WED-AM-064

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

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

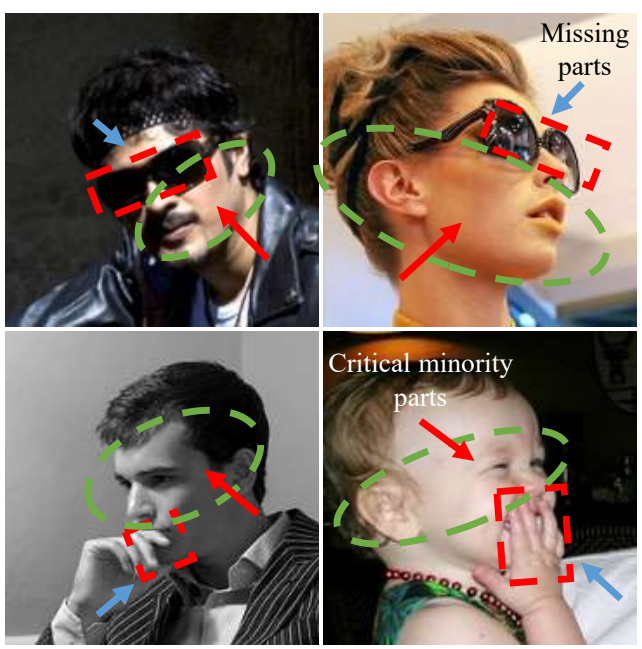
[1]Central China Normal University [2]Beijing University of Technology [3]City University of Hong Kong

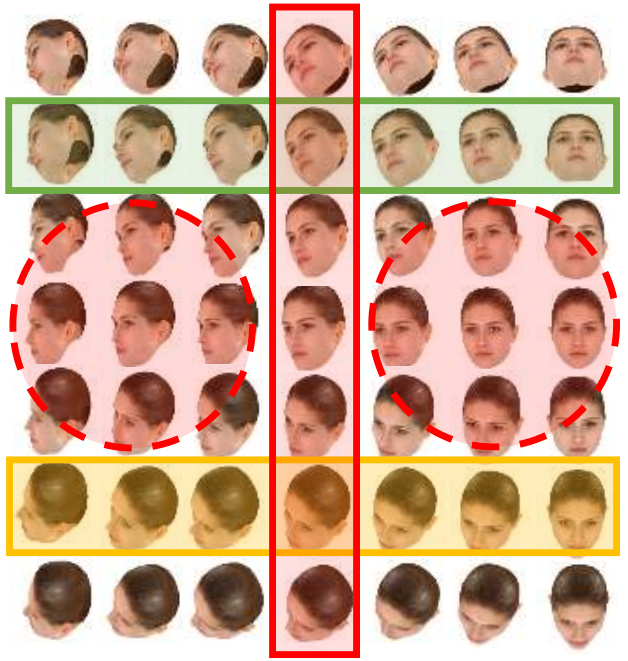
Quick Preview

Quick Preview

Findings:

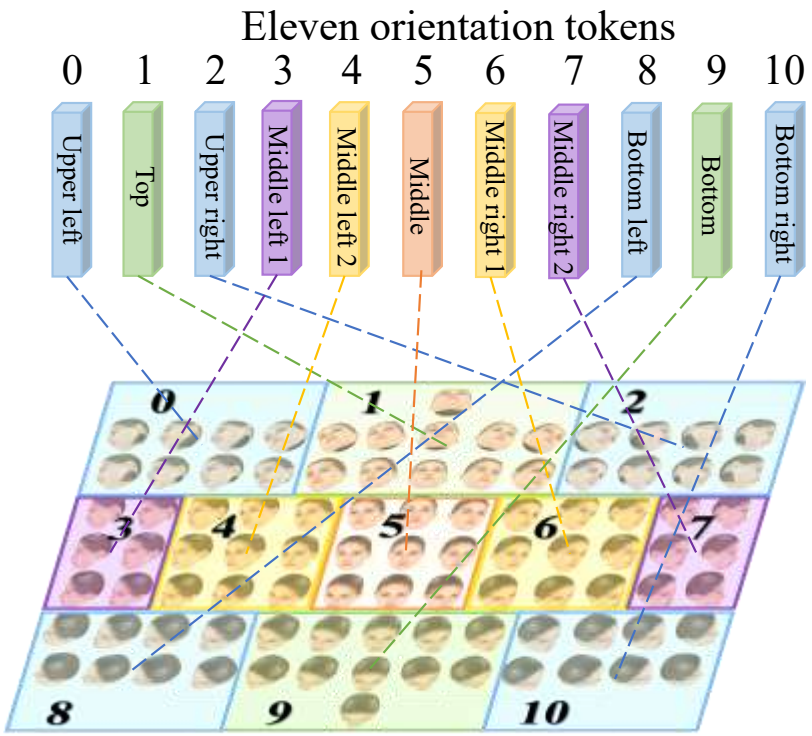
- Neighborhood similarities
- Significant facial changes
- Critical minority relationships

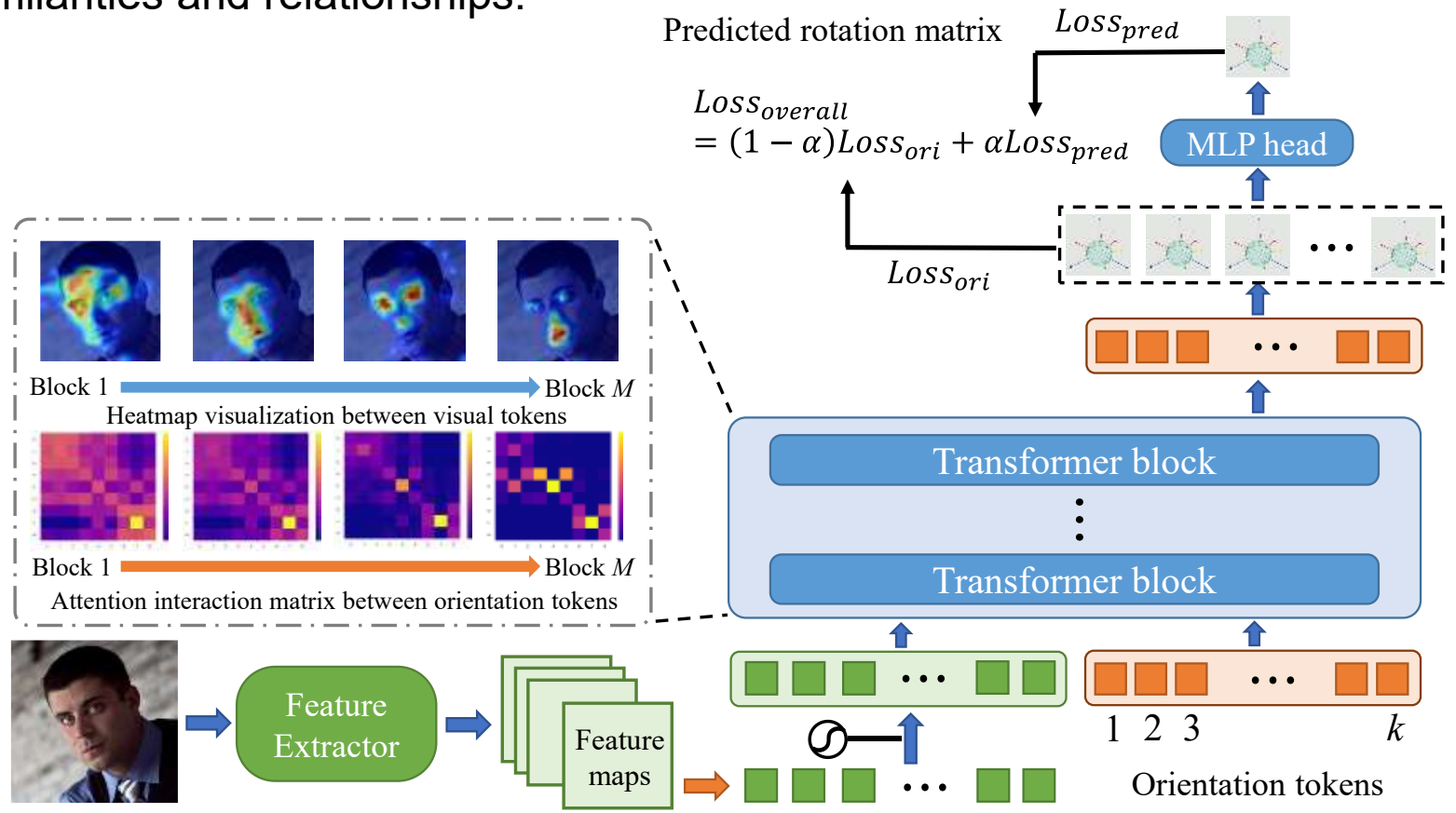




Quick Preview

- We propose a novel critical minority relationship-aware method based on the Transformer architecture
  - We design several orientation tokens to explicitly encode the basic orientation regions
  - A novel token guide multi-loss function is designed to guide the orientation tokens as they learn the desired regional similarities and relationships.





