Orientation Cues-Aware Facial Relationship Representation for Head Pose Estimation via Transformer

Hai Liu, *Senior Member, IEEE*, Cheng Zhang, *Student Member, IEEE*, Yongjian Deng, *Member, IEEE*, Tingting Liu, *Member, IEEE*, Zhaoli Zhang, *Senior Member, IEEE*, and You-Fu Li, *Fellow, IEEE*

***Abstract*— Head pose estimation (HPE) is an indispensable upstream task in the fields of human-machine interaction, self-driving, and attention detection. However, practical head pose applications suffer from several challenges, such as severe occlusion, low illumination, and extreme orientations. To address these challenges, we identify three cues from head images, namely, critical minority relationships, neighborhood orientation relationships, and significant facial changes. On the basis of the three cues, two key insights on head poses are revealed: 1) intra-orientation relationship and 2) cross-orientation relationship. To leverage two key insights above, a novel relationship-driven method is proposed based on the Transformer architecture, in which facial and orientation relationships can be learned. Specifically, we design several orientation tokens to explicitly encode basic orientation regions. Besides, a novel token guide multi-loss function is accordingly designed to guide the orientation tokens as they learn the desired regional similarities and relationships. Experimental results on three challenging benchmark HPE datasets show that our proposed TokenHPE achieves state-of-the-art performance. Moreover, qualitative visualizations are provided to verify the effectiveness of the token-learning methodology.**

***Index Terms*— Head pose estimation, attention mechanism, relationship perception, deep learning, transformer.**

I. Introduction

HEAD pose estimation (HPE) is a popular research area in image processing [1], [2], [3] and an indispensable upstream task in human-machine interaction [4], [5], [6], [7], driver assistance [8], virtual reality [9], [10], and attention detection [11]. In the past few years, the accuracy of HPE has been considerably improved in terms of utilizing extra facial landmark information [12], [13], extra RGB-depth information [14], [15], [16], [17], extra temporal information [18], stage-wise regression strategy [19], multitask learning [20], [21], and alternative parametrizations of orientation [22], [23], [24], [25], [26], [27]. However, several challenges still exist for practical application where occlusion, unstable illumination and extreme orientations are ubiquitous.

Fig. 1. Existing challenges on head pose estimation, including (a)–(b) serious occlusions, (c)–(d) poor illumination, and (e)–(f) extreme orientations. Some or even most of the facial parts are missing in these scenarios, resulting in difficulties for HPE.

A. Challenges

Currently, convolutional neural networks (CNNs) have become prevalent on computer vision tasks, and they are widely adopted on HPE. CNN-based HPE methods [19], [24], [27], [28], [29] have achieved impressive performance due to the powerful abilities of CNNs on representing superficial visual patterns. Nevertheless, the intrinsic relationships of head orientations and facial parts are usually neglected. A possible reason is that these relationships are theoretically difficult to learn by existing CNN architectures, which are based on pattern driven learning. In normal and easy-to-predict scenarios, highly accurate head pose predictions can be achieved by detecting facial patterns through CNNs. However, in some

Manuscript received 3 February 2023; revised 30 June 2023 and 21 September 2023; accepted 26 October 2023. Date of publication 14 November 2023; date of current version 20 November 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFC3340802; in part by the National Natural Science Foundation of China under Grant 62277041, Grant 62211530433, Grant 62177018, Grant 62173286, Grant 62077020, Grant 62005092, Grant 62203024, and Grant 92167102; and in part by the Research Grants Council of Hong Kong under Project CityU11213420 and Project CityU11206122. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Angeliki Katsenou. *(Corresponding authors: Cheng Zhang; You-Fu Li.)*

Hai Liu, Cheng Zhang, and Zhaoli Zhang are with the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China (e-mail: hailiu0204@ccnu.edu.cn; zc2021@mails.ccnu.edu.cn; zl.zhang@ccnu.edu.cn).

Yongjian Deng is with the College of Computer Science, Beijing University of Technology, Beijing 100124, China (e-mail: yjdeng@bjut.edu.cn).

Tingting Liu is with the School of Education, Hubei University, Wuhan, Hubei 430062, China, and also with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong (e-mail: tliu@hubu.edu.cn).

You-Fu Li is with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong (e-mail: meyfli@cityu.edu.hk).

Digital Object Identifier 10.1109/TIP.2023.3331309

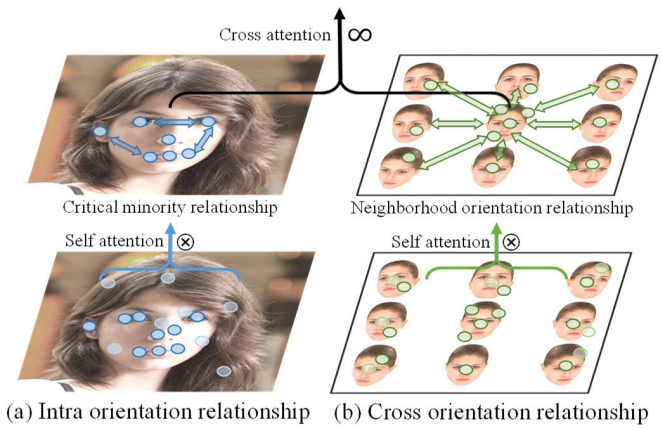


Fig. 2. Illustration of (a) intra-orientation relationship and (b) cross-orientation relationship. The critical minority relationships in a single image are deduced by self-attention among visual tokens, the neighborhood orientation information is encoded in orientation tokens, and their relationships are deduced by self-attention among orientation tokens. Finally, the intra- and cross-orientation relationships are exchanged by cross attention.

challenging scenarios (Fig. 1), such as severe occlusions, poor illumination, and extreme orientations, many remarkable facial parts are missing because of occlusion or low light, which is devastating for existing CNN-based methods that highly depend on facial patterns for prediction. Consequently, the few remaining facial parts and their geometric relationships must be leveraged to achieve robust and high-accuracy prediction. Furthermore, the latent relationships of neighborhood orientations also can be exploited when facial part missing happens in the current orientation. Recently, there are a stream of researches that explore Transformer-architecture as an alternative to CNN layers in their models and achieve compelling performance, nevertheless the main structures are still a CNN-style, leaving the long-range and semantic relationships virtually intact. Therefore, how to leverage the head orientation and facial part relationships is considerably attractive on the research topic of high accuracy and robust HPE.

B. Observation and Insights

For the purpose of utilizing the facial and orientational information to facilitate HPE, in this study, we identify three cues by careful observation. First, *critical minority relationships* of facial parts exist, and they can determine the orientation of a head pose despite possible occlusions and missing facial parts. For example, as shown in Fig. 1(a), if a person's eye is occluded, then the head orientation can be determined by the geometric spatial relationships of the remaining facial parts, such as ear, nose, and the outline of the face. In other words, by ingeniously leveraging the semantic relationships of the few remaining facial parts, accurate prediction can be achieved despite severe facial part missing. This important cue is defined as critical minority relationships. Second, a local similarity in neighbored orientation regions exists. As shown in the right part of Fig. 2, the facial appearances in neighbored orientations are similar, which indicates that neighborhood orientation information can be leveraged to improve accuracy. In a local orientational region, head poses and their corresponding latent facial characteristics enjoy high semantic similarities. Therefore, the neighborhood orientations contribute latent semantic information to the central orientation. Prediction can be facilitated by taking the neighborhood orientation information, which is defined as *neighborhood orientation relationships*. Third, several *significant facial part changes* are observed in specific orientations. For example, two facial regions can be distinguished by a significant facial part change, such as the appearance/disappearance of eye on one side, appearance of the nostril, and overlapping of the nose and mouth. The set of head orientations can be partitioned into several highly similar local facial regions by these significant facial part changes. Generally, we find three cues (critical minority relationships, neighborhood orientation relationships, and significant facial part changes) that veiled in head poses, which are necessary for efficient HPE on all scenarios.

Furthermore, on the basis of the three cues, we reveal two insights in head poses, namely, the intra-orientation relationship and cross-orientation relationship. We argue that the two novel insights on facial and orientational relationships are curial for efficient HPE from a different relationship-driven learning paradigm. The proposed two key insights are introduced as follows.

Key insight I: Intra-orientation relationship. There exists critical minority relationship in a specific head orientation (a single head image). The few facial parts and their relationships within a head image defined as intra-orientation relationships are crucial for prediction and more robust and reliable than merely superficial visual patterns. Figure 2(a) provides an illustration of the intra-orientation relationship. As can be observed, a single image has many informative facial patterns, but only the core facial parts and their relationships are determinative for prediction. On the basis of critical minority relationship learning, detriments of facial part missing from occlusion or poor illumination can be greatly alleviated.

