GLPose: Global-Local Attention Network with Feature Interpolation Regularization for Head Pose Estimation of People Wearing Facial Masks

Hsueh-Wei Chen[1]
r09922050@ntu.edu.tw
Yi Chen[1]
d10922004@csie.ntu.edu.tw
Pei-Yung Hsiao[2]
pyhsiao@nuk.edu.tw
Li-Chen Fu[1]
lichen@ntu.edu.tw
ZI-RONG DING[3]
rong@artc.org.tw

[1] Department of Computer Science and Information Engineering
National Taiwan University
Taiwan, ROC

[2] Department of Electrical Engineering
National University of Kaohsiung
Taiwan, ROC

[3] Automotive Research & Testing Center (ARTC)
Taiwan, ROC

Abstract

To precisely estimate head poses based on RGB images is essential and useful for many applications, such as understanding the vehicle drivers' status for driving safety, and passengers' action conditions. Recently, due to the impact of the COVID-19 pandemic, people are required to wear masks in almost all public places, sometimes even in a vehicle, but the existing research works on head pose estimation have become more challenging when the face is occluded. To tackle this issue, we propose a novel siamese structure network integrating the global-local attention mechanisms with data augmentation and a multi-task learning strategy. Specifically, we initially incorporate data augmentation for synthesizing facial masks on human faces and landmark prediction in the training stage to help the model be generalized and robust. Next, a global-local attention mechanism is designed so that the relationship in whole feature maps can be learned and the critical spatial-channel information can be enhanced to obtain a better feature representation. Lastly, the feature interpolation regularization module utilizes pairs of feature embedding from the siamese network to optimize the feature embedding. To validate our proposed work, the proposed method is evaluated on AFLW2000, BIWI, and MAFA datasets. Extensive experiments show that our method can achieve highly promising performance on those public datasets.

1 Introduction

In recent years, head pose estimation task have received more attention in computer vision community. This task aims to predict the three-dimensional angular information (yaw, pitch,

and roll) of human faces from images or videos. It is essential and useful to accurately estimate head poses for many applications, such as identifying vehicle drivers' status or passengers' action condition in human-vehicle interaction [22] systems, so that some necessary warming or assisting measures can be issued for driving safety or enjoyable ride. Instead of using depth images to implement the head pose estimation, the work using RGB images is more valuable since RGB images are easier to obtain and can easily be applied to real-world applications. The research works [5, 23, 27, 30] based on deep convolutional neural network using RGB images lately have made significant progress.

However, the facial mask occlusion situation for head pose estimation is still one of the challenging problems in real-world conditions. There are two reasons why the previous research works perform worse under those situations. Firstly, they did not design the specific modules in the model architecture to discriminate useful information from RGB images. Secondly, the existing datasets [2, 31] have fewer cases with mask occlusion, so the the data-driven model can not learn the better feature representation to predict the accurate head pose.

Considering the above issues, we propose the Global-Local Attention network with the siamese structure for head pose estimation. The main contribution of this paper are summarized as follows:

- In this paper, we introduce a network with the global-local attention mechanism and the multi-task learning strategy trained on the synthesized facial mask images and non-facial mask images to learn the discriminative and useful information.
- We propose the feature interpolation regularization module in the siamese network structure to optimize the embedded features generated from pairs of feature embedding.
- Extensive experiments show that our proposed method have achieved competitive performance compared to state-of-the-art methods on public datasets.

2 Related Work

In this section, we introduce the landmark-based methods, landmark-free methods and rotation representation for head pose estimation. Moreover, we present the attention mechanisms that are widely applied to computer vision.

Landmark-based methods: In these approaches, 2D facial landmarks are predicted first and then utilized for estimating head poses. The simple method in 3D vision technique is to solve the correspondence between 2D facial landmarks predicted by existing landmark detector [1, 15] and 3D head model through Perspective-n-Point (PnP) [8] algorithm. However, it highly depends on the generic head model and the accuracy of facial landmarks. Moreover, those results are generated independently and obtained suboptimal solutions. Therefore, approaches such as KEPLER [17] refines the facial landmarks iteratively by cascade regressor and predicts head pose as well. 3DDFA_V2 [10] optimizes the network for regression task of predicting 3DMM parameters including rotation information. OsGG [21] proposes the end-to-end model combining the Convolutional Neural Network (CNN) and Graph Convolutional Network (GCN) to regress 2D facial landmarks and head pose angles.