Quick Preview

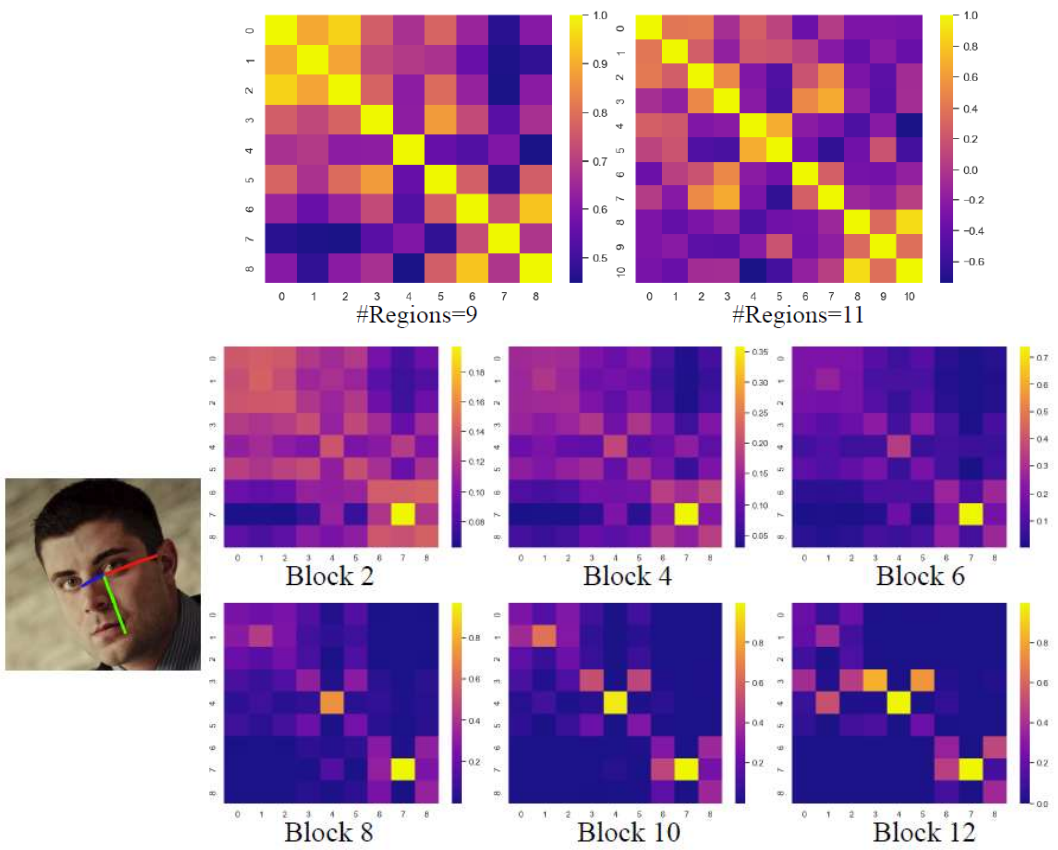
- Experiments show that our method achieves better performance compared with state-of-the-art methods.

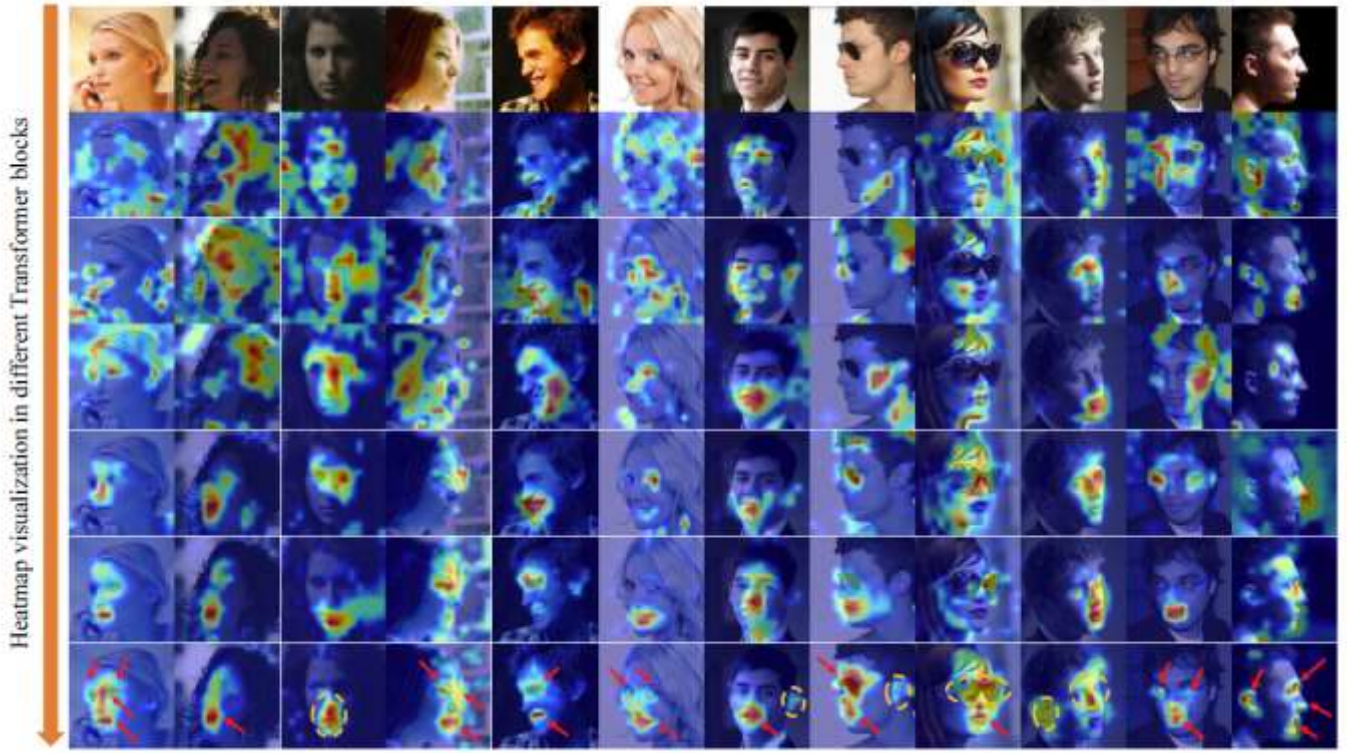
Table 1. Mean absolute errors of Euler angles and vectors on the AFLW2000 dataset. All methods are trained on the 300W-LP dataset.
[1]These methods take an RGB image as the input and can be trained free from extra annotations, such as landmarks.

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]	✔	5.62	3.97	3.92	**4.50**	-	-	-	-
TriNet [3]	✔	5.77	4.20	4.04	4.67	**5.78**	5.67	**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	4.66	6.01	**5.10**	6.82	**5.98**

Quick Preview

- We conduct experiments to verify that the proposed orientation tokens can encode the facial part relationships and orientation characteristics in the basic regions





Introduction

Introduction

Challenges:

- Extreme head pose randomness
- Serious occlusions

Introduction

Findings:

Intrinsic facial part relationships:

- **Neighborhood similarities**
- Significant facial changes
- Critical minority relationships

Introduction

Findings:

Intrinsic facial part relationships:

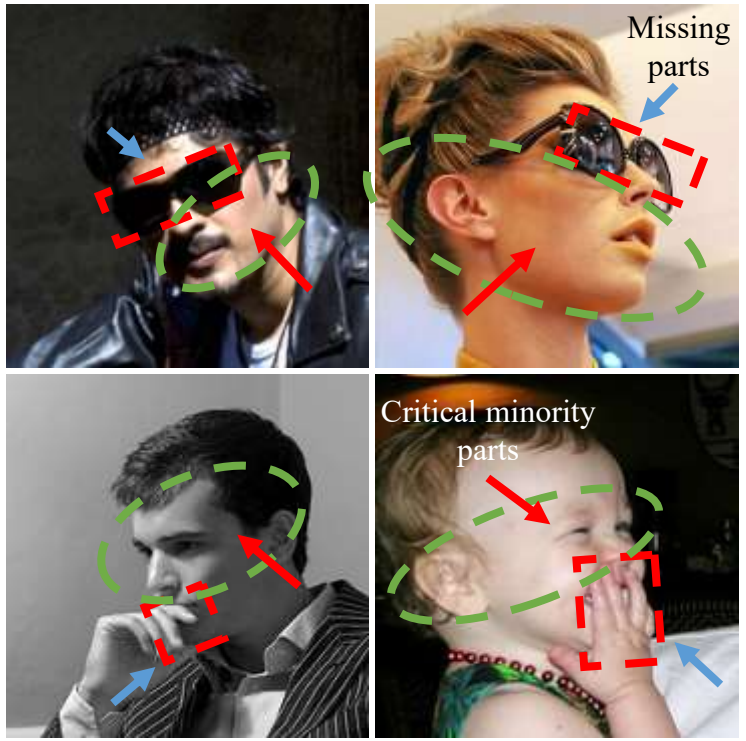
- Neighborhood similarities
- **Significant facial changes**
- Critical minority relationships

Introduction

Findings:

Intrinsic facial part relationships:

- Neighborhood similarities
- Significant facial changes
- **Critical minority relationships**



Introduction

Three findings on intrinsic facial part relationships

- Neighborhood similarities
- Significant facial changes
- Critical minority relationships

Basic orientation region partitioning

- Divide the panoramic overview into basic regions based on intrinsic facial part relationships