Key insight II: Cross-orientation relationship. We argue that the vicinal and symmetric orientation characteristics are informative to the central orientation due to their high similarities. This property is defined as cross-orientation relationship, because interrelated orientational features are learned and for prediction. As show in Fig. 2(b), the attention is distributed to the vicinal orientational regions, allowing a larger reception field than in a single image to collect more orientational information for prediction. However, this general relationship cannot be encoded by CNN because of its inherent architecture. Therefore, the cross-orientation relationships have kept as untapped treasures that are hardly leveraged by previous works on HPE.

Given the aforementioned key insights on head pose images, the question is how to design a model that can utilize this heuristic knowledge. The traditional CNN architecture cannot easily learn these relationships. By contrast, the Transformer architecture can effectively address this drawback of CNN. Recently, Vision Transformer (ViT) [30] emerged as a new choice for various computer vision tasks. The Transformer architecture is known for its extraordinary ability to learn long-distance, high-level relationships between image patches. Therefore, using Transformer to learn the intra-orientation

relationship is reasonable. Moreover, cross-orientation relationships can be well-represented by learnable tokens in the Transformer.

C. Contributions

Inspired by the two key insights and Transformer's properties, this study proposes TokenHPE, a method that can discover and leverage intra-orientation and cross-orientation relationships via the Transformer architecture. The proposed method can discover facial part geometric relationships via self-attention among visual tokens, and the orientation tokens can encode the characteristics of the neighborhood orientation regions. These relationships between visual and orientation tokens are learned by TokenHPE from abundant synthetic data. The learned information is encoded into the orientation tokens, which can be visualized by vector similarities. In addition, a special token guide multi-loss function is constructed to help the orientation token learn the general information. Notice that although currently there are several Transformer-based approaches for HPE, the superior properties of Transformer architecture than CNN architecture, which is the capability to reveal the long-range and semantic relationships of the input token sequences, has not been exploited. The gap between them and our method is that they utilize the Transformer encoder layers as a supplementary module for the main CNN structure, while we take Transformer blocks as our core design. Specifically, we divide the input image into patches, considering them as visual tokens that contains semantic information analogous to "words" in natural language processing. Then, we proposed several learnable orientation tokens that represent the orientation knowledge to interact with the visual tokens in the Transformer blocks via attention mechanism. Overall, our main contributions can be summarized as follows:

- Three cues are derived on head images, including critical minority relationships, neighborhood orientation relationships, and significant facial part changes. Furthermore, to leverage our findings and cope with challenging scenarios, a novel token-learning model based on Transformer for HPE is presented.
- We reveal two key insights in head poses, namely, the intra-orientation relationship and the cross-orientation relationship. Several learnable orientation tokens are designed to encode the general information of cross-orientation relationships. Moreover, a novel token guide multi-loss function is designed to train the model.
- Experiments are conducted on three benchmark HPE datasets. Results show TokenHPE achieves state-of-the-art performance with a novel token-learning concept compared with its existing CNN-based counterparts. Besides, we conduct abundant visualizations to illustrate the effectiveness of the proposed orientation tokens.

The remainder of this article is organized as follows. In Section II, we review the head pose estimation-related works. Section III introduces the proposed model for head pose angle inference and experimental results are provided in Section IV. In Section V, we conclude this study.

II. Related Works

A. Head Pose Estimation

Generally, the traditional HPE models can be classified into three kinds, such as Euler angle regression (EAR)-based models [19], [29], [31], [32], extra information-utilized (EIU) models [20], [21], [33], [34], [35], and alternative parametrizations of orientation (APO) model [23], [24], [25], [26]. For the EAR-based head pose estimation method, three Euler angles need to be regressed progressively. The paradigm in early studies was to consider HPE as a regression problem [3], [19], [29], [36]. Abate et al. [36] proposed a web-shaped model algorithm to encode the pose of the face and then applied regression algorithms to predict the pose of the face. Recently, CNNs have been adopted for HPE and remained dominant for many years because convolution can efficiently reveal the visual patterns on human faces. Ruiz et al. [29] was the first to propose an end-to-end method which independently predicts three Euler angles by using a multi-loss network based CNN. In [19], Yang et al. proposed FSA-Net, a novel architecture that consists of progressive stage fusions and fine-grained spatial structures. The spatial information can be preliminarily learned by setting a learnable or fixed importance over the spatial location. However, becaues of the incapacity of CNN to learn the relationships among visual patterns, further facial part relationships are not explored in this category.

For the EIU approaches, extra facial information is exploited to facilitate angle estimation. With graph neural network (GCN) being generalized to various natural language processing (NLP) and computer vision tasks [37], [38], [39], [40], [41], Xin et al. [33] proposed a GCN-based method which learns through the facial landmark graph. However, the precision of the model depends largely on the precision of the additional landmark detector. Kazemi et al. [12] proposed a general framework based on gradient boosting for learning an ensemble of regression trees that optimizes the sum of square error loss and naturally handles missing or partially labelled landmarks. Wu et al. [20] proposed a multi-task approach named SynergyNet that predicts complete 3D facial geometry. Through synergistic learning of 3D landmarks and 3D morphable models (3DMM) parameters, improved performance is achieved by the collaborative contribution. In these methods, facial part relationships can be learned from landmarks or other extra information. However, many manual annotations are required for training, which is laborious and inefficient.

For the APO models, Euler angles representation are usually substituted with other representations. Most contributions to HPE in recent years have focused on alternative parametrizations of head orientation because traditional Euler angle labels inevitably have some problems at specific orientations. Geng et al. [26] proposed a multivariate label distribution as a substitute of Euler angles. In this manner, inaccutate manual annnotation can be alleviated and the original label is softened, making the training easy. In [25], a vector-based head pose representation is proposed which handles the issue of discontinuity of Euler angle annnotation. Recently, Hepel et al. [24] proposed a rotation matrix-based representation for HPE. In this way, the ambiguity problem was perfectly resolved

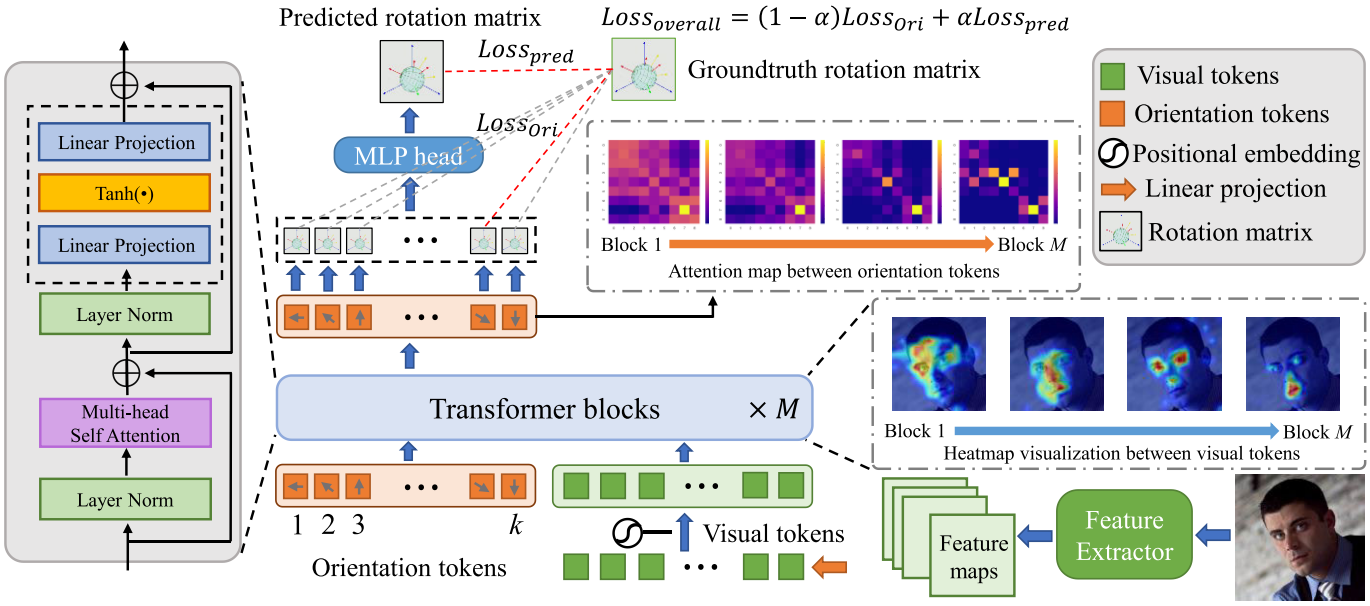


Fig. 3. Pipeline of our TokenHPE model.

by full pose regression based on rotation matrices. In our previous work [22], the head orientations are represented by matrix Fisher distribution based on rotation matrix, which greatly avoids the ambiguity problem of Euler angle labels. Furthermore, in [27], the characteristics of head pose image varitations in different directions are revealed and leveraged by constructing the anisotropic angle distributions. Although these methods have achieved impressive results, the intrinsic facial and orientational relationships are not fully exploited.