Landmark-free methods: Landmark-free methods directly predict the head pose without the requirement of the 2D facial landmark information from RGB images. HopeNet [23]

combines ResNet50 with multi-loss techniques where each angle has binned classification and regression for its individual loss. FSA-Net [27] is the lightweight model adopting attention mechanisms with the stage-wise regression. WHENet [30] extends the yaw angle to full 360 degree range with designed loss. FDN [28] proposes a feature decoupling network which obtains compact and distinct features for each pose angle. Inspired by the success of Transformer [25] in computer vision, HeadPosr [4] and LwPosr [5] utilize transformer encoders in the network to predict head pose estimation. Nowadays, landmark-free methods still outperform landmark-based methods.

Rotation representation: The representation of a rotation in a 3D world can take many forms. For the head pose estimation task, Euler angle and quaternion are commonly predicted by deep neural network in [3, 13]. However, both representation have the ambiguity problem shown in [29]. Additionally, it demonstrates that a continuous representation of rotation in 3D space can be obtained using at least the five-dimensional representation, which is more suitable for learning. Therefore, research works [2, 12, 20] utilize the rotation matrix as the rotation representation for model training. TriNet [2] adopts additional orthogonality loss to regularize the predictions. 6DRepNet [12] uses the geodesic loss to calculate the distance between two rotation matrices.

Attention mechanism: It is well known that many works have applied attention mechanisms to their models to improve performance. SENet [14] is the first attention network that proposed the channel attention. The later-on CBAM [26] achieves considerable performance improvements while keeping less computational and parameter overhead by generating channel attention and spatial attention sequentially. However, using convolution to generate local spatial attention limits the receptive field to the size of the convolution kernel. The global self-attention module is a component of the Transformer [25] that models long-range dependencies to generate the attention maps based on the input. Vision Transformer [6] applies Transformer architecture on many computer vision tasks and achieve impressive performance. The global self-attention operation is proper for modeling the spatial relationships of the feature map. Previous research works [5, 27, 28] apply attention mechanisms to the head pose estimation task and obtain promising performance. Therefore, we utilizes an attention module that can capture the features across the local and global dimensions.

3 Method

In this section, we will introduce a novel deep learning network with global-local attention mechanisms called GLPose in detail. The overall architecture is described in section 3.1. In section 3.2, the global-local attention mechanism in GLPose is introduced. The feature interpolation regularization module is described in section 3.3. The loss function used in GLPose is shown in section 3.4.

3.1 Overall Architecture

The model architecture with the siamese structure is shown in Figure 1. The proposed network GLPose is ResNet50 [11] integrating global-local attention mechanisms, the facial landmark detection module as an auxiliary task for predicting the 2D facial landmarks and the head pose prediction module as a main task for predicting head poses. In the head pose prediction module, the six-dimensional representation is first predicted, and this representation is then transformed to the rotation matrix. The detail of the transformation is shown in

