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

Abstract

Head pose estimation (HPE) has been widely used in the fields of human machine interaction, self-driving, and attention estimation. However, existing methods cannot deal with extreme head pose randomness and serious occlusions. To address these challenges, we identify three cues from head images, namely, neighborhood similarities, significant facial changes, and critical minority relationships. To leverage the observed findings, we propose a novel critical minority relationship-aware method based on the Transformer architecture in which the facial part relationships can be learned. Specifically, we design several orientation tokens to explicitly encode the basic orientation regions. Meanwhile, a novel token guide multi-loss function is designed to guide the orientation tokens as they learn the desired regional similarities and relationships. We evaluate the proposed method on three challenging benchmark HPE datasets. Experiments show that our method achieves better performance compared with state-of-the-art methods. Our code is publicly available at https://github.com/zc2023/TokenHPE.

1. Introduction

Head pose estimation (HPE) is a popular research area in computer vision and has been widely applied to driver assistance [29], human–computer interaction [36], virtual reality [22], and attention detection [5]. In recent years, HPE has been actively studied and the accuracy has been considerably improved in terms of utilizing extra facial landmark information [2,20], extra RGB-depth information [13,26–28], extra temporal information [16], stage-wise regression strategy [42], multi-task learning [1,38], and alternative parameterization of orientation [3,15,18,19,24]. Currently, many methods focus on the representation of the head pose orientation and have achieved impressive performance, but the intrinsic facial part relationships are usually neglected. A possible reason is that these relationships are difficult to learn by existing CNN architectures. However, in some challenging scenarios, as shown in the left part of Fig. 1, many remarkable facial parts are missing. Consequently, the remaining facial parts and their geometric relationships must be leveraged to achieve robust and high-accuracy prediction. Therefore, how to leverage the facial part relationships for high-accuracy HPE is an attractive research topic.

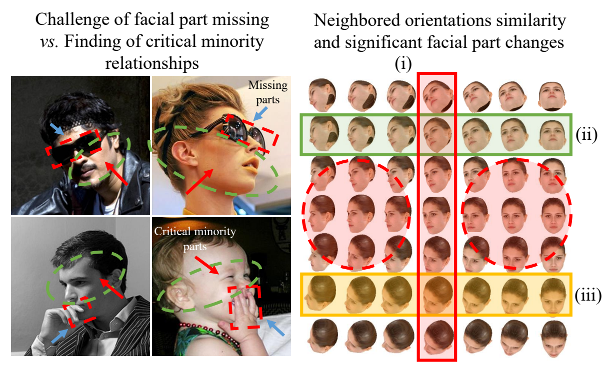

Figure 1. Left part: Missing facial parts and our finding on critical minority relationships. Although some of the facial parts are missing or occluded (marked with a red rectangle), the pose orientation still can be inferred from the existing critical minority facial parts (marked with a green circle). Right part: different head orientation images in a panoramic overview. The rectangular boxes highlight several significant facial changes, such as i) appearance of the eye on one side, ii) appearance of the nostril, and iii) overlapping of the nose and mouth. The circled areas show some regions in which the facial part features are similar.

To begin with, we firstly identify three implicit facial part relationships in head poses. First, a local similarity in specific spatial orientation exists. Inside the circled region in Fig. 1, the facial appearances are similar. Second, sev-

*: Corresponding author (Hai Liu, Youfu Li)

eral significant facial part changes are observed in specific orientations. For example, in Fig. 1, the two circled facial regions can be distinguished by a significant facial part change, which is the appearance of the right eye. Third, critical minority relationships of facial parts exist, and they can determine the orientation of a head pose despite possible occlusions. As Fig. 1 shows, if a person's mouth is occluded, the head pose can be determined by the geometric spatial relationships of the eyes, nose, and the outline of the face. In these scenarios, the remaining minor facial parts and their relationships are crucial for high-accuracy HPE.

Given the aforementioned facial part relationships, the question is how to design a model that can utilize this heuristic knowledge. The traditional CNN architecture cannot easily learn these relationships. In contrast, the Transformer architecture can effectively address this drawback of CNN. Recently, Vision Transformer (ViT) [11] 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 relationships among critical minority facial parts is reasonable. Moreover, the basic orientation regions can be well represented by learnable tokens in Transformer.

Inspired by the three findings and Transformer's properties, we propose TokenHPE, a method that can discover and leverage facial part relationships and regional similarities 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 basic orientation regions. The latent relationships between visual and orientation tokens can be learned from large HPE datasets. Then, 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. Our main contributions can be summarized as follows:

(1) Three findings are derived on head images, including facial part relationships and neighborhood orientation similarities. Furthermore, to leverage our findings and cope with challenging scenarios, a novel token-learning model based on Transformer for HPE is presented for the first time.