B. Vision Transformer and Its Applications

ViT is a variant model of Transformer [42], which is originally utilized in NLP. In ViT model, an input image is divided into patches and projected into 1D vectors called tokens. These visual patches can be viewed as words. In addition, a learnable class token is concatenated similarly to the original Transformer. The success of ViT quickly focused researchers' attention to applying the Transformer architecture in various vision tasks, including fine-grained classification [43], [44], [45], object detection [46], facial expression recognition [47], human pose estimation [48], and image segmentation [49]. Li et al. [48] proposed the utilization of learnable tokens to represent each human keypoint entity on the basis of prior knowledge. Through this sensible token construction, constraint cues and visual cues are expicitly learned and incorporated through the Transformer architecture. Dhingra [32] preliminarily utilized Transformer encoder after a CNN backbone for HPE. However, Transformer's ability to learn the semantic relationships was not fully exploited. In [50], Cordonnier et al. provided a theoretical explanation of the long-distance information learned in Transformer. Based on their theoretical foundation, we believe that the intra- and cross-orientation relationship revealed in this work can be learned based on Transformer architecture. Specifically, the intra orientation relationship can be learned by attention mechanism, and cross-orientation information can be encoded into learnable orientation tokens.

III. Proposed TokenHPE Model

In this section, the overview of the proposed TokenHPE is provided first. Second, the details of the four parts of the model are elaborated. Lastly, the supplementary architecture details of the TokenHPE are provided.

A. Architecture Overview

Our method's overview is shown in Fig. 3. The TokenHPE model comprises four parts. The first one is visual token construction, where the input image is transformed into visual tokens through multiple approaches. The second part is orientation token construction. We provide two strategies to construct orientation tokens based on our finding on head image panoramic overview. The third part is the Transformer module, wherein the relationships of facial parts and orientation characteristics in the basic regions are learned by the Transformer mechanism. The fourth part is token learning-based prediction. A novel token guide multi-loss that can help the orientation tokens encode general information is also introduced in this part.

B. Visual Token Construction

In this part, an original input RGB image is transformed into visual tokens. We provide three options to obtain the visual tokens: by patch division of the original image (Option 1), by extracting feature maps from a CNN (Option 2), and by selecting the tokens from a ViT backbone (Option 3) [30]. For Option 1, suppose we have an input image I of size $H \times W \times C$. The image is divided into patches with patch size $P_h \times P_w$. each patch is subsequently resized into a 1-dimensional vector of size $P_h \times P_w \times C$. Linear projection

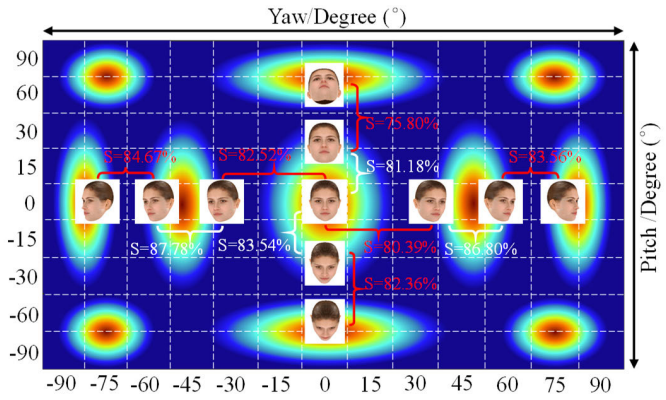


Fig. 4. Illustration of significant facial part change on neighbored orientations measured by cosine similarity scores, which are denoted as "S".

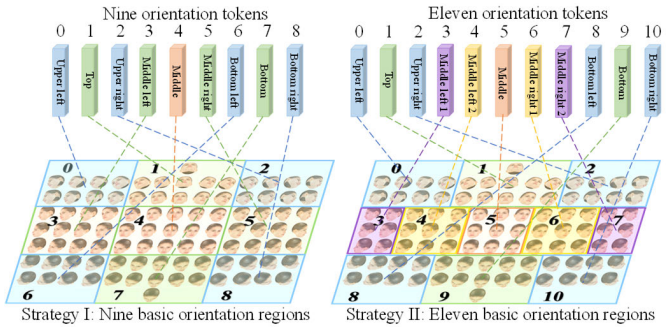


Fig. 5. Construction of orientation tokens. We discover that the head pose panoramic overview can be roughly divided into several basic orientation regions according the neighbor image similarities. As the division granularity varies, the number of basic orientation regions also varies.

is applied to obtain a visual token. This operation is expressed as:

$$O : p \rightarrow s \in \mathbb{R}^d, \tag{1}$$

where p refers to a 1D patch vector and s is a visual token with a dimension of d. For Option 2, the output of the CNN extractor is a set of feature maps with a size of $H \times W \times C'$. The remaining operations are similar to those in Option 1. For Option 3, the visual tokens can be simply selected from the output of a Transformer backbone.

Given that spatial relationships are essential for accurate HPE, positional embedding, pos, is added to the visual tokens to reserve spatial relationships, which can be expressed as:

$$[visual] = \{s_1 + pos, s_2 + pos, \cdots, s_n + pos\}, \tag{2}$$

where n is the number of patches. Then, we obtain n1D vectors symbolically presented by [visual] tokens.

C. Orientation Token Construction

1) Basic Orientation Region Partitioning: The cross-orientation relationship information is encoded into learnable orientation tokens. To construct the orientation tokens, the panoramic overview is divided into several basic orientation regions. Within a specific orientation region, the orientations have high similarities on head pose characteristics.

The significant facial part change angle threshold can be observed by calculating the cosine similarities between the feature maps generated from the feature extractor in different head pose images. As shown in Fig. 4, the cosine similarity is relatively lower when significant facial part change happens, such as the appearance/disappearance of eye on one side and appearance/disappearance of ear. For example, when the pitch angle is constantly at 0° and the yaw angle moves from −90° to 0°, two main significant facial part changes are marked by variation on cosine similarities from 84.67% to 87.78% and then to 83.54%. Therefore, three basic orientation regions on the left central part can be defined according to the discontinuity on neighborhood cosine similarities. Following this rule, the remaining basic orientation regions can be easily defined.

Based on quantitative results on cosine similarities of neighborhood head pose images, we introduce two partitioning strategies, as shown in Fig. 5. In Strategy I, the panoramic overview is divided into nine basic orientation regions according to the appearance of the eyes and the overlapping of the nose and mouth. 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 depicted in the left part of Fig. 5, head poses in the same region are similar, and the opposite head poses are symmetric. In Strategy II, the panoramic overview is divided into 11 regions, with a fine-grained partition in the yaw direction. 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 the right part of Fig. 5, 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.

2) Orientation Token: After the quantitative analysis on Basic orientation region partitioning, we construct the same number of d dimensional learnable vectors to represent k basic orientation regions. These vectors are symbolized as [dir] tokens. Then, the [visual] tokens are concatenated with the [dir] tokens as the input of Transformer blocks. After that, the processed [dir] tokens are chosen as the output of Transformer.

D. Transformer Blocks

Inputted with the [visual] and [dir] tokens, the Transformer blocks learns the relationships among facial parts and head orientations. The Transformer is constructed by stacking M identical unit blocks. 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. Self-attention (SA) is defined as:

$$SA\left(R^t\right) = softmax\left(\frac{R^t W_Q \left(R^t W_K\right)^T}{\sqrt{\theta}}\right)\left(R^t W_V\right), \quad (3)$$

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. R^t is the output of the t-th Transformer layer. θ is part of the scaling factor $1/\sqrt{\theta}$. In SA, s equals the dimension d of the tokens. MSA is an extension of SA with h self-attention operations, which are named heads. In MSA, θ is typically set as d/h. Thus, MSA can be formulated as:

$$MSA\left(R^t\right) = \left[SA_1\left(R^t\right); SA_2\left(R^t\right); \cdots; SA_h\left(R^t\right)\right] W_P, \quad (4)$$

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

$$\tilde{R}^{t-1} = MSA\left(LN\left(R^{t-1}\right)\right) + R^{t-1}, \quad (5)$$

$$R^t = MLP\left(LN\left(\tilde{R}^{t-1}\right)\right) + \tilde{R}^{t-1}. \quad (6)$$

The MLP module is constructed by two linear projections, with a Tanh(•) activation function and dropout operations in between.

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 $\{R_1^M, R_2^M, \cdots, R_k^M\}$, where k is the number of [dir] tokens.

E. Token Learning-Based Prediction

Suppose a set of orientation tokens outputted by part three in our model is denoted as $\{R_1^M, R_2^M, \cdots, R_k^M\}$, where k is the number of orientation tokens. The orientation tokens need to be transformed to rotation matrices for training and prediction. We adopt similar transformation strategy as used in [24]. The transformation is elaborated as follows.