Token learning methodology for HPE

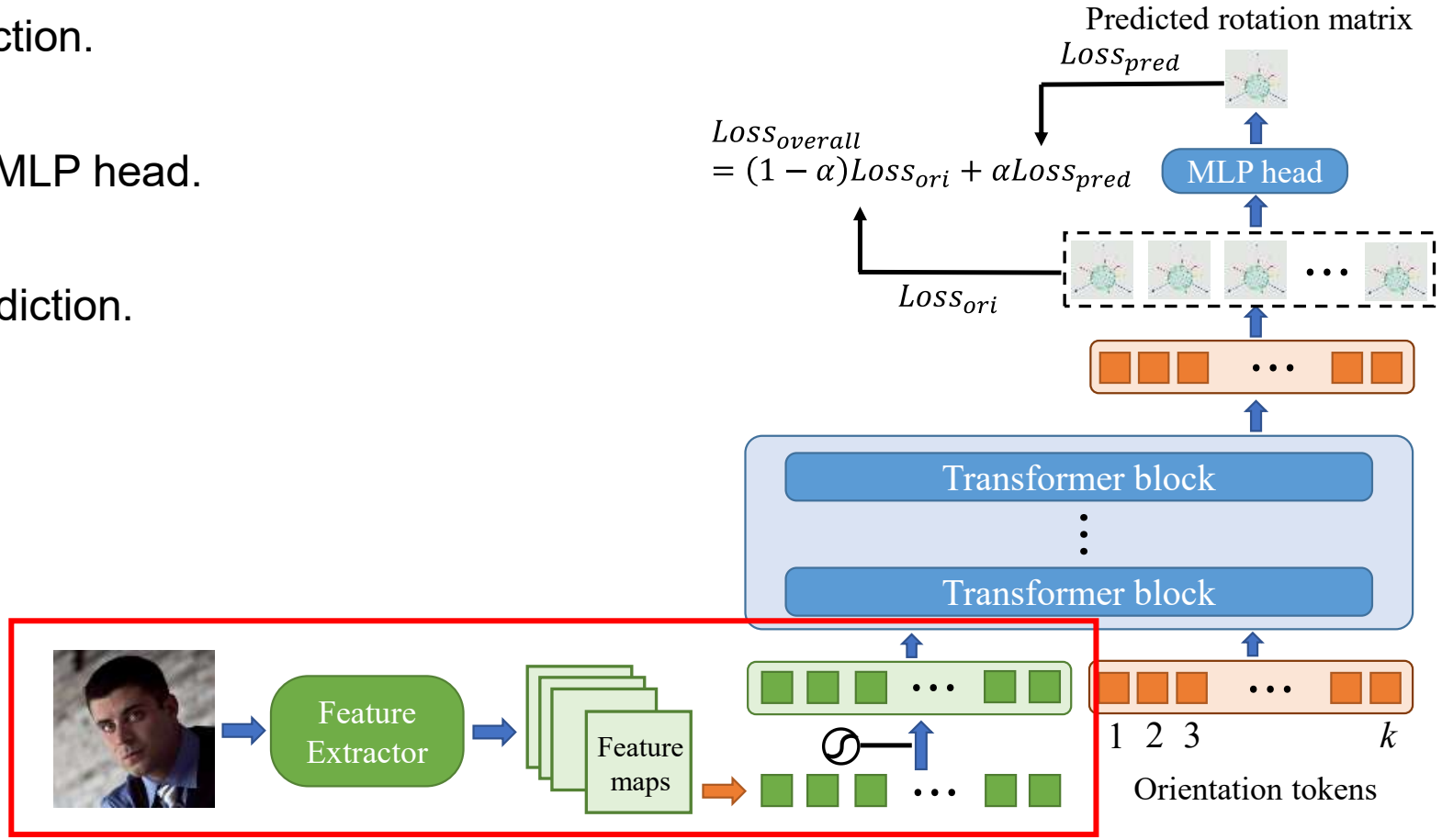
- Construct learnable orientation tokens according the Basic orientation region partitioning.
- Design a token guide multi-loss function to guide the orientation tokens learn the desired regional similarities and relationships

Methodology

Methodology

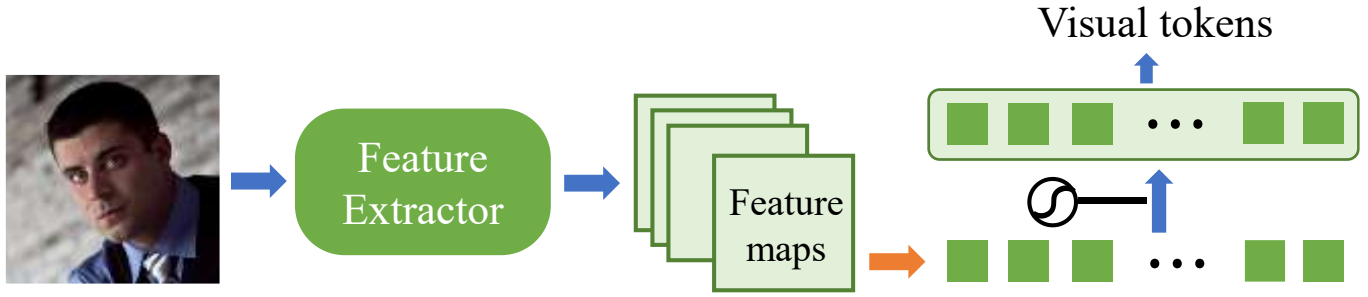
The TokenHPE model consists of four parts:

- **Visual token construction.**
- Orientation token construction.
- Transformer module and MLP head.
- Token learning-based prediction.



Methodology

Visual Token Construction



An original input RBG image is transformed into visual tokens. This operation can be expressed as:

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

where p refers to a 1D patch vector and v is a visual token with a dimension of d.

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:

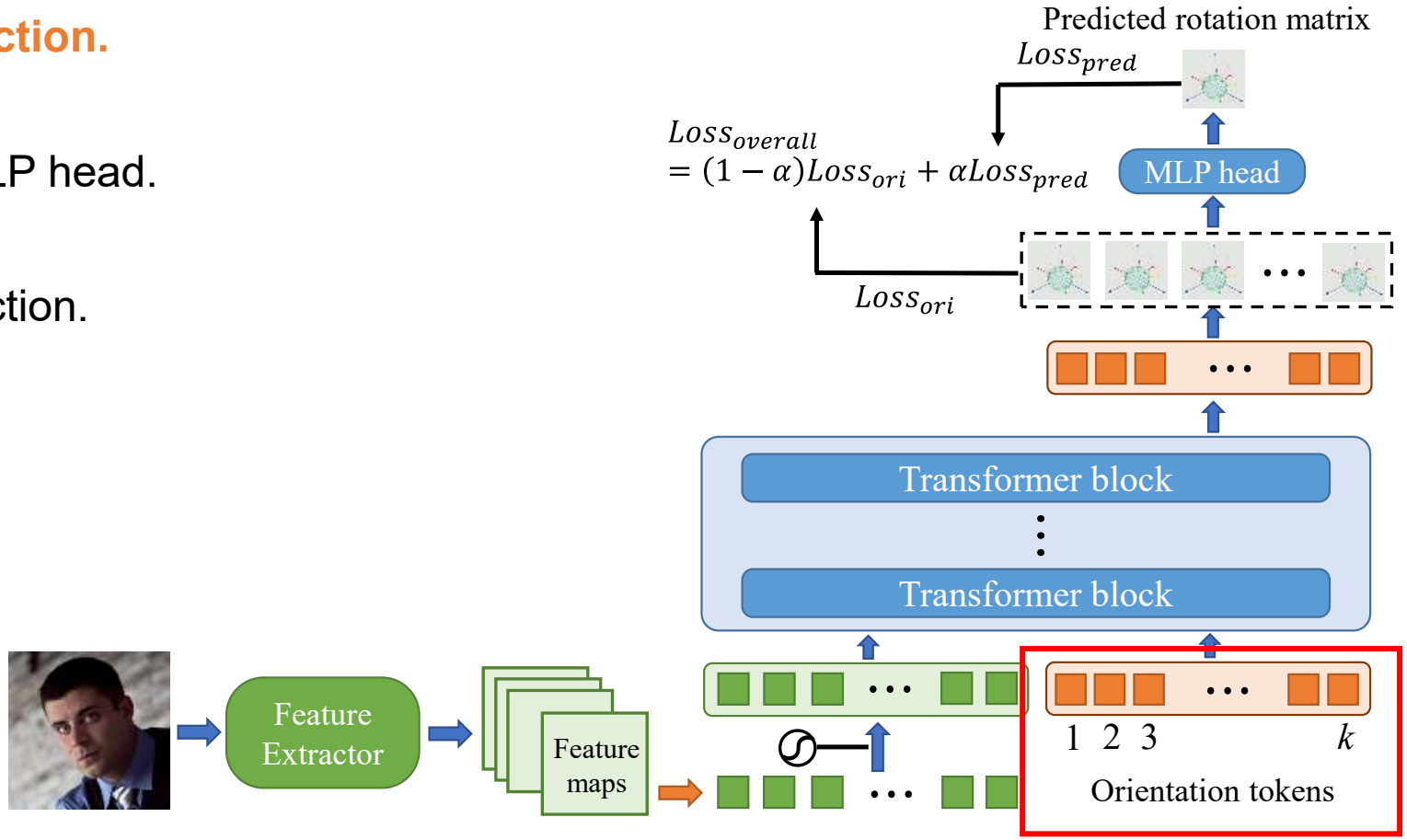
$$[\text{visual}] = \{v_1 + pos, v_2 + pos, \cdots, p_n + pos\},$$

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

Methodology

The TokenHPE model consists of four parts:

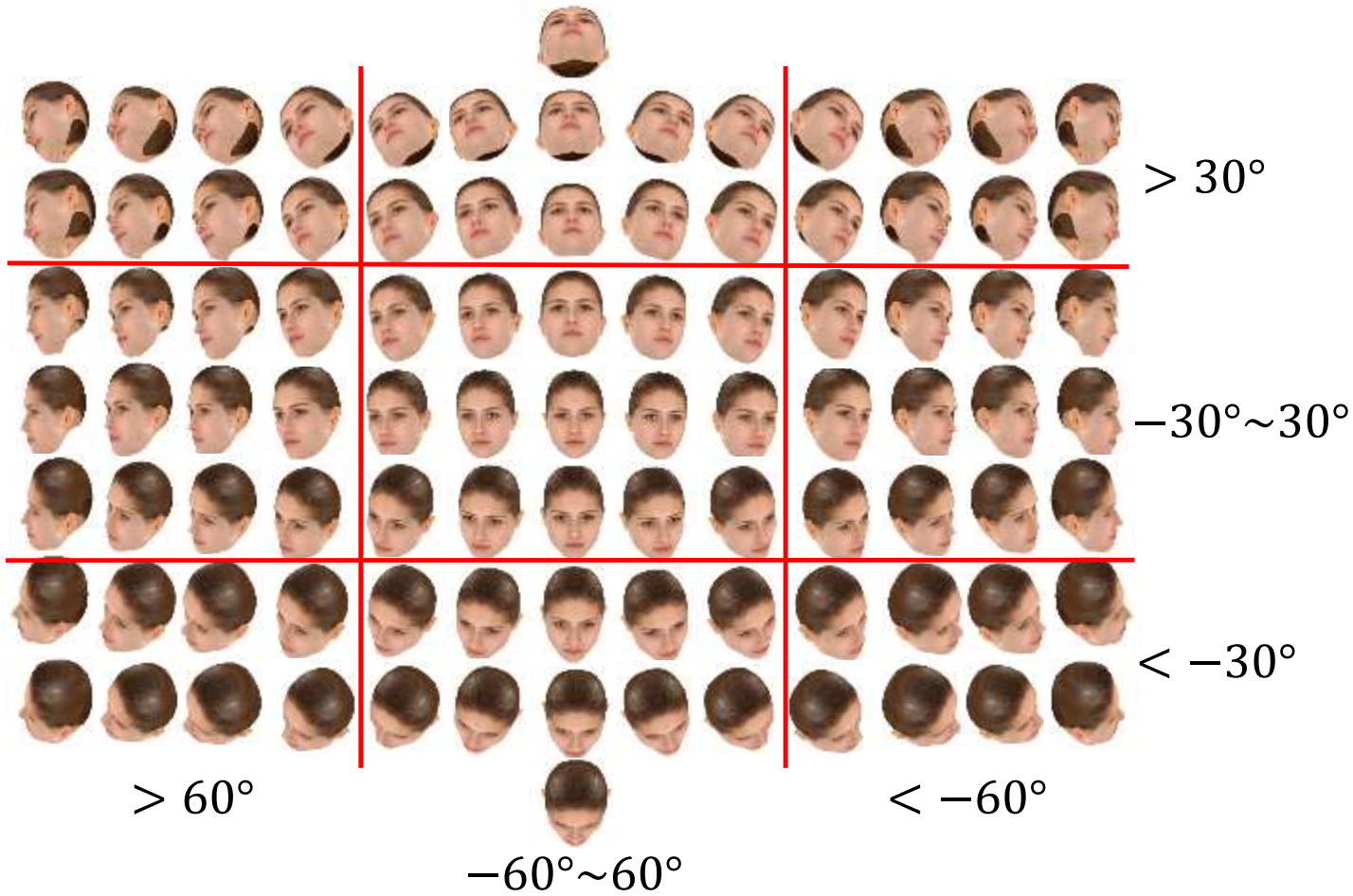
- Visual token construction.
- **Orientation token construction.**
- Transformer module and MLP head.
- Token learning-based prediction.



Methodology

Orientation Token Construction

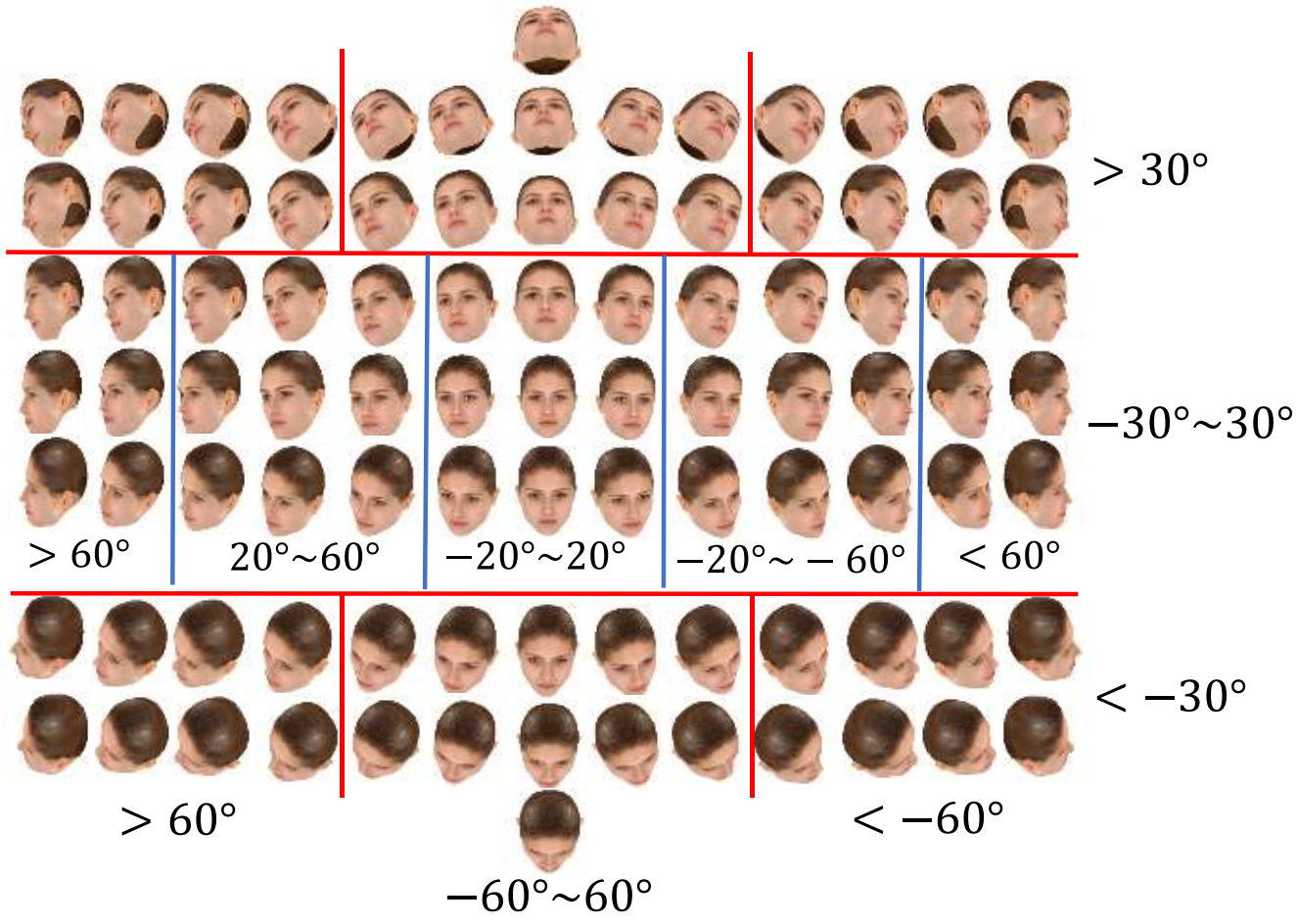
- Partition strategy I with nine basic orientation regions



Methodology

Orientation Token Construction

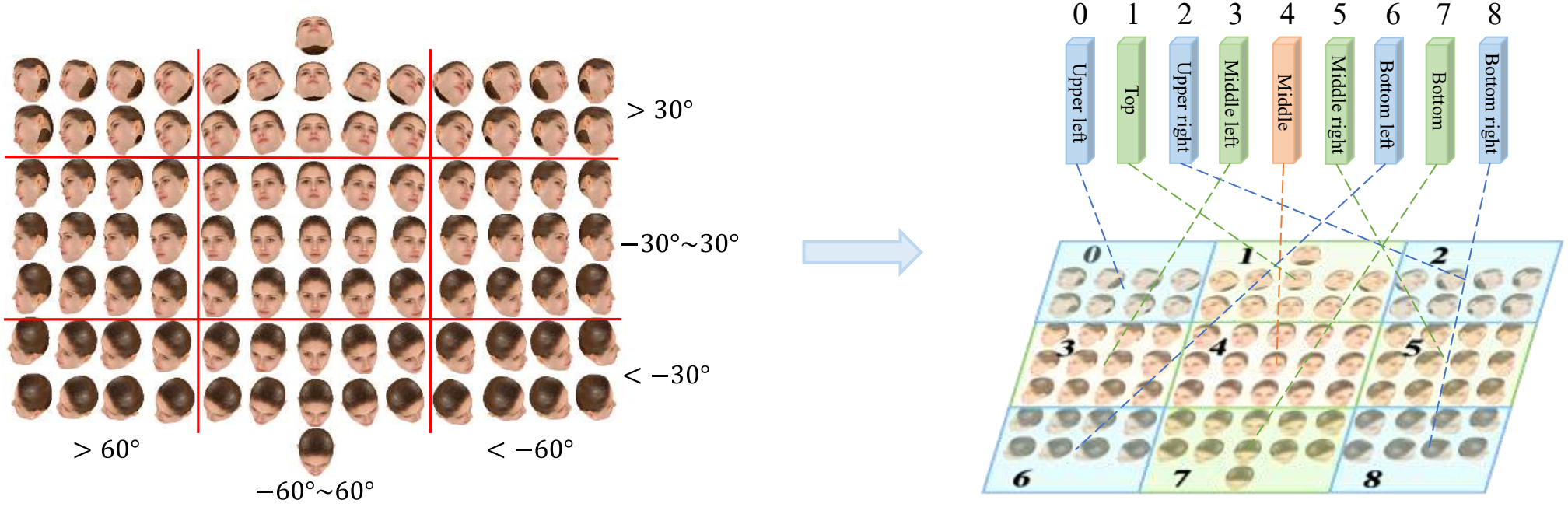
- Partition strategy II with 11 basic orientation regions.