(2) We find that the head pose panoramic overview can be partitioned into several basic regions according to the orientation characteristics. The same number of learnable orientation tokens are utilized to encode this general information. Moreover, a novel token guide multi-loss function is designed to train the model.

(3) We conduct experiments on three widely used HPE datasets. TokenHPE achieves state-of-the-art performance with a novel token-learning concept compared with its existing CNN-based counterparts. Abundant visualizations are also provided to illustrate the effectiveness of the proposed orientation tokens.

2. Related Work

2.1. Head Pose Estimation

Existing HPE methods can be roughly divided into three categories: (1) *Euler angle regression* approaches [9, 10, 31, 42] that regress the three Euler angles progressively, (2) *extra information-utilized* approaches [1, 7, 30, 38, 39] that exploit extra facial information to facilitate HPE, and (3) *Alternative orientation parametrization* approaches [3, 15, 18, 19] that substitute Euler angle representation with other representations.

Euler angle regression approaches. The paradigm in early studies was to consider HPE as a regression problem. CNNs have been adopted for HPE [31, 42] and remained dominant for many years because convolution can efficiently reveal the visual patterns on human faces. Ruiz *et al.* [31] applied CNN to HPE in an end-to-end manner to independently predict three Euler angles by using a multi-loss network. In [42], Yang *et al.* proposed FSA-Net, which reveals aggregated features with fine-grained spatial structures and progressive stage fusions. However, becaues of the incapacity of CNN to learn the relationships among visual patterns, further facial part relationships are not explored in this category.

Extra information-utilized approaches. With graph convolutional network (GCN) being leveraged in many NLP and computer vision tasks [8, 14, 21, 41, 45], Xin *et al.* [39] proposed a novel method that learns through the facial landmark graph. However, the precision of the model depends largely on the precision of the additional landmark detector. Wu *et al.* [38] proposed a multi-task model called SynergyNet that predicts complete 3D facial geometry. Improved performance is achieved by synergistic learning of 3D landmarks and 3D morphable model parameters. 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.

Alternative orientation parametrization approaches. Most contributions to HPE in recent years have focused on alternative parametrization of head pose labels because traditional Euler angle labels inevitably have some problems at specific orientations. Geng *et al.* [15] 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 [3], Cao *et al.* proposed a vector-based head pose representation that handles the issue of discontinuity of Euler angle annnotation. Recently, Hempel *et al.* [18] proposed

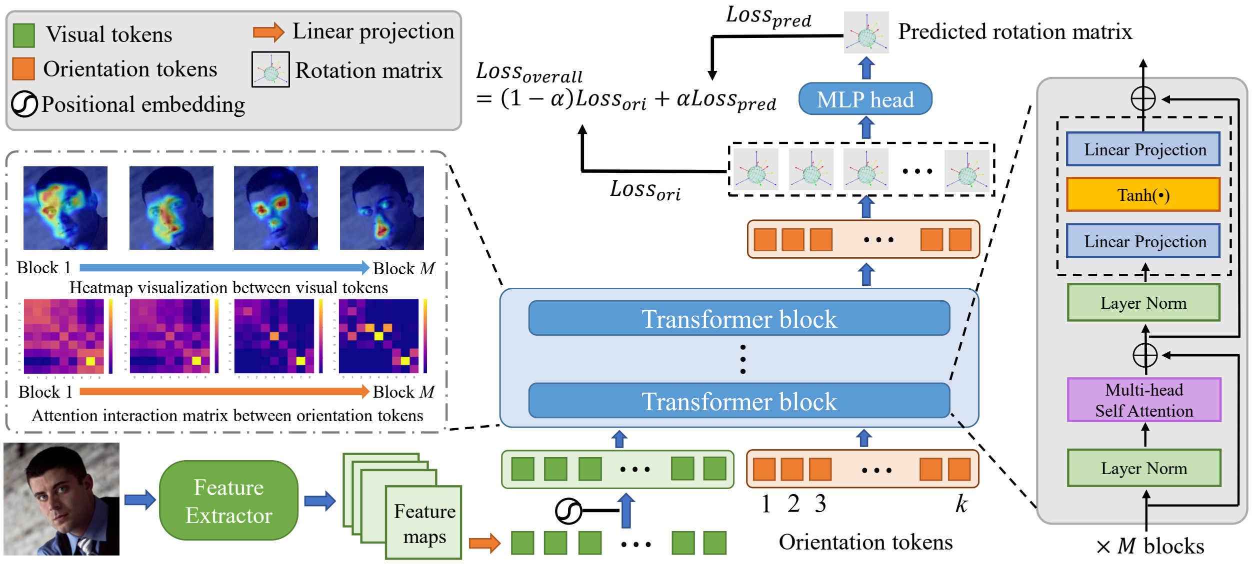

Figure 2. Architecture of the proposed TokenHPE model. The four major parts are visual token construction, orientation token construction, Transformer module, and token learning-based prediction. A given image is transformed into visual tokens and added with a positional embedding. Then, the visual tokens and learnable orientation tokens are concatenated as the input of the Transformer blocks. The orientation tokens outputted by the last Transformer block are used to predict the head poses.

a rotation matrix-based representation for HPE. The rotation matrix enables full pose regression without suffering from ambiguity problems. Although these methods have achieved impressive results, the intrinsic facial part relationships are still not fully exploited.