First, a linear projection is applied to R_i^M 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:

$$\hat{A}_i = F_{GS}\left(W R_i^M\right), \quad (7)$$

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

$$F_{GS}\left(p_1, p_2\right) = \left[\, q_1 \quad q_2 \quad q_3 \,\right], \quad (8)$$

where $p_1, p_2 \in \mathbb{R}^3$ are 3D column vectors of a rotation matrix. q_i is 3D column vector of the rotation matrix defined as:

$$\begin{cases} q_1 = \dfrac{p_1}{\|p_1\|}, \\ u_2 = p_2 - (q_1 \cdot p_2)\, q_1, \\ q_2 = \dfrac{u_2}{\|u_2\|}, \\ q_3 = q_1 \times q_2. \end{cases} \quad (9)$$

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

To obtain the final prediction rotation matrix, $\{\hat{A}_1, A_2, \cdots, \hat{A}_k\}$ is concatenated and flattened as the input of the MLP head, which can be formulated as:

$$\hat{A} = F_{GS}\left(W_2\left(\tanh\left(W_1 \cdot \tilde{A} + b_1\right)\right) + b_2\right), \quad (10)$$

where $\tilde{A} \in \mathbb{R}^{9\cdot k}$ is a vector of flattened rotation matrices. W_i and b_i are the weight matrix and bias vector of the MLP module, respectively. In the training stage, the intermediate rotation matrices and the final prediction rotation matrix are used for calculating the loss for back propagation. In the prediction stage, only the prediction rotation matrix $\hat{A}$ is outputted as the model prediction.

F. Total Loss Function

The prediction of the proposed TokenHPE is a rotation matrix representation denoted as $\hat{A}$. Suppose that the groundtruth rotation matrix is A. The geodesic distance is used as the loss between two 3D rotations, similar to that used in [24]. The geodesic distance loss is formulated as:

$$L_g\left(A, \hat{A}\right) = \cos^{-1}\left(\frac{tr\left(A\hat{A}^T\right) - 1}{2}\right). \quad (11)$$

1) Orientation Token Loss: Information can be encoded into the orientation tokens through the orientation token loss. It is defined as a mean squared error, which is formulated as:

$$L_{Ori} = \sum_{i=1}^{k} I(A, i) \cdot L_g\left(A, \hat{A}_i\right), \quad (12)$$

where k is the number of basic orientation regions, A is the ground truth rotation matrix, $\hat{A}_i$ is the predicted rotation matrix, and $\mathbb{I}(A, i)$ is an identity function that determines if a ground truth head pose lies in the i-th basic region. $\mathbb{I}(A, i)$ can be written as:

$$\mathbb{I}(A, i) = \begin{cases} 1, & if\ A\ in\ region\ i, \\ 0, & if\ A\ not\ in\ region\ i. \end{cases} \quad (13)$$

2) Prediction Loss: The predictions from the orientation tokens are aggregated to form the final prediction of our model. This is optimized by the prediction loss, which is formulated as:

$$L_{pred} = L_g\left(A, \hat{A}\right), \quad (14)$$

where $\hat{A}$ is the prediction of the model.

3) Overall Loss: The overall loss consists of the orientation token loss and the prediction loss. It is formulated as:

$$L_{overall} = \gamma L_{pred} + (1-\gamma) L_{Ori}, \tag{15}$$

where γ is a hyperparameter that balances prediction loss and orientation token loss.

G. Network Parameters

To obtain the visual tokens, three options aforementioned previously can be utilized. A few extra CNN layers can efficiently extract low-level superficial features. In our different versions, a feature extractor is added, or the raw image patches are manipulated directly. In the version added with a feature extractor, many low-level features are utilized for prediction. In Option 2, we adopt the widely used stem-net, which can quickly downsample the feature map into 1/4 input resolution in a very shallow convolutional structure. In Option 3, we adopt ViT-B/16 as the feature extractor for a tradeoff between model size and performance. The outputs of ViT are the visual tokens that can be directly used in the second part of the proposed model. Option 3 is set by default in our TokenHPE model if not specially mentioned.

IV. Experimental Results and Discussion

A. Generally Setting

1) Evaluation Metrics: Two common evaluation metrics are selected to validate the performance of the comparing methods. It includes the mean absolute errors of Euler angles (MAE), and mean absolute errors of vectors (MAEV). For the MAE, it is usually assumed that pose angles are known. Namely, the Euler angles $\{yaw, pitch, roll\}$ of an image are considered as the ground-truth. The symbols yaw, $pitch$, and $roll$ represent pitch, yaw, and roll angle, respectively. The predicted set of Euler angles from a model is denoted as $\left\{\widehat{yaw}, \widehat{pitch}, \widehat{roll}\right\}$. Then, MAE is defined as:

$$MAE = \frac{1}{3}\left(|yaw - \widehat{yaw}| + \left|pitch - \widehat{pitch}\right| + \left|roll - \widehat{roll}\right|\right). \tag{16}$$

We adopt MAE as an evaluation metric. However, because this metric is unreliable, the MAEV results are given at the same time for a more accurate measurement of the models.

For the MAEV, it is usually based on rotation matrix representation. For an image, suppose that the groundtruth rotation angle matrix is $A = [a_1, a_2, a_3]$, where a_i is a 3D vector that indicates a spatial direction. The predicted rotation matrix from a model is denoted as $\hat{A} = [\hat{a}_1, \hat{a}_2, \hat{a}_3]$. MAEV can be formulated as:

$$MAEV = \frac{1}{3}\sum_{i=1}^{3}\left\|a_i - \hat{a}_i\right\|_1. \tag{17}$$

2) Datasets: Three datasets are employed in our experiments as listed below.

1) *BIWI dataset* [51]: It includes 15, 678 images of 20 individuals (14 individuals are males and the rest are females) with 4 of whom recorded twice., a RGB-depth image (640 × 480 pixels), and the corresponding head pose annotation are recorded for each video frame. The head pose range covers about ±60° pitch and ±75° yaw. The 3D location of the head and its Euler angle are provided as the ground truth labels of each frame.
2) *AFLW2000 dataset* [52]: It contains 2000 images that have been annotated with 68-point 3D facial landmarks at image-level. The dataset is typically adopted as the evaluation benchmark of 3D facial landmark detection task. The head poses in this dataset are diverse and always difficult to be detected by traditional CNN-based face detectors. Notice that the 2D landmark annotations are discarded in the dataset because some of the data do not have complete landmark points, as mentioned in the original paper.
3) *300W-LP dataset* [52]: It is an expanded version of 300W dataset, which collects multiple alignment databases with 68 landmarks, including IBUG, XM2VTS, LFPW, AFW, and HELEN. With 300W dataset, 300W-LP adopts the proposed face profiling to generate about 61k samples across large poses. The dataset is usually employed as the training set for HPE.

B. Training Details

1) Training: In our experiments, the TokenHPE is trained end-to-end. The batch size is set as 64, and γ is set to 0.65 by default. We train the proposed TokenHPE model for 120 epochs. The learning rate is initialized as 0.0001, which is further decayed by a factor of 10 at the 30^{th} and 60^{th} epochs.

2) Initialization: All the images are resized into 240 × 240 pixels. A random crop is then applied to make the input image size 224×224 pixels. Our method is implemented with the Pytorch toolbox with a single TITAN V GPU. All the parameters in our model are trained with random initialization.

3) Computational Time: The training time is about six hours on GPU. In the inference stage, our model can inference in real time on GPU at over 400 fps. When ran on CPU, the inference speed is about 10 fps.

C. Compare to State-of-the-Art

We compare our proposed TokenHPE with 13 state-of-the-art (SOTA) methods, including Euler angle regression methods (HopeNet, FSA-Net, FAN)), extra information-utilized methods (3DDFA, Dlib, EVA-GCN, img2pose, SynergyNet), and alternative parametrizations of orientation methods (QuatNet, TriNet, 6DRepNet). In our two experiments, we follow the convention by FSA-Net [19]. We conduct experiments on two versions of our model: TokenHPE-N with nine basic orientation regions and TokenHPE-E with eleven basic orientation regions. TokenHPE-E is our standard model referred as TokenHPE.

