension of 64, GELU activation function, and 8 heads in MSA. The results are shown in Table 1. The best dimension of token is 128, the best activation function is Tanh, and best number of heads is 12.

Number of Transformer blocks. Different numbers of Transformer blocks are evaluated to check their effect on HPE. The experimental results are shown in Fig. 3. As the number of blocks increases, the MAE first decreases then increases. When the number of blocks is small, the model has less capacity to learn the complicated facial relationships. When the number of blocks is too large, the model is difficult to converge, thus resulting in the increase of MAE. The best result is achieved when the number of blocks is set to 12.

Number of orientation tokens. The basic orientation regions have various granularities. We propose two sets of basic orientation region granularities to compare the ef-

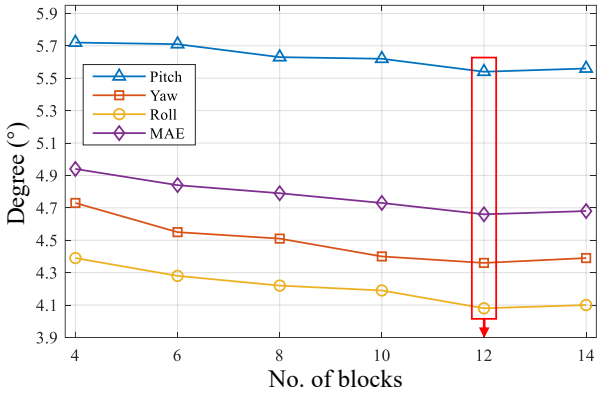

Figure 3. Effect of the number of Transformer blocks. The models are trained on 300W-LP dataset and tested on AFLW2000 dataset.

Table 2. Effect of the number of direction tokens. The models are trained on the 300W-LP dataset and tested on the AFLW2000 dataset. One [dir] token situation has no regional partition.

#[dir] tokens	Pitch	Yaw	Roll	MAE	MAEV
1 (holistic)	5.79	5.01	4.44	5.08	6.48
9	5.73	4.53	4.29	4.85	6.11
11	5.54	4.36	4.08	4.66	5.98

fect of the orientation tokens. For the first set, facial orientations are divided into nine basic regions in a 3×3 grid. For the other set, orientations in the yaw directions are further divided into five basic regions, resulting in eleven basic regions. In Table 2, we report the results of the two sets, in addition to the non-partition in the basic regions (#[dir]=1). In the situation where exists only one [dir] token, the model is similar to ViT [2] with one [cls] token. The results indicate that eleven [dir] tokens exhibit the best performance. This finding indicates that the eleven region partitions can effectively represent the basic orientations.

4. Extra Visualizations

Orientation token learning during training. We calculate the cosine similarity between the orientation tokens in different training epochs. As Fig. 4 shows, in early stages, no distinct relationship is learned by the orientation tokens. As the training epochs increase, general information is learned gradually by the orientation tokens. The orientation relationships can be observed in the later training epochs. In partition strategy I (nine basic orientation regions), take the middle left (region 3) orientation token in the 30th epoch for example. The similarity scores are higher in its neighborhood regions (upper left (region 0), bottom left (region 6)) and spatial symmetric regions, such as middle right (region 5). Similar results can be observed when the number of basic orientation regions is set to eleven. Visualization of orientation token learning in the training stage validates that general orientation information can be learned by the orientation tokens.

Heatmap visualization in the inference stage. Grad-CAM [3] is utilized to investigate the attention distribution of TokenHPE. The heatmaps in the 2nd, 4th, 6th, 8th, 10th, 12th Transformer blocks are visualized in Fig. 5. In the first few blocks, the model pays attention to elementary facial pattens across the face. As the network goes deeper, the crucial facial parts, such as eyes, nose, mouth and ear, have higher attention values than other parts. In the last few blocks, the model pays most attention to the critical minority facial parts to yield head pose predictions. Therefore, the heatmap visualization validates that our proposed TokenHPE can learn the critical minority relationships for high accuracy HPE.