2.2. Vision Transformer

ViT [11] is a variant model of Transformer [34], which is originally used in NLP. In ViT , an input image is divided into patches that can be viewed as words. The success of ViT has led to its wide application in various vision tasks, including fine-grained classification [17, 25, 35], object detection [43], facial expression recognition [40], human pose estimation [23], and image segmentation [37]. Li *et al.* [23] proposed the use of learnable tokens to represent each human keypoint entity on the basis of prior knowledge. In this way, visual cue and constraint cue learning are expicitly incorporated through the Transformer architecture. In [6], Cordonnier *et al.* provided a theoretical explanation of the long-distance information learned in Transformer. Therefore, Transformer is capable to learn the facial part relationships, and neighborhood orientation similarities can be encoded into learnable orientation tokens.

3. Our Method

In this section, we first provide an overview of the proposed TokenHPE. Then, the details of the four parts of the model are elaborated. Lastly, we report the implementation details.

3.1. Overview

An overview of our method is shown in Fig. 2. The TokenHPE model consists of four parts. The first part 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 self-attention mechanism. The fourth part is token learning-based prediction. A novel token guide multi-loss function is introduced to help the orientation tokens encode general information.

3.2. Visual Token Construction

In this part, an original input RBG 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), or by directly selecting the tokens from a Transformer extractor, such as ViT [11] (Option 3). For Option 1, suppose we have an input image I with size $H \times W \times C$. The image is divided into patches with patch size $P_h \times P_w$. Then, each patch is resized into a 1D vector of size $P_h \times P_w \times C$ and a linear projection is applied to obtain a visual token. This operation can be expressed as:

$$f : p \rightarrow v \in \mathbb{R}^d, \tag{1}$$

where p refers to a 1D patch vector and v is a visual token

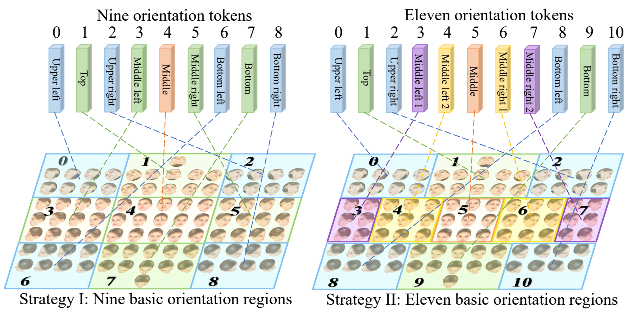

Figure 3. Construction of orientation tokens. We discover that the head pose panoramic overview can be roughly divided into several basic orientation region s according the neighbor image similarities. As the division granularity varies, the number of basic orientation regions also varies.

with a dimension of d. For Option 2, the output of the CNN extractor is considered as 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 are simply gained from the output of a Transformer extractor.

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:

$$[\texttt{visual}] = \{v_1 + pos, v_2 + pos, ..., v_n + pos\}, \quad (2)$$

where n is the number of patches. Then, we obtain n 1D vectors symbolically presented by [visual] tokens.

3.3. Orientation Token Construction

Basic orientation region partitioning. We introduce two heuristic partitioning strategies, as shown in Fig. 3. 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 Strategy II, the panoramic overview is divided into 11 regions, with a fine-grained partition in the yaw direction. A detailed description of the partition strategies is included in the supplementary material.

Orientation token. We prepend k learnable d dimensional vectors to represent k basic orientation regions. These vectors are symbolized as [dir] tokens. The [dir] tokens, together with the [visual] tokens, are accepted as the input of Transformer. In the end, the processed [dir] tokens are chosen as the output of Transformer.

3.4. Transformer Blocks and MLP Head

With the [visual] and [dir] tokens as the input, the Transformer blocks can learn the relationships between tokens. For each Transformer block, we adopt the classical structure (cf. [11, 23]), which can be briefly expressed as:

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

where MSA denotes multi-head self-attention, MLP means multi-layer perception and LN is layernorm operation. We modify the MLP module by setting the $Tanh(\cdot)$ as the activation function. 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 $\{X_1^M, X_2^M, ..., X_k^M\}$.