Firstly, we follow the conventional protocol 1 [19], in which the models are all trained on the 300W-LP dataset and tested on AFLW2000 and BIWI datasets. Tables I and II show the results of the first experiment. An extra column is added to indicate which methods are free from extra annotation for

TABLE I
COMPARISON WITH SOTA METHODS ON THE AFLW2000 DATASET WITH PROTOCOL 1

HPE Models	w/o Annotation information	Vector errors				Prediction errors (°)			
		Front	Down	Left	MAEV	Roll	Yaw	Pitch	MAE
3DDFA [52]	No	18.52	39.05	30.57	29.38	28.43	4.71	27.05	20.08
Dlib [12]	No	14.31	28.51	26.56	23.13	22.83	8.50	11.25	14.19
FAN [13]	No	-	-	-	-	8.71	6.36	12.3	9.12
EVA-GCN [33]	No	-	-	-	-	4.11	4.46	5.34	4.64
SynergyNet [20]	No	-	-	-	-	2.55	3.42	4.09	3.35
img2pose [21]	No	-	-	-	-	3.28	3.43	5.03	3.91
HopeNet [29]	Yes	7.50	5.98	7.07	6.85	6.13	5.31	7.12	6.20
Xia *et al.* [28]	Yes	-	-	-	-	6.50	3.99	7.32	5.94
FSA-Net [19]	Yes	7.35	6.22	6.75	6.77	4.78	4.96	6.34	5.36
LwPosr [31]	Yes	-	-	-	-	4.88	4.80	6.38	5.35
HeadPosr EH64 [32]	Yes	-	-	-	-	4.30	4.64	5.84	4.92
QuatNet [23]	Yes	-	-	-	-	3.92	3.97	5.62	**4.50**
TriNet [25]	Yes	**6.52**	5.67	**5.78**	5.99	4.04	4.20	5.77	4.67
TokenHPE-N (ours)	Yes	6.97	5.21	6.16	6.11	4.29	4.53	5.73	4.85
TokenHPE-E (ours)	Yes	6.82	**5.10**	6.01	**5.98**	4.08	4.36	**5.54**	4.66

TABLE II
COMPARISON SOTA METHODS ON THE BIWI DATASET WITH PROTOCOL 1

HPE models	w/o Annotation information	Vector errors				Prediction errors (°)			
		Front	Down	Left	MAEV	Roll	Yaw	Pitch	MAE
HopeNet [29]	Yes	8.68	6.73	7.65	7.69	3.72	6.01	5.89	5.20
FSA-Net [19]	Yes	7.22	5.96	6.03	6.40	3.07	4.56	5.21	4.28
QuatNet [23]	Yes	-	-	-	-	2.94	4.01	5.49	4.15
TriNet [25]	Yes	6.57	5.46	5.57	5.86	4.11	**3.05**	4.76	3.97
EVA-GCN [33]	No	-	-	-	-	2.98	4.01	4.78	3.92
HeadPosr EH64 [32]	Yes	-	-	-	-	**2.69**	3.37	5.44	3.83
WHENet [53]	Yes	-	-	-	-	3.06	3.99	**4.39**	3.81
TokenHPE-E (ours)	Yes	**6.23**	**5.17**	**5.41**	**5.60**	**2.71**	3.95	4.51	**3.72**

fair comparison. Results show that the proposed TokenHPE is on par with SOTA methods on AFLW2000 dataset and achieves SOTA results on MAEV on BIWI dataset. Among the compared methods, HopeNet [29] is normally considered the baseline of HPE. Compared with it in Table I, our model achieves a 24.8% decrease in MAE and a 12.7% decrease in MAEV, which shows the high accuracy of our method. Xia et al. [28] proposed a method that applies an affine transformation to simplify the input and combines landmarks information into a CNN feature extractor, leading to 4.20% improvement from baseline. TriNet [25] is a vector-based model, in which the head pose is represented by vectors instead of Euler angles to solve the discontinuity problem. The MAE is 0.69 lower than the baseline. A new MAEV metric is also introduced. We adopt this metric for our comparison. Compared with TriNet, our method obtains a lower MAEV value, which indicates that the proposed relationship-learning approach has the potential to achieve SOTA performance. Some extra information-utilized methods (e.g., SynergyNet, img2pose) are also compared in Table I. FAN, Dlib, SynergyNet and img2pose, which are better known to perform landmarks prediction, are not specially designed for HPE but can be readily modified for HPE as a downstream task. EVA-GCN [33] is a facial landmark graph-based method, which takes the detected landmark graph as the input The GCN can learn the landmark relationships for HPE thus the model result has an impressive improvement. SynergyNet is a multi-task model, and HPE is a subtask. The model is trained by synergistic learning. Therefore, abundant information, including 3DMM parameters and 3D landmarks, is utilized to enhance the performance. Img2pose [21] achieves the best result among annotation utilized methods by applying the six degress of freedom 3D face pose estimation. HeadPosr achieves excellent performance by introducing a Transformer encoder to a CNN backbone, which is built on the CNN learning methodology. In general, compared to other methods that mainly based on CNN and its variants, our model is the only Transformer-based token learning method, thus has a stronger ability to learn the facial relationships and the orientation characteristics in the basic regions. Therefore, even on the challenging AFLW2000 dataset that has many difficult-to-predict images, our method still outperforms the majority of the compared methods by a large margin. The excellent performance verifies the orientation learning capacity of the proposed TokenHPE.

TABLE III
COMPARISON WITH SOTA METHODS ON BIWI DATASET WITH PROTOCOL 2

HPE models	Prediction errors (°)			
	Roll	Yaw	Pitch	MAE
FSA-Net [19]	3.60	2.89	4.29	3.60
Xia *et al.* [28]	3.09	2.39	4.92	3.47
FDN [54]	2.88	3.00	3.98	3.29
Hopenet [29]	3.00	3.29	3.39	3.23
TriNet [25]	2.44	2.93	3.04	2.80
6DRepNet [24]	2.36	2.69	**2.92**	2.66
TokenHPE-E (ours)	**2.01**	**2.28**	3.01	**2.49**

Afterward, the authors follow the protocol 2 in [19], in which the model performace is evaluated on the BIWI dataset alone with a ramdom 7:3 separation for training and testing. Experiments are conducted on the TokenHPE and the compared SOTA methods with protocol 2. Results shown in Table III demonstrates the proposed TokenHPE outperforms other methods by a large margin both on MAE and three Euler angles. 6DRepNet [24] uses the rotation matrix representation with a CNN backbone. Compared to 6DRepNet, our TokenHPE can learn the general regional information and facial relationships through Transformer architecture, resulting in a 6.39% drop on MAE. The similar results on two experiments show that our method is robust and stable, and its impressive performance is independent from the training and testing datasets.

D. Ablation Study

In this section, we conduct ablation study on different segment of our TokenHPE model. The models are trained on 300W-LP dataset and tested on AFLW2000 dataset by default if there is no explicit declaration.

1) On the Token Guide Multi-Loss Function: The proposed model is trained by a token guide multi-loss function. We conduct ablation study on the orientation token loss and prediction loss, which is controlled by γ. When γ is set to 1.0, the model is trained solely on the prediction loss, meaning the model learns the basic orientation regions by itself. As the value of γ decreases, orientation token loss plays an increasingly important role in helping the model learn the orientation information. Since the final prediction head an indispensable part in our method, the γ only can be set to a very small value near zero but cannot be zero. Preliminarily, we remove each component of the multi-loss individually and evaluate the importance. In setting I, since the prediction loss is indispensable, we set it to a very small weight (5%) thus the orientation loss is predominant.

Results presented in Table IV shows that both sub-losses are significant for model performance. Furthermore, we set a sequence of γ values for a thorough investigation of the contribution of two sub-loss to the model performance. The experimental results are shown in Fig. 6. When γ decreases, MAE initially decreases then increases. The best result is obtained when γ is set to 0.6. This situation indicates that the token guide loss indeed helps the model encode the basic orientation regions. As γ decreases, the flexibility of the model is constrained, resulting in poor performance. When γ is set to 0.6, where prediction loss and orientation token loss jointly contribute to the overall loss, the model reaches the best performance.

TABLE IV
ABLATION STUDY ON THE TOKEN GUIDE MULTI-LOSS FUNCTION

Setting	L_{pred}	L_{ori}	#Regions=9		#Regions=11	
			MAE	Δ	MAE	Δ
Appr. only L_{ori}		✓	5.33	-9.90%	5.26	-12.88%
Only L_{pred}	✓		4.99	-2.89%	4.87	-4.51%
Joint	✓	✓	4.85	-	4.66	-

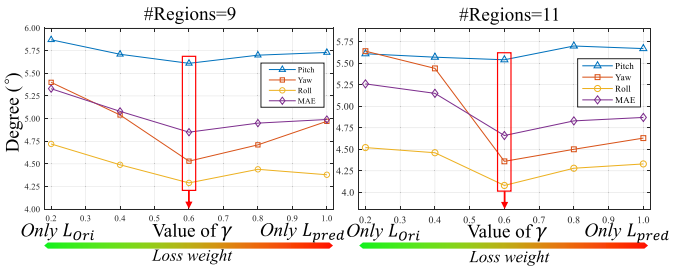


Fig. 6. Effect of the token guide multi-loss function. The hyperparameter γ balances the importance of prediction loss and the orientation token loss. When γ is set to 1.0, the model is solely trained with prediction loss, while when γ approaches zero, the model is solely trained with orientation token loss.