Methodology

Orientation Token Construction

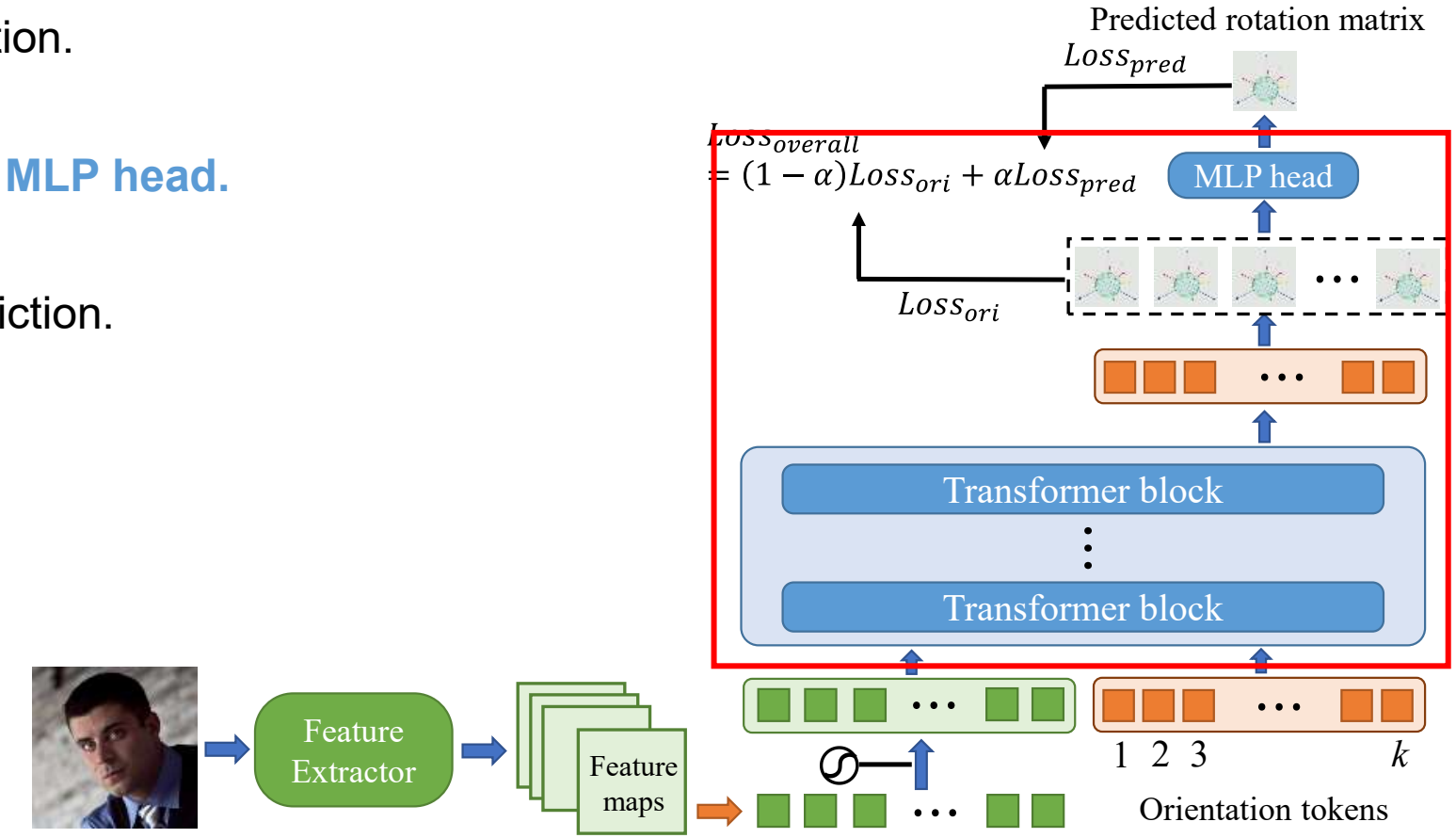
We prepend k learnable d dimensional vectors to represent k basic orientation regions. These vectors are symbolized as [dir] tokens.



Methodology

The TokenHPE model consists of four parts:

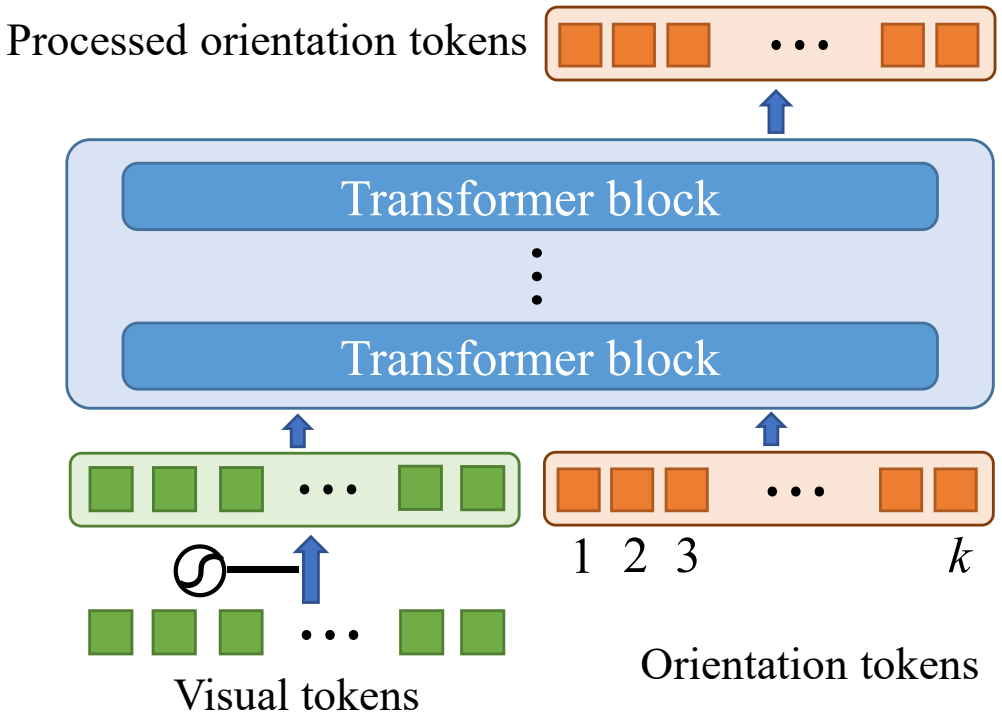
- Visual token construction.
- Orientation token construction.
- **Transformer module and MLP head.**
- Token learning-based prediction.



Methodology

Transformer Blocks

- The [dir] tokens, together with the [visual] tokens, are accepted as the input of Transformer.
- After the last Transformer layer, the [dir] tokens are selected as the output, whereas the [visual] tokens are not used in the following steps.



Methodology

MLP head

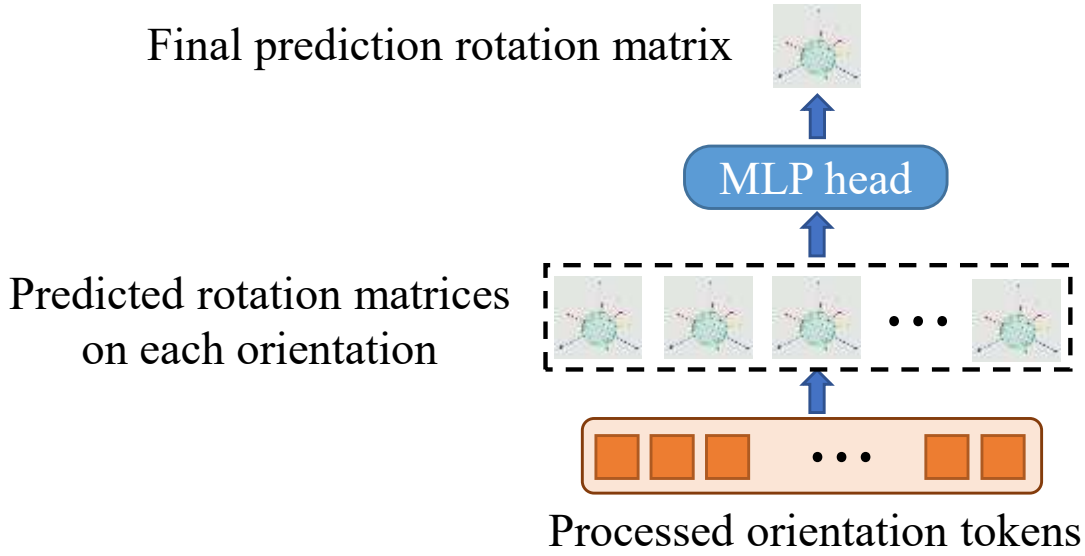
- First, a linear projection is applied to each [dir] token to obtain a 6D representation. Then the orientation matrix is obtained by the Gram–Schmidt process.

$$\hat{R}_i = F_{GS}\left(WX_i^M\right),$$

- A set of intermediate rotation matrices $\mathcal{C} = \{\hat{R}_1, \hat{R}_2, \cdots, \hat{R}_k\}$ 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 $\tilde{R}$ as the input of the MLP head.