The orientation tokens need to be transformed to rotation matrices for training and prediction. We adopt similar transformation strategy as used in [18], which is formulated as:

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

where W is a projection matrix to obtain a 6D representation of head pose, and $\hat{R}_i$ is the predicted rotation matrix of the i-th basic orientation region. $F_{GS}(\cdot)$ denotes the Gram–Schmidt process. For more details, please refer to the supplementary material.

A set of intermediate rotation matrices $\mathcal{C} = \{\hat{R}_1, \hat{R}_2, ..., \hat{R}_k\}$ can be generated by the transformation above. In order to obtain the final prediction rotation matrix, $\mathcal{C}$ is concatenated and flattened into a vector $\tilde{R} \in \mathbb{R}^{9 \cdot k}$ as the input of the MLP head, which can be formulated as:

$$\hat{R} = F_{GS}(W_2(tanh(W_1 \cdot \tilde{R} + b_1)) + b_2), \quad (5)$$

where W_i and b_i are the parameters of the MLP head. In the training stage, the intermediate rotation matrices and the final prediction rotation matrix are used for calculating the loss for back propagation while in the prediction stage, only the prediction rotation matrix $\hat{R}$ is used for the model prediction.

3.5. Token Guide Multi-loss Function

The prediction of the proposed model is a rotation matrix representation denoted as $\hat{R}$. Suppose that the groundtruth rotation matrix is R. The geodesic distance [18] is used as the loss between two 3D rotations. The geodesic distance loss is formulated as:

$$L_g(R, \hat{R}) = cos^{-1}\left(\frac{tr(R\hat{R}^T) - 1}{2}\right). \quad (6)$$

Orientation token loss. Information can be encoded into the orientation tokens through the orientation token loss, which is defined as the geodesic distance with respect to their corresponding orientation regions. Therefore, the orientation token loss is written as:

$$Loss_{ori} = \sum_{i=1}^{k} \mathbb{I}(R, i) \cdot L_g(R, \hat{R}_i), \tag{7}$$

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

$$\mathbb{I}(R, i) = \begin{cases} 1, & if\ R\ in\ region\ i, \\ 0, & if\ R\ not\ in\ region\ i. \end{cases} \tag{8}$$

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:

$$Loss_{pred} = L_g(R, \hat{R}), \tag{9}$$

where $\hat{R}$ is the model prediction.

Overall loss. The overall loss consists of the orientation token loss and the prediction loss. It can be formulated as:

$$Loss_{overall} = \alpha Loss_{pred} + (1 - \alpha) Loss_{ori}, \tag{10}$$

where α is a hyper-parameter that balances prediction loss and orientation token loss.

3.6. Architecture Details

The three options mentioned previously can be used to obtain the visual tokens. In Option 1, the raw image patches are directly transformed into visual tokens. In the version added with a , many low-level features are utilized for prediction. In Option 2, a CNN feature extractor is added to efficiently extract low-level features, we adopt the widely used stem-net, which quickly downsamples the feature map into 1/4 input resolution in a very shallow convolutional structure [4, 33]. Option 3 is applied in our TokenHPE model by default, in which the ViT-B/16 is set 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.

3.7. Implementation Details

Pre-processing. In our experiments, the image is 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 one TITAN V GPU. All the parameters in our model are trained with random initialization.

Training. We train our TokenHPE in an end-to-end manner. The batch size is set to 64, and α is set to 0.6 by default. We train our model for 60 epochs. The learning rate is initialized as 0.00001, which is further decayed by a factor of 10 at the 20th and 40th epochs.

4. Experiments

This section describes the three datasets used for training and testing, the evaluation metrics, the experiment results and comparison with several methods, the ablation study, and the model visualization.

4.1. Datasets and Evaluation Metrics

Datasets. BIWI dataset [12] includes 15,678 images of 20 individuals (6 females and 14 males, 4 individuals are recorded twice). The head pose range covers about ±75° yaw and ±60° pitch. **AFLW2000 dataset** [47] contains 2000 images and is typically used for the evaluation of 3D facial landmark detection models. The head poses are diverse and often difficult to be detected by a CNN-based face detector. **300W-LP dataset** [47] adopts the proposed face profiling to generate about 61k samples across large poses. The dataset is usually employed as the training set for HPE.