TABLE V
EFFECT OF ORIENTATION TOKENS

#[dir] Tokens	Guidance of orientation information	Prediction errors (°)				MAEV
		Roll	Yaw	Pitch	MAE	
None	No	4.44	5.01	5.79	5.08	6.48
9	No	4.38	4.97	5.61	4.99	6.36
9	Yes	4.29	4.53	5.73	4.85	6.11
11	No	4.33	4.63	5.67	4.87	6.14
11	Yes	**4.08**	**4.36**	**5.54**	**4.66**	**5.98**

2) On the Orientation Tokens: Since the number of orientation tokens is derived from the partitioning strategies which are rigorously defined, we evaluate their effectiveness on three settings. In the first setting, we removed all orientation tokens and leave a learnable token similar to the [cls] token in ViT for prediction. In the second and third settings, we use nine tokens (Strategy I) and eleven tokens (Strategy II), respectively. When there is no guidance of orientation information, the model is trained only with the prediction loss thus the reginal orientation characters are not learned in this scheme. Results in Table V show that nine orientation tokens and eleven tokens bring 5.7% and 7.7% improvement on the MAEV, 4.5% and 8.2% improvement on MAE compared to the version that has no orientation token. Besides, the guidance of orientation information brings 2.8% and 4.3% improvement on MAE in nine and eleven token settings compared to the control groups that don not have the orientation information from the orientation loss. Overall, the quantitative results verify the effectiveness of the proposed token guide multi-loss and the contribution of orientation tokens.

TABLE VI
EFFECT OF THE FEATURE EXTRACTOR

Feature extractor	Prediction errors (°)				MAEV
	Roll	Yaw	Pitch	MAE	
None	5.11	4.96	6.07	5.38	6.65
CNN	4.48	5.71	**4.68**	4.96	6.04
ViT	**4.08**	**4.36**	5.54	**4.66**	**5.98**

TABLE VII
RESULTS OF DIFFERENT POSITIONAL EMBEDDING STRATEGIES

Positional embedding	Prediction errors (°)			
	Roll	Yaw	Pitch	MAE
None	4.39	4.51	7.11	5.33
Learnable	4.33	4.63	5.67	4.87
2D sine	**4.08**	**4.36**	**5.54**	**4.66**

TABLE VIII
ABLATION STUDY ON TRANSFORMER BLOCK HYPERPARAMETERS

	Prediction errors (°)				MAEV
	Roll	Yaw	Pitch	MAE	
Activation function in P module					
Tanh	**4.08**	4.36	**5.54**	**4.66**	5.99
ReLU	4.28	4.43	5.71	4.80	6.01
GELU	4.19	**4.35**	5.63	4.72	**5.98**
Token dimension					
386	4.22	4.45	5.64	4.77	6.00
128	**4.08**	**4.36**	**5.54**	**4.66**	**5.98**
64	4.29	4.54	5.70	4.85	6.04
Number of heads in MSA					
16	4.25	4.42	5.62	4.76	5.99
12	**4.08**	**4.36**	**5.54**	**4.66**	**5.98**
8	4.32	4.46	5.77	4.85	6.05

3) Feature Extractor: Since the visual tokens are generated from the feature extractor, the performance of the model partially depends on it. Therefore, we conduct experiments on CNN and ViT feature extractors to reveal the extent to which performance is affected by the feature extractor. As shown in Table VI, we test three versions with or without feature extractors. Results show that feature extractor improves performance to a specific extent compared with the version without a feature extractor. The model with ViT feature extractor has the best performance. This mainly contributes to the better capability of Transformer to encode semantic visual information than CNNs.

E. Discussion

1) Positional Embedding: Different from classification tasks, spatial relationships play an important role in HPE. Given that the self-attention operation is positionally invariant, normally, 2D sine positional embedding is added to reserve the spatial relationships for computer vision tasks. Therefore, we conduct experiments on our TokenHPE model with three positional embedding types (i.e., no positional embedding, learnable positional embedding, and 2D sine positional embedding) to investigate the effect of positional embedding. As shown in Table VII, the model with 2D sine positional embedding demonstrates the best performance. The learnable positional embedding version has a low prediction accuracy.

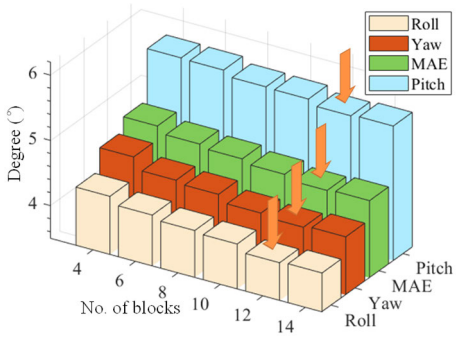


Fig. 7. Effect of the number of Transformer blocks.

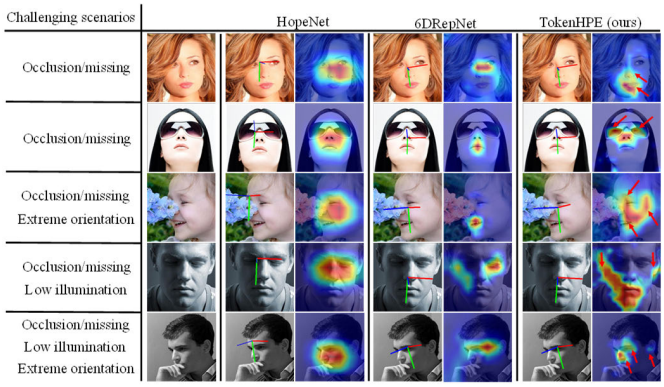


Fig. 8. Heatmap visualization of three models, namely, HopeNet (left), 6DRepNet (middle), and our proposed model (right) in challenging scenarios, including occlusion and extreme orientation, occlusion, low illumination. The red-color areas mean that the model provides high attention to these facial parts.

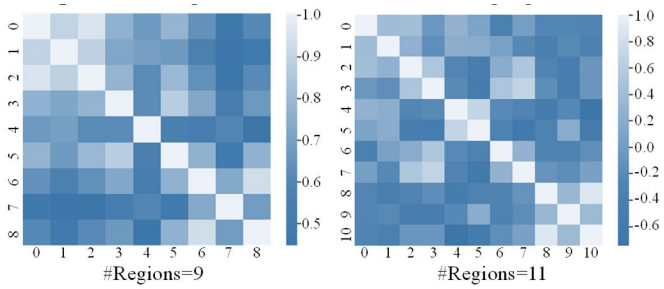


Fig. 9. Cosine similarity matrix between the learned orientation tokens. (a) Strategy I: nine basic orientation regions. (b) Strategy II: eleven basic orientation regions.

The model without positional embedding performs the worst. Therefore, fixed positional embedding is important for a model to learn the facial part relationships. Meanwhile, the absence of positional embedding results in the loss of spatial geometric relationships between visual tokens.

2) Transformer Block Parameters: Transformer parameters have effect in a certain extent on model performance. Therefore, we investigate different options of Transformer block parameters. For comparison, only the investigated parameter varies while the others are set to default configuration (token dimension of 64, GELU activation function, and 8 heads in MSA). The experimental results are shown in Table VIII. The best dimension of token is 128, the best activation function is Tanh, and best number of heads is 12.

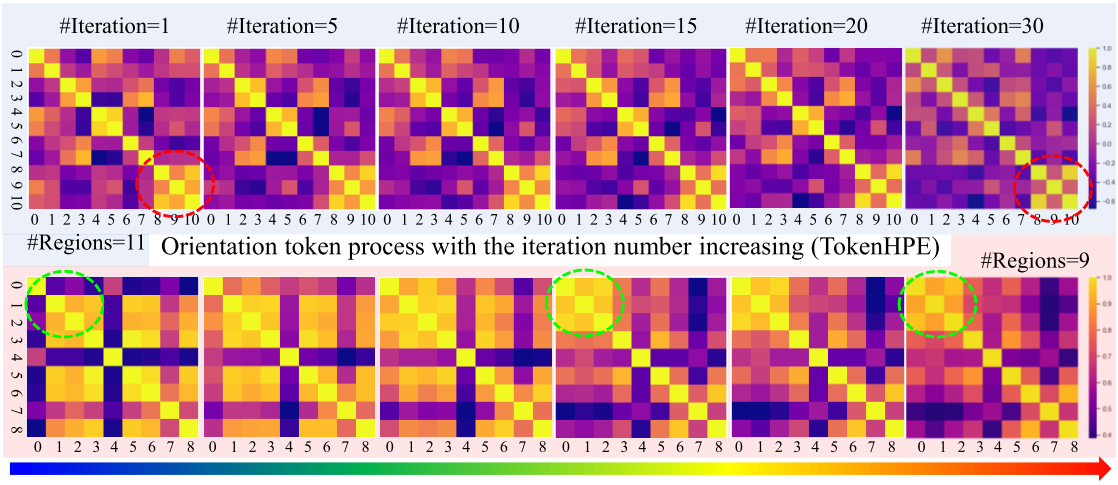


Fig. 10. Cosine similarity matrix between orientation tokens during model training. The orientation information is learned gradually by the orientation tokens.