References

[1] Naina Dhingra. Lwposr: Lightweight efficient fine grained head pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1495–1505, 2022. 2

[2] 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. 3

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 3

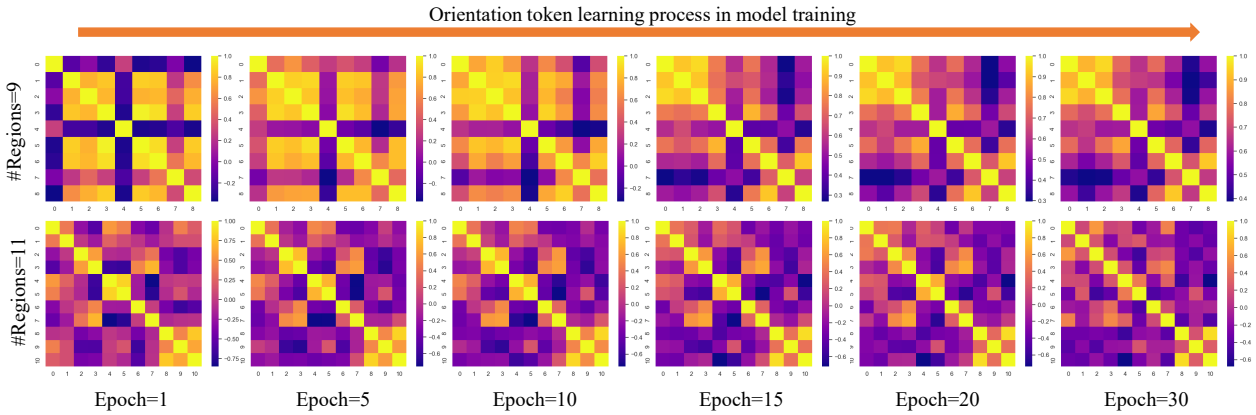

Figure 4. Cosine similarity matrix between orientation tokens during model training. The orientation information is learned gradually by the orientation tokens.

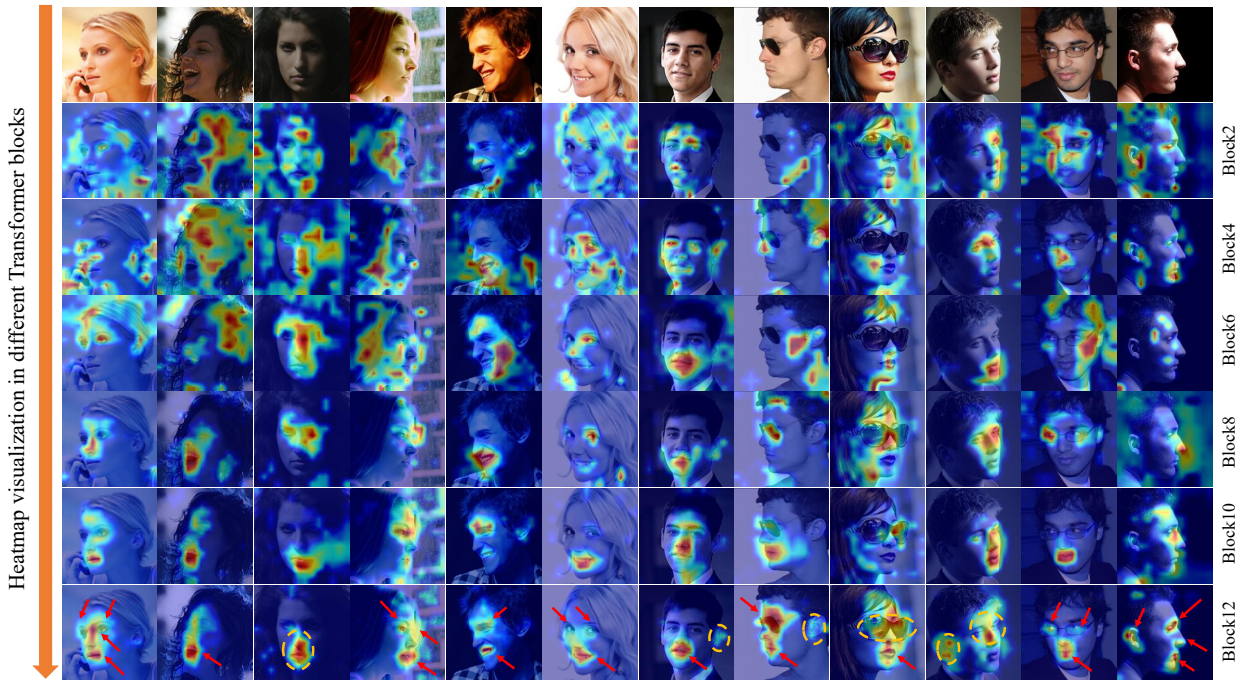

Figure 5. Heatmap visualization in the 2nd, 4th, 6th, 8th, 10th, 12th Transformer blocks of the TokenHPE model. Arrows and circles indicate the crucial facial parts to which the model pays attention for the head pose prediction.