Evaluation metric 1: Mean absolute errors of Euler angles (MAE). MAE is a standard metric for HPE. Assume a given set of groundtruth Euler angles $\{\alpha, \beta, \gamma\}$ of an image, in which α,β,and γ represent pitch, yaw, and roll angle, respectively. The predicted set of Euler angles from a model is denoted as $\{\hat{\alpha}, \hat{\beta}, \hat{\gamma}\}$. Then, MAE is defined as:

$$MAE = \frac{1}{3}(|\alpha - \hat{\alpha}| + |\beta - \hat{\beta}| + |\gamma - \hat{\gamma}|). \tag{11}$$

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

Evaluation metric 2: Mean absolute errors of vectors (MAEV). MAEV is based on rotation matrix representation. For an image, suppose that the groundtruth rotation matrix is $R = [r_1, r_2, r_3]$, where r_i is a 3D vector that indicates a spatial direction. The predicted rotation matrix from a model is denoted as $\hat{R} = [\hat{r_1}, \hat{r_2}, \hat{r_3}]$. MAEV can be formulated as:

$$MAEV = \frac{1}{3}\sum_{i=1}^{3} \|r_i - \hat{r_i}\|_1. \tag{12}$$

4.2. Comparison with State-of-the-art Methods

We compare our method with state-of-the-art methods, including Euler angle regression methods (HopeNet, FSA-Net, FAN)), extra information-utilized methods (3DDFA, Dlib, EVA-GCN, img2pose, SynergyNet), and alternative orientation parametrization methods (Quatnet, TriNet, 6DRepNet). In our two experiments, we follow the conventional protocols in FSA-Net [42]. We conduct experiments on two versions of our model: TokenHPE-v1 with nine basic orientation regions and TokenHPE-v2 with eleven basic orientation regions.

Experiment 1. In our first experiment, we follow the protocol 1 in [42] to train our model on the 300W-LP dataset

Table 1. Mean absolute errors of Euler angles and vectors on the AFLW2000 dataset. All methods are trained on the 300W-LP dataset. [1]These methods take an RGB image as the input and can be trained free from extra annotations, such as landmarks.

Methods	Extra annotation free[1]	Euler angle errors (°)				Vector errors			
		Pitch	Yaw	Roll	MAE	Left	Down	Front	MAEV
3DDFA [47]	✗	27.05	4.71	28.43	20.08	30.57	39.05	18.52	29.38
Dlib [20]	✗	11.25	8.50	22.83	14.19	26.56	28.51	14.31	23.13
FAN [2]	✗	12.3	6.36	8.71	9.12	-	-	-	-
EVA-GCN [39]	✗	5.34	4.46	4.11	4.64	-	-	-	-
SynergyNet [38]	✗	4.09	3.42	2.55	3.35	-	-	-	-
img2pose [1]	✗	5.03	3.43	3.28	3.91	-	-	-	-
HopeNet [31]	✔	7.12	5.31	6.13	6.20	7.07	5.98	7.50	6.85
FSA-Net [42]	✔	6.34	4.96	4.78	5.36	6.75	6.22	7.35	6.77
LwPosr [10]	✔	6.38	4.80	4.88	5.35	-	-	-	-
Quatnet [19]	✔	<u>5.62</u>	3.97	3.92	**4.50**	-	-	-	-
TriNet [3]	✔	5.77	4.20	4.04	4.67	**5.78**	<u>5.67</u>	**6.52**	5.99
TokenHPE-v1 (ours)	✔	5.73	4.53	4.29	4.85	6.16	5.21	6.97	6.11
TokenHPE-v2 (ours)	✔	**5.54**	4.36	4.08	<u>4.66</u>	<u>6.01</u>	**5.10**	<u>6.82</u>	**5.98**

Table 2. Mean absolute errors of Euler angles and vectors on the BIWI dataset. All methods are trained on the 300W-LP dataset.

Methods	Extra annotation free	Euler angle errors (°)				Vector errors			
		Pitch	Yaw	Roll	MAE	Left	Down	Front	MAEV
EVA-GCN [39]	✗	4.78	4.01	2.98	3.92	-	-	-	-
HopeNet [31]	✔	5.89	6.01	3.72	5.20	7.65	6.73	8.68	7.69
FSA-Net [42]	✔	5.21	4.56	3.07	4.28	6.03	5.96	7.22	6.40
Quatnet [19]	✔	5.49	4.01	<u>2.94</u>	4.15	-	-	-	-
TriNet [3]	✔	4.76	**3.05**	4.11	3.97	<u>5.57</u>	<u>5.46</u>	<u>6.57</u>	<u>5.86</u>
WHENet [46]	✔	**4.39**	3.99	3.06	<u>3.81</u>	-	-	-	-
6DRepNet [18]	✔	4.48	3.24	2.68	3.47	-	-	-	-
TokenHPE-v2 (ours)	✔	<u>4.51</u>	<u>3.95</u>	**2.71**	**3.72**	**5.41**	**5.17**	**6.23**	**5.60**