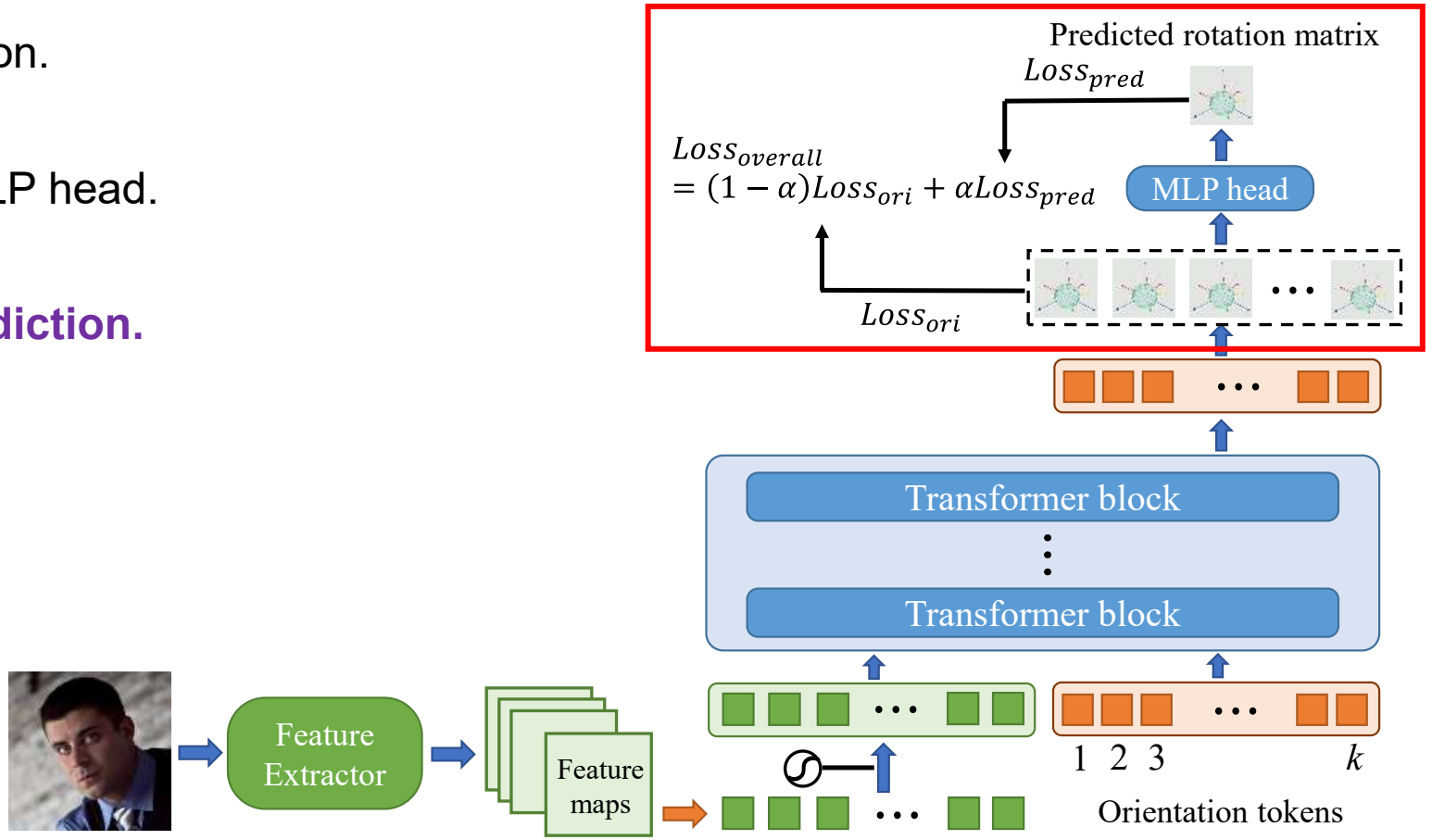
$$\hat{R} = F_{GS}\left(W_2\left(\tanh\left(W_1 \cdot \tilde{R} + b_1\right)\right) + b_2\right),$$



Methodology

The TokenHPE model consists of four parts:

- Visual token construction.
- Orientation token construction.
- Transformer module and MLP head.
- **Token learning-based prediction.**



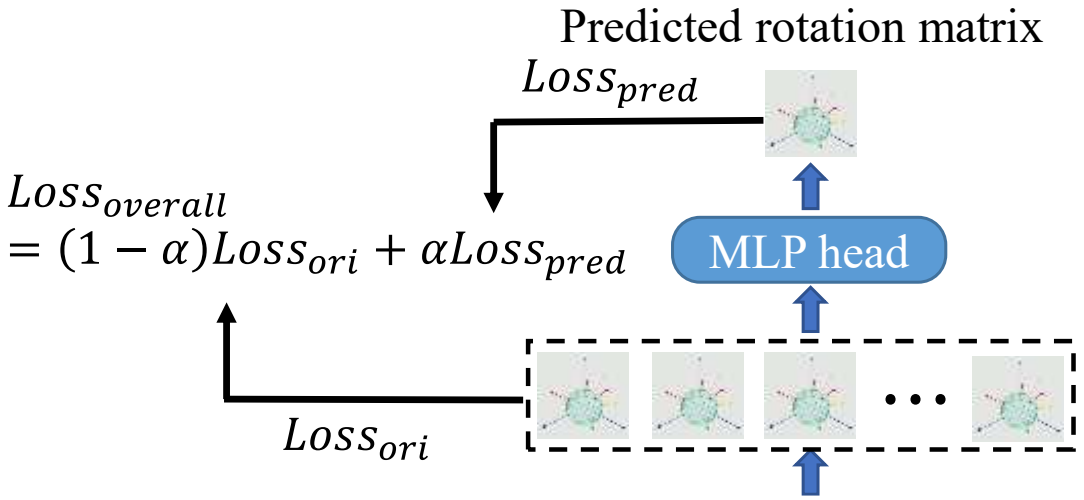
Methodology

Token Guide Multi-loss Function

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

$$Loss_{overall} = \alpha Loss_{pred} + (1 - \alpha) Loss_{Ori},$$

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



Methodology

Token Guide Multi-loss Function

Geodesic distance loss: 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 is used as the loss between two 3D rotations. The geodesic distance loss is formulated as:

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

Methodology

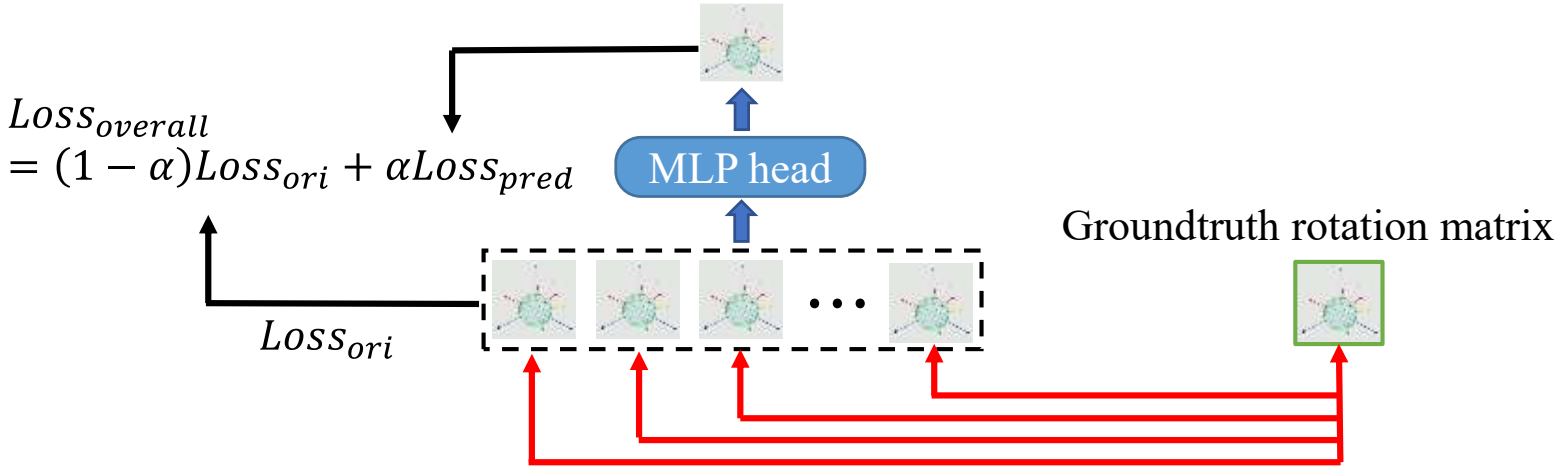
Token Guide Multi-loss Function

Orientation token loss. Information can be encoded into the orientation tokens through the orientation token loss.

$$Loss_{Ori} = \sum_{i=1}^{k} \mathbb{I}(R, i) \cdot L_g\left(R, \hat{R}_i\right),$$