3) Number of Transformer Blocks: Different numbers of Transformer blocks are evaluated to check their effect on HPE. The results are shown in Fig. 7. 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.

F. Visualizations

We visualize the inference details to investigate how the TokenHPE explicitly utilizes orientation tokens to find the facial part relationships and orientation characteristics in the basic regions. Notice that on most common images, our proposed TokenHPE exhibits similar behaviors and all images in Figs. 8-12 are randomly chosen from the AFLW 2000 dataset in order to visualize the details.

1) Visualization in Challenging Scenarios: To confirm that our model can learn critical minority facial part relationships and tackle challenging scenarios, we use Grad-CAM [55] to visualize the attention of head pose predictions in a challenging subset of AFLW2000. Two representative methods (HopeNet and 6DRepNet) are adopted for a comparison with our proposed model. As Fig. 8 shows, our method can learn the crucial minority relationships of facial parts, such as the eyes, nose, and ears in challenging scenarios (e.g., occlusion, extreme orientation, low illumination) where some facial parts are missing and hard to estimate. In these scenarios, the compared methods performed poorly when abundant facial information is missed. Row 2 indicates that our method can deduce the spatial location of the eyes to achieve accurate prediction compared with the other methods that only attend to the facial parts that appear. As shown on Row 4, our method presents an impressive capability to reveal the symmetric relationships of the face even though the entire left side of the face is dark due to low illumination. On Row 5, the attention heatmaps show that our method can find the critical minority relationships (nose, eyes, and ears) in the most challenging scenario. In summary, the heatmap visualization proves that our method can learn facial part relationships and can deduce the spatial relationships of facial parts to mitigate the obstacles in challenging scenarios.

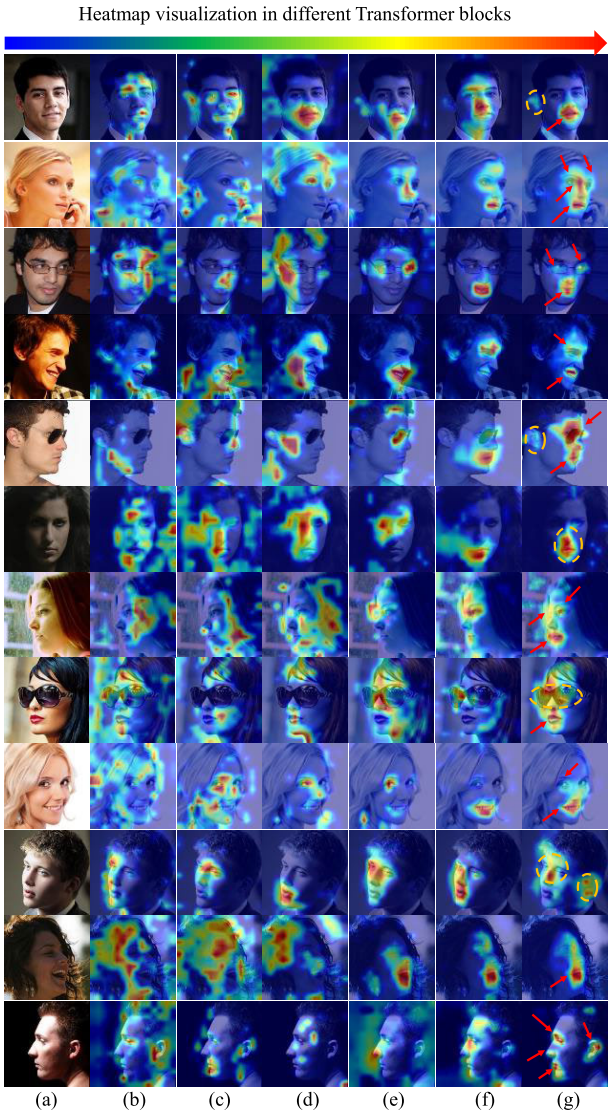


Fig. 11. Heatmap visualization in different Transformer blocks of the TokenHPE model. Arrows and circles indicate the crucial facial parts to which the model pays attention for the head pose prediction.

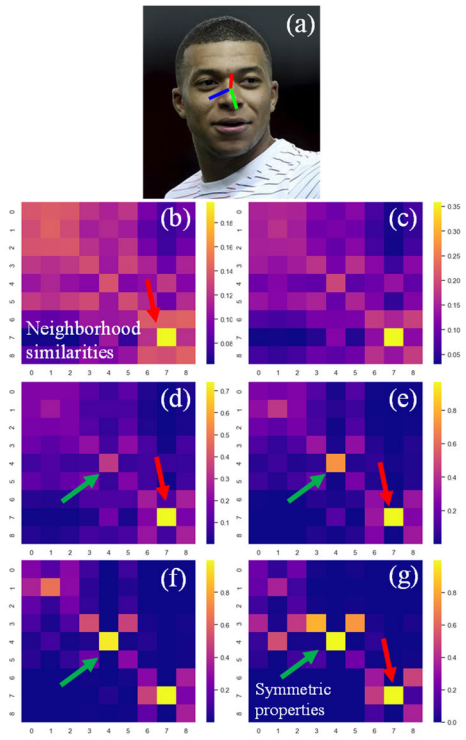


Fig. 12. Attention maps of orientation tokens of our TokenHPE model with the Transformer blocks number increasing.

2) Similarity Matrix of Orientation Tokens: We visualize the cosine similarities of the orientation tokens. As shown in Fig. 9, the neighbored orientation tokens are highly similar. The orientation tokens that represent symmetric facial regions have higher similarity scores than the tokens that represent the other unrelated regions. Therefore, the results of the similarity matrix verify that the general information and the cross-orientation relationships are learned by the orientation tokens.

3) Orientation Token Learning During Training: We calculate the cosine similarity between the orientation tokens in different training epochs. As Fig. 10 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 30^{th} 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 and the cross-orientation relationships can be learned by the orientation tokens.

4) Region Information Learned by Orientation Tokens: The attention maps of orientation tokens are visualized in Fig. 12. It can be observed that in shallow blocks, each orientation token pays similar attention to the rest in order to construct the global perception of the image. By contrast, in deeper blocks, each orientation token pays most attention on its neighborhood region tokens and spatial symmetric tokens to yield the final prediction. As indicated in Fig. 12, at the deeper Transformer blocks, the attention score is higher between neighbor regions (the diagonal) and symmetric regions, such as regions 0 and 2, regions 3 and 5, and regions 6 and 8. In Fig. 12, the attention score is higher in regions 3, 4, 6, and 7, indicating that the predicted head pose has more probability in the left–bottom direction, similar to the groundtruth. Therefore, from the visualization shown in Fig. 12, we can conclude that our model has the ability to encode the cross-orientation relationships of the basic regional orientation characteristics, including neighborhood similarities and symmetric properties.

V. Conclusion

In this work, we proposed an orientation cues-aware facial relationship representation learning method for head pose estimation. We revealed intra-orientation relationships and cross-orientation relationships on head images. To leverage these significant properties of head images, Transformer architecture was utilized to learn intra-orientation relationships, and several orientation tokens were designed to encode cross-orientation relationships according to panoramic overview partitions. The experimental results showed that TokenHPE achieves state-of-the-art performance and is capable to resolve the challenges of low illumination, occlusion, and extreme orientations. Besides, the success of TokenHPE reveals the significance of facial and orientational relationships for head pose estimation, which have been ignored in previous researches. Moreover, we hope this initial work can inspire further research on token-learning methods for HPE and other head related fields, such as attention detection, facial expression recognition, and gaze estimation.

References

[1] A. F. Abate, C. Bisogni, A. Castiglione, and M. Nappi, "Head pose estimation: An extensive survey on recent techniques and applications," *Pattern Recognit.*, vol. 127, Jul. 2022, Art. no. 108591.

[2] P. Barra, S. Barra, C. Bisogni, M. De Marsico, and M. Nappi, "Web-shaped model for head pose estimation: An approach for best exemplar selection," *IEEE Trans. Image Process.*, vol. 29, pp. 5457–5468, 2020.

[3] V. Drouard, R. Horaud, A. Deleforge, S. Ba, and G. Evangelidis, "Robust head-pose estimation based on partially-latent mixture of linear regressions," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1428–1440, Mar. 2017.

[4] J. Liu, Z. Wang, H. Qin, K. Xu, B. Ji, and H. Liu, "Free-head pose estimation under low-resolution scenarios," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Jul. 2020, pp. 2277–2283.

[5] Y. Wang, W. Liang, J. Shen, Y. Jia, and L.-F. Yu, "A deep coarse-to-fine network for head pose estimation from synthetic data," *Pattern Recognit.*, vol. 94, pp. 196–206, Oct. 2019.

[6] C. Bisogni, M. Nappi, C. Pero, and S. Ricciardi, "FASHE: A fractal based strategy for head pose estimation," *IEEE Trans. Image Process.*, vol. 30, pp. 3192–3203, 2021.