and test it on AFLW2000 and BIWI datasets . Tables 1 and 2 show the results of the first experiment. An extra column is added to indicate which methods are free from extra annotation for fair comparison. Results show that our method is on par with state-of-the-art methods on AFLW2000 dataset and achieves state-of-the-art results in MAEV on BIWI dataset. Among the compared methods, HopeNet [31] is normally considered the baseline of HPE. Compared with it (Table 1), our model achieves a 24.8% decrease in MAE and a 12.7% decrease in MAEV. TriNet [3] is a vector-based model, in which the head pose is represented by vectors. Its 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 slightly lower MAEV value, which indicates that our method is competitive to state-of-the-art methods. Some extra information-utilized methods (i.e., 3DDFA, Dlib, EVA-GCN, SynergyNet, img2pose) are also compared in Table 1. EVA-GCN [39] is a facial landmark graph-based method. A landmark detector is applied to the original image, and EVA-GCN takes the detected landmark graph as the input. The graph convolutional network learns the landmark relationships for HPE. Thus, the model result has an impressive improvement compared with the baseline. 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. 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 other methods by a large margin. The excellent performance verifies the orientation learning capacity of our proposed TokenHPE.

Experiment 2. In our second experiment, we follow the protocol 2 in [42] for fair comparison. 70% of the videos in the BIWI dataset are used for training and the others for testing. Table 3 shows the results of our method compared with those of other state-of-the-art methods that follow the same training–testing protocol. Our method outperform all other methods by a large margin both on MAE and three Euler angles. Compared to 6DRepNet [18] that uses the rotation matrix representation with a CNN backbone, our TokenHPE can learn the general regional information and facail relationships through Transformer architecture, resulting in a 6.4% improve on MAE. The similar results on two experiments show that our method is robust and stable, and its impressive results do not depend on the training dataset but on the method itself.

Table 3. Mean absolute errors of Euler angles on the BIWI dataset. The dataset is split at a ratio of 7:3 for training and testing.

Methods	Euler angle errors (°)			
	Pitch	Yaw	Roll	MAE
FSA-Net [42]	4.29	2.89	3.60	3.60
FDN [44]	3.98	3.00	2.88	3.29
Hopenet [31]	3.39	3.29	3.00	3.23
TriNet [3]	3.04	2.93	2.44	2.80
6DRepNet [18]	**2.92**	2.69	2.36	2.66
TokenHPE-v2 (ours)	3.01	**2.28**	**2.01**	**2.49**

Table 4. Effect of the feature extractor. The models are trained on the 300W-LP dataset and tested on the AFLW2000 dataset.

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

4.3. Ablation Study

Feature extractor. The visual tokens are generated from the feature extractor. Therefore, the performance of the model partially depends on the feature extractor. We conduct experiments on different feature extractors to reveal the extent to which performance is affected by the feature extractor. As shown in Table 4, we test versions with two different feature extractors and a version without a feature extractor. The results show that the feature extractor improves performance to a specific extent compared with the version with no feature extractor. The ViT feature extractor has the best performance.

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. To illustrate the effect of positional embedding, we conduct experiments on our TokenHPE model with different positional embedding types (i.e., no positional embedding, learnable positional embedding, and 2D sine positional embedding). As shown in Table 5, the model with 2D sine positional embedding demonstrates the best performance. The learnable positional embedding version has a lower prediction accuracy and 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.

Effect of the token guide multi-loss function. The proposed model is trained by a token guide multi-loss function. The hyper-parameter α in the multi-loss function controls the importance of the direction loss. When α is set to 1, the model learns the basic orientation regions by itself. As the value of α decreases, orientation token loss plays an increasingly important role in helping the model learn the directional information. The experimental results are shown in Fig. 4. When α decreases, MAE initially decreases then increases. The best result is obtained when α is set to 0.6. This situation indicates that the token guide loss indeed helps the model encode the basic orientation regions. As α decreases, the flexibility of the model is constrained, resulting in poor performance.

Table 5. Results of different positional embedding strategies. The models are trained on the 300W-LP dataset and tested on the AFLW2000 dataset.

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

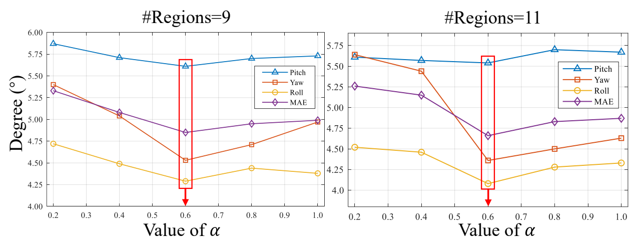

Figure 4. Effect of the value of α in the multi-loss function. The models are trained on the 300W-LP dataset and tested on the AFLW2000 dataset.

4.4. Visualization

In order to illustrate how the proposed TokenHPE explicitly utilizes orientation tokens to find the facial part relationships and orientation characteristics in the basic regions, we visualize the details during inference. We observe that our model exhibits similar behaviors for most common examples. Therefore, we randomly choose some samples from the AFLW 2000 dataset and visualize the details in Figs. 5 to 7.