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

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



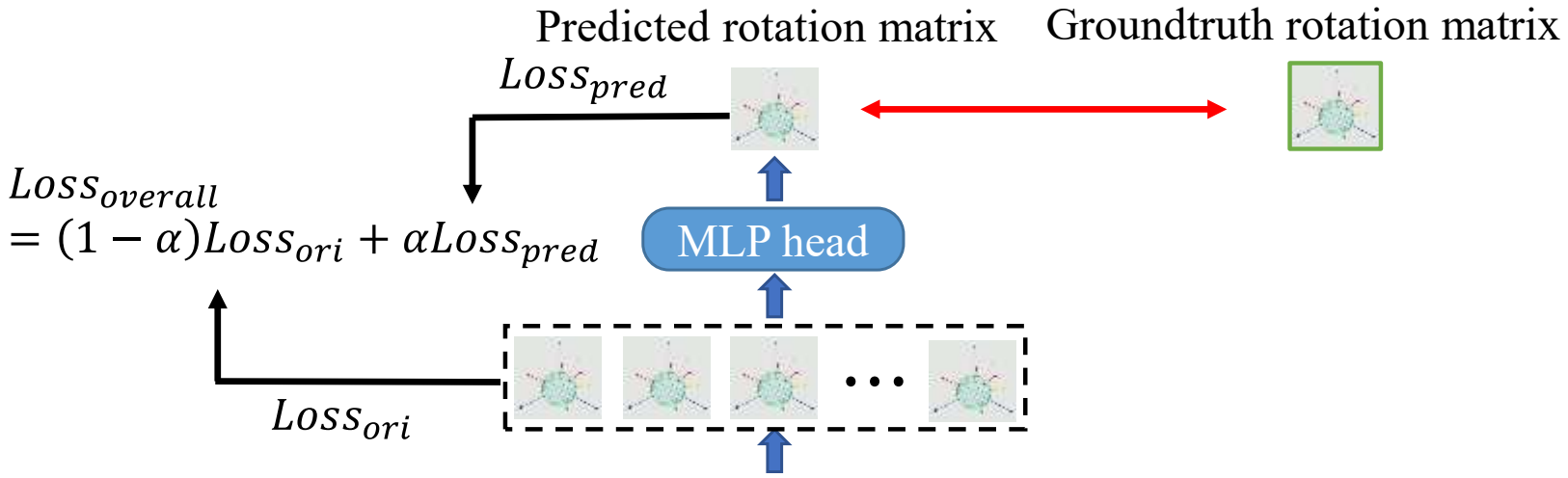
Methodology

Token Guide Multi-loss Function

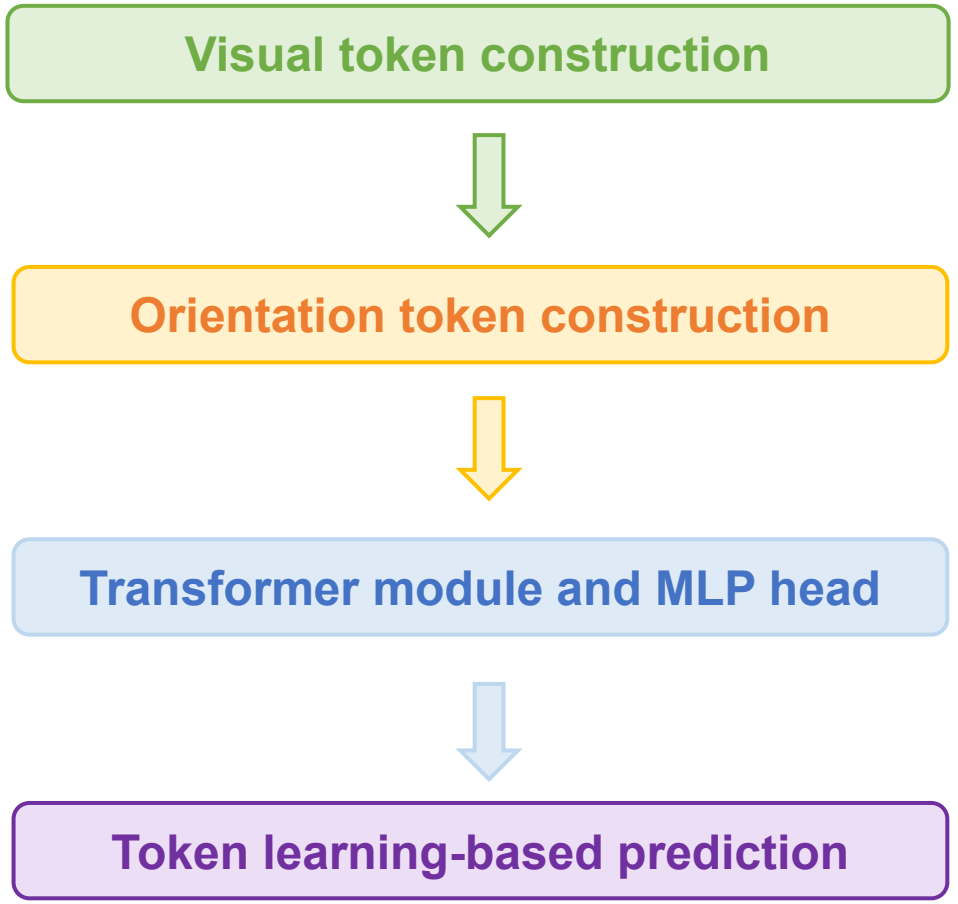
Prediction loss. The predictions from the orientation tokens are aggregated to form the final prediction of our model. This is optimized by the prediction loss, which is formulated as:

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

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



Methodology



Experimental results and Visualization

Experimental results

Datasets

- **BIWI dataset** 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** 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** adopts the proposed face profiling to generate about 61k samples across large poses. The dataset is usually employed as the training set for HPE.

Experimental results

Evaluation metrics

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

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

Evaluation metric 2: Mean absolute errors of vectors (MAEV). MAEV is based on rotation matrix representation. For an image, suppose that the ground truth 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 $\hat{R} = [\hat{r}_1, \hat{r}_2, \hat{r}_3]$.

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

Experimental results

Evaluation protocol 1: All models are trained on 300-LP dataset and tested on AFLW2000 dataset and BIWI dataset, respectively.

Evaluation protocol 2: All models are trained and tested on BIWI dataset with a 7:3 train-test split.

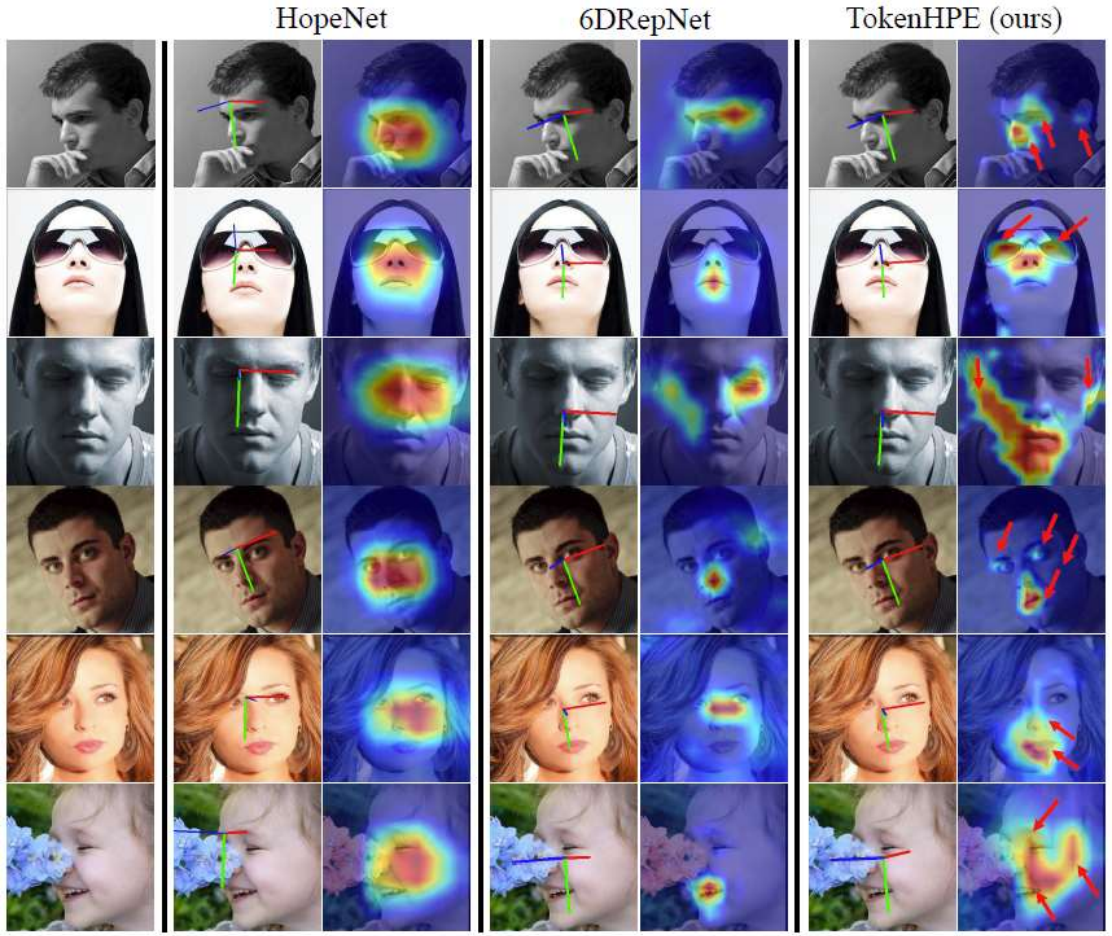
Table 1. Mean absolute errors of Euler angles and vectors on the AFLW2000 dataset. All methods are trained on the 300W-LP dataset.
[1]These methods take an RGB image as the input and can be trained free from extra annotations, such as landmarks.

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]	✔	5.62	3.97	3.92	**4.50**	-	-	-	-
TriNet [3]	✔	5.77	4.20	4.04	4.67	**5.78**	5.67	**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	4.66	6.01	**5.10**	6.82	**5.98**

Visualization

Heatmap visualization

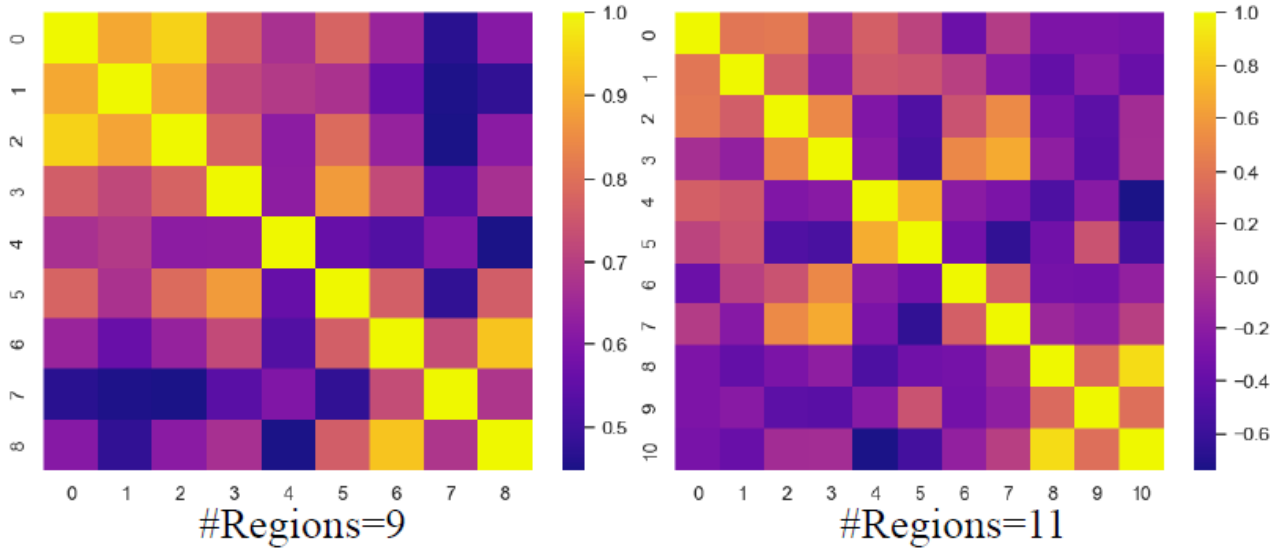
We visualize the attention of head pose predictions to confirm that our model can learn critical minority facial part relationships.