[7] W.-Y. Hsu and C.-J. Chung, "A novel eye center localization method for head poses with large rotations," *IEEE Trans. Image Process.*, vol. 30, pp. 1369–1381, 2021.

[8] E. Murphy-Chutorian, A. Doshi, and M. M. Trivedi, "Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation," in *Proc. IEEE Intell. Transp. Syst. Conf.*, Sep. 2007, pp. 709–714.

[9] A. Kumar, A. Alavi, and R. Chellappa, "KEPLER: Keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 258–265.

[10] D. Bicho, P. Girão, J. Paulo, L. Garrote, U. J. Nunes, and P. Peixoto, "Markerless multi-view-based multi-user head tracking system for virtual reality applications," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 2645–2652.

[11] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg, "Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 383–398.

[12] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2014, pp. 1867–1874.

[13] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.

[14] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2094–2107, Nov. 2015.

[15] M. Martin, F. Van De Camp, and R. Stiefelhagen, "Real time head model creation and head pose estimation on consumer depth cameras," in *Proc. 2nd Int. Conf. 3D Vis.*, vol. 1, Dec. 2014, pp. 641–648.

[16] G. Fanelli, T. Weise, J. Gall, and L. V. Gool, "Real time head pose estimation from consumer depth cameras," in *Proc. Joint Pattern Recognit. Symp.*, 2011, pp. 101–110.

[17] G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, and J. Kautz, "Robust model-based 3D head pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2015, pp. 3649–3657.

[18] J. Gu, X. Yang, S. De Mello, and J. Kautz, "Dynamic facial analysis: From Bayesian filtering to recurrent neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 2017, pp. 1548–1557.

[19] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "FSA-net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1087–1096.

[20] C.-Y. Wu, Q. Xu, and U. Neumann, "Synergy between 3DMM and 3D landmarks for accurate 3D facial geometry," in *Proc. Int. Conf. 3D Vis. (3DV)*, Dec. 2021, pp. 453–463.

[21] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, "img2pose: Face alignment and detection via 6DoF, face pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7617–7627.

[22] H. Liu, S. Fang, Z. Zhang, D. Li, K. Lin, and J. Wang, "MFDNet: Collaborative poses perception and matrix Fisher distribution for head pose estimation," *IEEE Trans. Multimedia*, vol. 24, pp. 2449–2460, 2022.

[23] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, "QuatNet: Quaternion-based head pose estimation with multiregression loss," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1035–1046, Apr. 2019.

[24] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "6D rotation representation for unconstrained head pose estimation," 2022, *arXiv:2022.12555*.

[25] Z. Cao, Z. Chu, D. Liu, and Y. Chen, "A vector-based representation to enhance head pose estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1187–1196.

[26] X. Geng and Y. Xia, "Head pose estimation based on multivariate label distribution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1837–1842.

[27] C. Zhang, H. Liu, Y. Deng, B. Xie, and Y. Li, "TokenHPE: Learning orientation tokens for efficient head pose estimation via transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 1–9.

[28] J. Xia, L. Cao, G. Zhang, and J. Liao, "Head pose estimation in the wild assisted by facial landmarks based on convolutional neural networks," *IEEE Access*, vol. 7, pp. 48470–48483, 2019.

[29] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2018, pp. 2074–2083.

[30] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[31] N. Dhingra, "LwPosr: Lightweight efficient fine grained head pose estimation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1495–1505.

[32] N. Dhingra, "HeadPosr: End-to-end trainable head pose estimation using transformer encoders," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 1–8.

[33] M. Xin, S. Mo, and Y. Lin, "EVA-GCN: Head pose estimation based on graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1462–1471.

[34] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2020, pp. 5203–5212.

[35] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[36] A. F. Abate, P. Barra, C. Pero, and M. Tucci, "Head pose estimation by regression algorithm," *Pattern Recognit. Lett.*, vol. 140, pp. 179–185, Dec. 2020.

[37] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[38] Y. Deng, H. Chen, H. Liu, and Y. Li, "A voxel graph CNN for object classification with event cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1172–1181.

[39] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.

[40] L. Ge et al., "3D hand shape and pose estimation from a single RGB image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2019, pp. 10833–10842.

[41] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2019, pp. 1237–1246.

[42] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[43] J. He et al., "TransFG: A transformer architecture for fine-grained recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 852–860.

[44] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[45] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[46] H. Zhang et al., "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.

[47] F. Xue, Q. Wang, and G. Guo, "TransFER: Learning relation-aware facial expression representations with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 3601–3610.

[48] Y. Li et al., "TokenPose: Learning keypoint tokens for human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11313–11322.

[49] Y. Wang et al., "End-to-end video instance segmentation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8741–8750.

[50] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," 2019, *arXiv:1911.03584*.

[51] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, Feb. 2013.

[52] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 146–155.

[53] Y. Zhou and J. Gregson, "WHENet: Real-time fine-grained estimation for wide range head pose," 2020, *arXiv:2005.10353*.

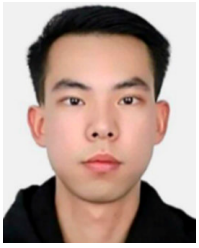
[54] H. Zhang, M. Wang, Y. Liu, and Y. Yuan, "FDN: Feature decoupling network for head pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12789–12796.

[55] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

Hai Liu (Senior Member, IEEE) received the M.S. degree in applied mathematics and the Ph.D. degree in pattern recognition and artificial intelligence from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2010 and 2014, respectively.

Since June 2017, he has been an Assistant Professor with the Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan. From 2023 to 2024, he was selected as the "China-European Commission Talent Program" under the National Natural Science Foundation of China (NSFC). He was a Senior Researcher with the UCL Interaction Centre, University College London, London, U.K., where he was hosted by Prof. Sriram Subramanian. He has authored more than 100 peer-reviewed articles in international journals from multiple domains. More than 20 articles were selected as the ESI highly cited articles and eight papers were selected as the hot papers. His current research interests include human pose estimation, gaze tracking, head pose estimation, facial expression recognition, deep learning, artificial intelligence, and self-regulated learning.

Dr. Liu won the First Prize for the Science and Technology Progress Award by the Hubei Province of China in 2020. He has been serving as a reviewer for more than six international journals, including IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSLATIONS ON INDUSTRIAL INFORMATICS, and IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. Since 2016, he has been a Communication Evaluation Expert of the NSFC.



Cheng Zhang (Student Member, IEEE) is currently pursuing the B.E. degree in artificial intelligence with Central China Normal University, Wuhan, China. His research interests include deep learning, object detection, and pattern recognition.

Yongjian Deng (Member, IEEE) received the Ph.D. degree from the City University of Hong Kong in 2021. He is currently an Assistant Professor with the College of Computer Science, Beijing University of Technology. His research interests include pattern recognition and machine learning with event cameras.

Tingting Liu (Member, IEEE) received the M.S. degree in natural language processing from the Huazhong University of Science and Technology in 2014 and the Ph.D. degree in education information technology from Central China Normal University (CCNU), Wuhan, China, in 2019.

She joined Hubei University, Wuhan, in 2020, where she is currently an Assistant Professor with the School of Education. From November 2022 to November 2024, she was selected as a Visiting Scholar with the School of Computer Science, City University of Hong Kong, Hong Kong, China, where she was hosted by Prof. You-Fu Li. Her current research interests include learning behavior analysis, human pose estimation, label distribution learning, and graph neural networks.

Dr. Liu has been serving as a reviewer for several international journals, including IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON MULTIMEDIA. She has been a Communication Evaluation Expert at the National Natural Science Foundation of China since 2020.

Zhaoli Zhang (Senior Member, IEEE) received the M.S. degree in computer science from Central China Normal University, Wuhan, China, in 2004, and the Ph.D. degree in computer science from the Huazhong University of Science and Technology in 2008.

He is currently a Professor with the National Engineering Research Center for E-Learning, Central China Normal University. His research interests include self-regulated learning, human–computer interaction, deep learning, image processing, knowledge services, and software engineering. He is a member of the China Computer Federation (CCF).

You-Fu Li (Fellow, IEEE) received the Ph.D. degree in robotics from the Department of Engineering Science, University of Oxford, U.K., in 1993.

He was a Research Staff with the Computer Science Department, University of Wales, Aberystwyth, U.K., from 1993 to 1995. He is currently a Professor with the Department of Mechanical Engineering, City University of Hong Kong. His research interests include robot sensing, robot vision, and attitude estimation. In these areas, he has published over 150 articles in refereed international journals.

Dr. Li has served as an Associate Editor for IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING. He is an Associate Editor of *IEEE Robotics and Automation Magazine*. He is an Editor of the IEEE Robotics and Automation Society Conference Editorial Board and the IEEE Conference on Robotics and Automation.