Heatmap visualization. To confirm that our model can learn critical minority facial part relationships, we use Grad-CAM [32] to visualize the attention of a head pose prediction. Two representative methods (HopeNet and 6DRepNet) are adopted for a comparison with our proposed model. As Fig. 5 shows, our method can learn the crucial minority relationships of facial parts, such as the eyes, nose, and ears when the mouth is being occluded or the nose, mouth, and ears when the eyes are occluded by sunglasses. In these scenarios, the compared methods performed poorly when abundant facial information is missed. On Row 1 in Fig. 5, the attention heatmaps show that our method can find the critical minority relationships (nose, eyes, and ears). Row 2 indicates that our method can deduce the spatial location of the eyes to achieve accurate prediction compared with

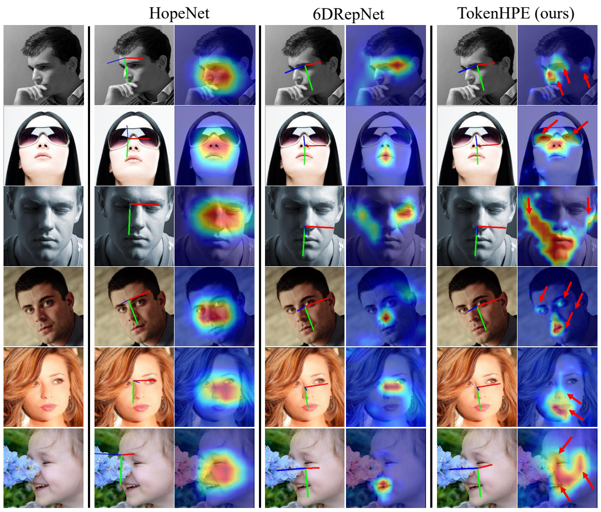

Figure 5. Heatmap visualization of three models, namely, HopeNet (left), 6DRepNet (middle), and our proposed model (right). The red-color areas mean that the model provides high attention to these facial parts. We select three challenging scenarios in which the mouth (Row 1), the eyes (Row 2), and the right half of the face (Row 3) are missing.

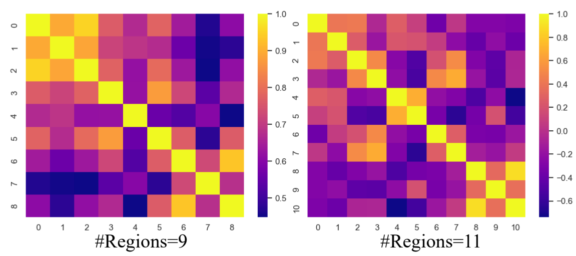

Figure 6. Cosine similarity matrix between the learned orientation tokens. (a) Strategy I: nine basic orientation regions. (b) Strategy II: eleven basic orientation regions.

the other methods that only attend to the facial parts that appear. As shown on Row 3, our method presents an impressive capability to reveal the symmetric relationships of the face even though the entire right side the face is dark. In summary, the heatmap visualization proves that our method can learn facial part relationships and can deduce the spatial relationships of facial parts.

Similarity matrix of orientation tokens. We visualize the cosine similarities of the orientation tokens. As shown in Fig. 6, the neighbor 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 is learned by the orientation tokens.

Region information learnt by orientation tokens. The attention maps of orientation tokens are visualized in Fig. 7. In the first few layers, each orientation token pays attention to almost all the other ones to construct the global context. As the network deepens, each orientation token tends to rely on its neighbor region tokens and spatial symmetric tokens to yield the final prediction. As indicated in Fig. 7, 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 7, 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 ground truth. Therefore, from the visualization shown in Fig. 7, we can conclude that our model has the ability to encode the general information of the basic regional orientation characteristics, including neighborhood similarities and symmetric properties.

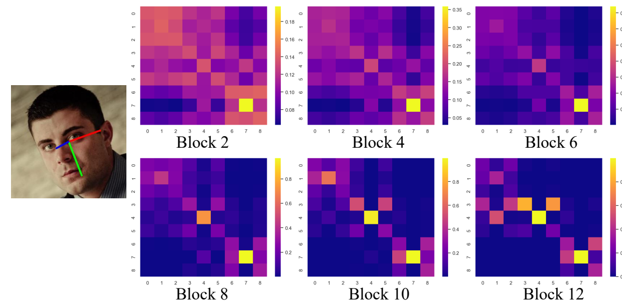

Figure 7. Attention interactions between orientation tokens in the 2nd, 4th, 6th, 8th, 10th, and 12th Transformer blocks of the proposed TokenHPE model.