Visualization

Similarity matrix of orientation tokens

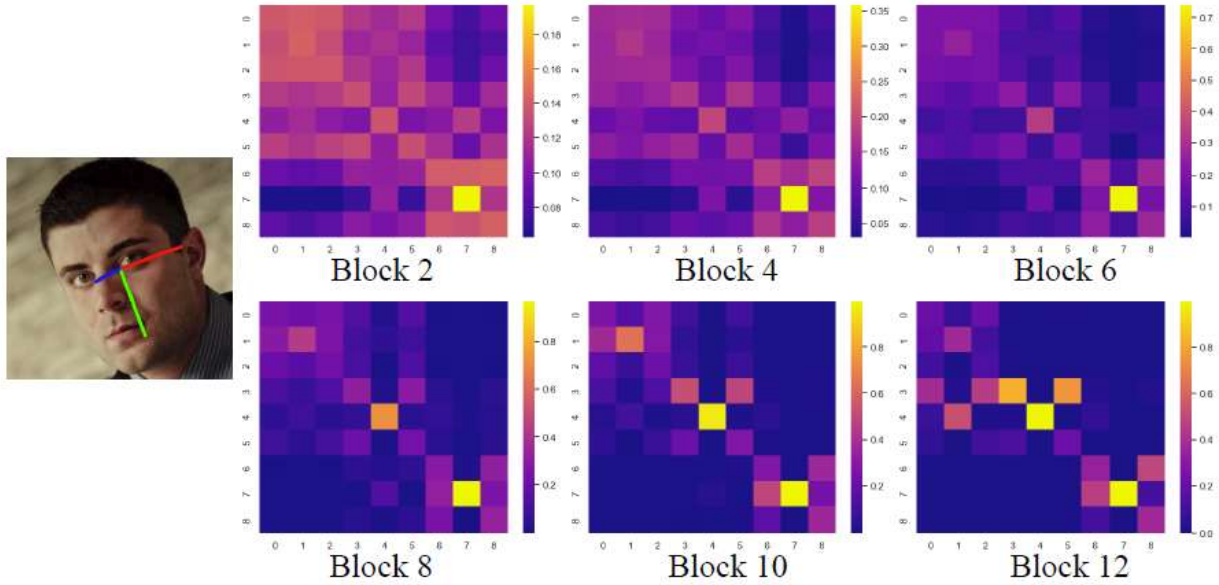
- The neighbor orientation tokens are highly similar.
- The orientation tokens that represent symmetric facial regions have higher similarity scores than the tokens that represent the other unrelated regions.



Visualization

Region information learnt by orientation tokens

- 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.
- At the deeper Transformer blocks, the attention score is higher between neighbor regions (near diagonal) and symmetric regions.

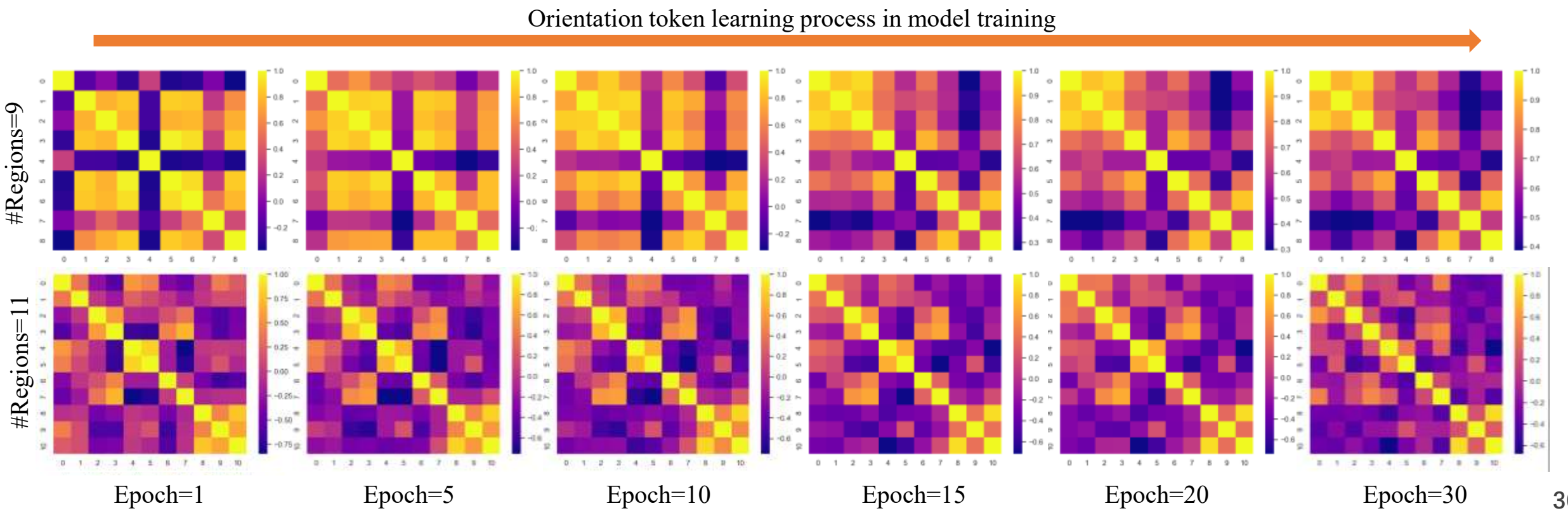


Visualization

Orientation token learning during training

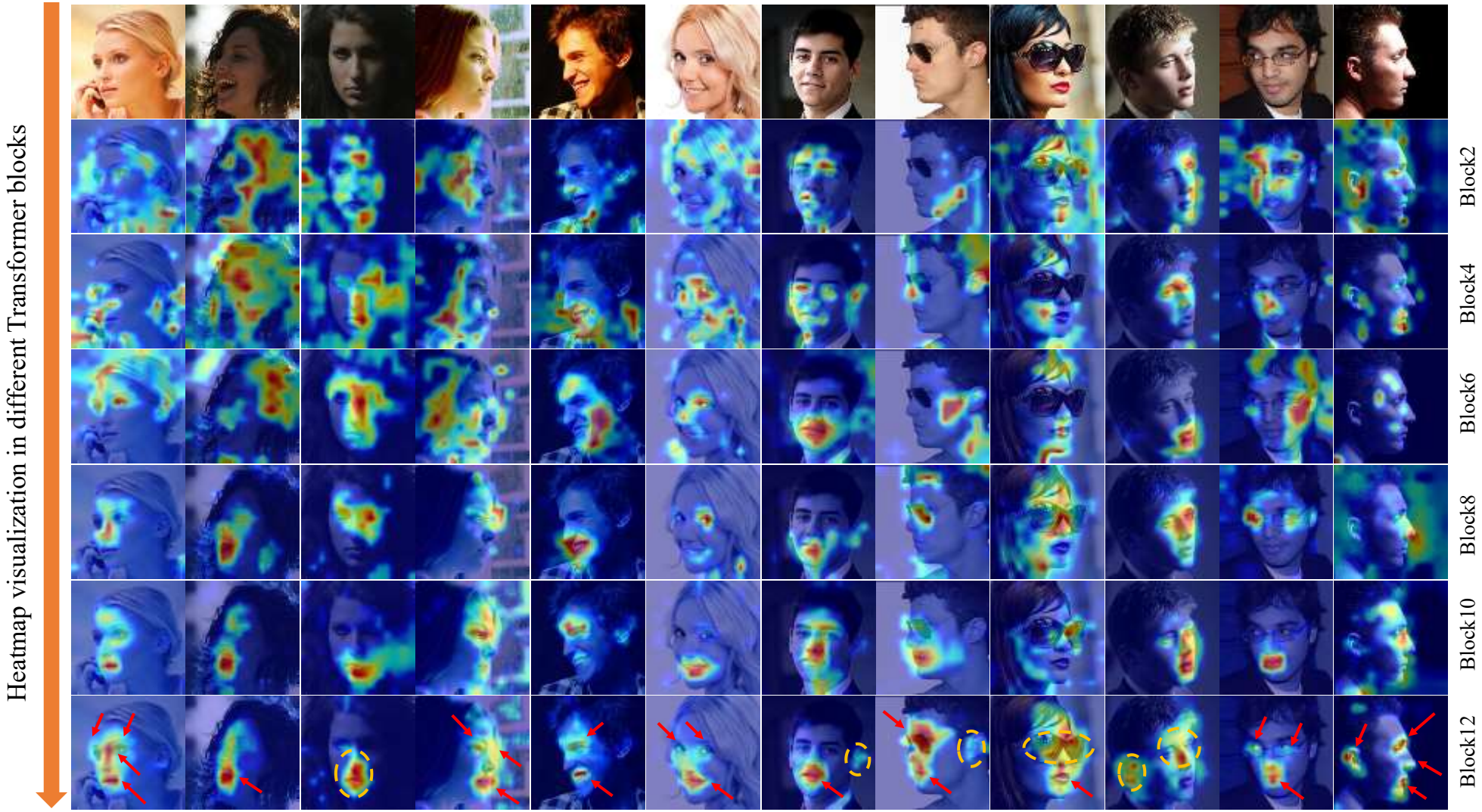
- As the training epochs increase, general information is learned gradually by the orientation tokens. The orientation relationships can be observed in the later training epochs.

- The similarity scores are higher in the neighbored regions and spatial symmetric regions.



Visualization

Heatmap visualization in the inference stage



Conclusion

Conclusion

- 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.

- In addition, the success of TokenHPE demonstrates the importance of orientation cues in the head pose estimation task. This initial work shed light on further research on token learning methods for HPE.

Thank you for listening!