5. Conclusion

In this work, we proposed a novel token-driven learning method for HPE called TokenHPE. We introduced three findings on head images, namely, neighborhood similarities, significant facial changes, and critical minority relationships. To leverage these properties of head images, we utilized the Transformer architecture to learn the facial part relationships and designed several orientation tokens according to panoramic overview partitions. Experimental results showed that TokenHPE can address the problem of ambiguity and occlusion in HPE and achieves a state-of-the-art performance compared with that of the existing method. In addition, the success of TokenHPE demonstrates the importance of orientation cues in the head pose estimation task, which was ignored by previous research. We hope this initial work can inspire further research on token learning methods for HPE.

Acknowledgement This work was supported in part by the National Key R&D Program of China under Grant 2021YFC3340802, in part by the National Natural Science Foundation of China under Grant (62277041, 62211530433, 62177018, 62173286, 62077020, 62005092, 62203024, 92167102).

References

[1] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7617–7627, 2021. 1, 2, 6

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 1, 6

[3] Zhiwen Cao, Zongcheng Chu, Dongfang Liu, and Yingjie Chen. A vector-based representation to enhance head pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1188–1197, 2021. 1, 2, 6, 7

[4] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020. 5

[5] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 383–398, 2018. 1

[6] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019. 3

[7] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 2

[8] Yongjian Deng, Hao Chen, Hai Liu, and Youfu Li. A voxel graph cnn for object classification with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1172–1181, 2022. 2

[9] Naina Dhingra. Headposr: End-to-end trainable head pose estimation using transformer encoders. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021. 2

[10] 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, 6

[11] 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. 2, 3, 4

[12] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101:437–458, 2013. 5

[13] Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc Van Gool. Real time head pose estimation from consumer depth cameras. In *Pattern Recognition: 33rd DAGM Symposium, Frankfurt/Main, Germany, August 31–September 2, 2011. Proceedings 33*, pages 101–110. Springer, 2011. 1

[14] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019. 2

[15] Xin Geng and Yu Xia. Head pose estimation based on multivariate label distribution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1837–1842, 2014. 1, 2

[16] Jinwei Gu, Xiaodong Yang, Shalini De Mello, and Jan Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1548–1557, 2017. 1

[17] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 852–860, 2022. 3

[18] Thorsten Hempel, Ahmed A Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500. IEEE, 2022. 1, 2, 4, 6, 7

[19] Heng-Wei Hsu, Tung-Yu Wu, Sheng Wan, Wing Hung Wong, and Chen-Yi Lee. Quatnet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia*, 21(4):1035–1046, 2018. 1, 2, 6

[20] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. 1, 6

[21] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2

[22] Amit Kumar, Azadeh Alavi, and Rama Chellappa. Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In *2017 12th IEEE International Conference on Automatic Face & gesture recognition (fg 2017)*, pages 258–265. IEEE, 2017. 1

[23] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11313–11322, 2021. 3, 4

[24] Hai Liu, Shuai Fang, Zhaoli Zhang, Duantengchuan Li, Ke Lin, and Jiazhang Wang. Mfdnet: Collaborative poses perception and matrix fisher distribution for head pose estimation. *IEEE Transactions on Multimedia*, 24:2449–2460, 2021. 1

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[26] Manuel Martin, Florian Van De Camp, and Rainer Stiefelhagen. Real time head model creation and head pose estimation on consumer depth cameras. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 641–648. IEEE, 2014. 1

[27] Gregory P Meyer, Shalini Gupta, Iuri Frosio, Dikpal Reddy, and Jan Kautz. Robust model-based 3d head pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3649–3657, 2015. 1

[28] Sankha S Mukherjee and Neil Martin Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015. 1

[29] Erik Murphy-Chutorian, Anup Doshi, and Mohan Manubhai Trivedi. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *2007 IEEE Intelligent Transportation Systems Conference*, pages 709–714. IEEE, 2007. 1

[30] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2017. 2

[31] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2083, 2018. 2, 6, 7

[32] 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. 7

[33] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 5

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[35] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 3

[36] Yujia Wang, Wei Liang, Jianbing Shen, Yunde Jia, and Lap-Fai Yu. A deep coarse-to-fine network for head pose estimation from synthetic data. *Pattern Recognition*, 94:196–206, 2019. 1

[37] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 3

[38] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *2021 International Conference on 3D Vision (3DV)*, pages 453–463. IEEE, 2021. 1, 2, 6

[39] Miao Xin, Shentong Mo, and Yuanze Lin. Eva-gcn: Head pose estimation based on graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1462–1471, 2021. 2, 6

[40] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610, 2021. 3

[41] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[42] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1087–1096, 2019. 1, 2, 5, 6, 7

[43] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[44] Hao Zhang, Mengmeng Wang, Yong Liu, and Yi Yuan. Fdn: Feature decoupling network for head pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12789–12796, 2020. 7

[45] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1237–1246, 2019. 2

[46] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. *arXiv preprint arXiv:2005.10353*, 2020. 6

[47] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016. 5, 6