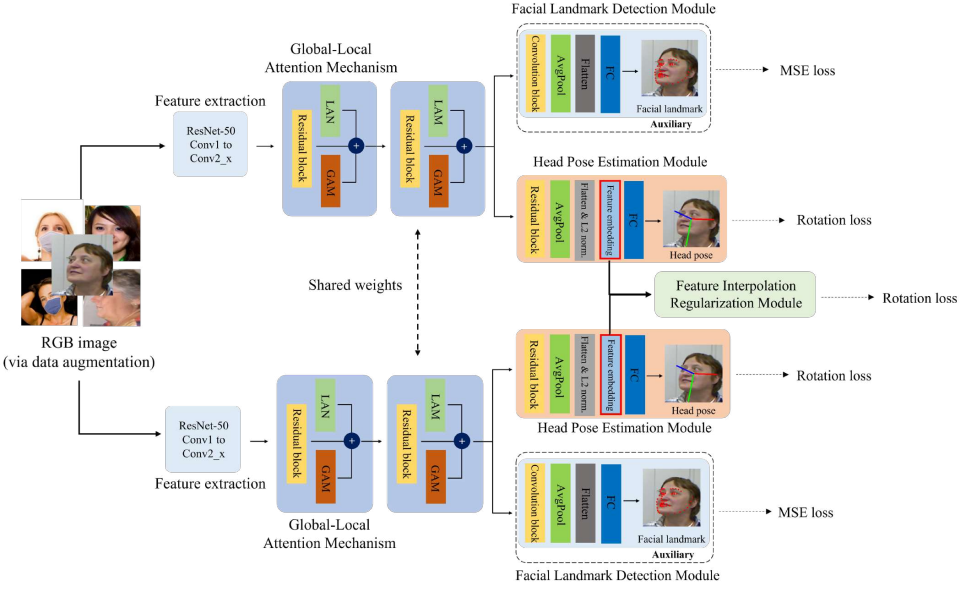

Figure 1: Illustration of the proposed network GLPose with the siamese structure. The network has five components: feature extraction module for low-level features, global-local attention mechanism for extracting useful information, feature interpolation regularization module for optimizing the feature embedding, facial landmark detection module for predicting the 2D facial landmarks, and head pose prediction module for predicting head pose.

the supplementary material. In the training stage, pairs RGB images with human faces are fed into two network branches, which share the network weights. Both generate respective prediction outputs. Furthermore, the feature embeddings from two branches are fed into the feature interpolation regularization module for feature embedding optimization. In the reference stage, the feature interpolation regularization module and the facial landmark detection module can easily remove and the network can perform the head pose estimation only.

3.2 Global-Local Attention Mechanism

Inspired by [24, 26] and due to the computational complexity of self-attention for large size of feature maps, we design global-local attention mechanisms which are integrated at the middle between two residual blocks as shown in Figure 1. The detail of the global-local attention mechanism is demonstrated in Figure 2. The global-local attention mechanism consists of two attention modules: local attention and global attention modules. The local attention enhances the channel information and spatial information sequentially, and the global attention models the spatial relationships of whole feature maps.

Local attention modules: Following CBAM [26], given the feature map $F \in \mathbb{R}^{C\times H\times W}$ from the backbone, where C indicates the number of channels, and the $H \times W$ is the spatial resolution, after adopting the channel attention module and the spatial attention module, the refined feature map $F^l \in \mathbb{R}^{C\times H\times W}$ is obtained.

Global attention modules: The global self-attention mechanism applies on the given feature map $F \in \mathbb{R}^{C\times H\times W}$ from the backbone, where C indicates the number of channels, and

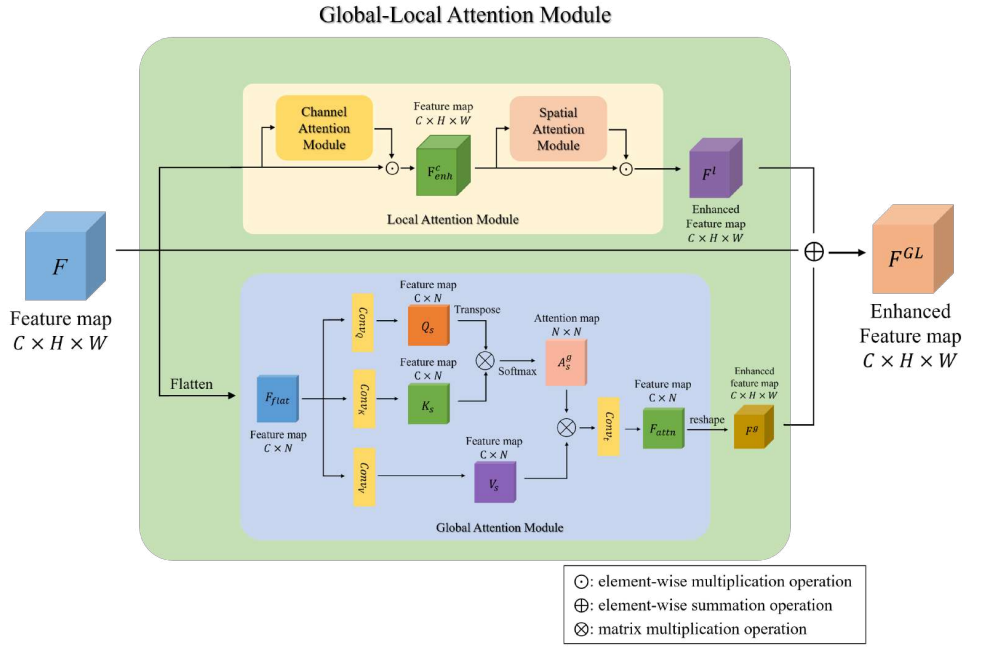

Figure 2: The structure of the global-local attention mechanism. There are two attention modules: local attention module and global attention module.

the $H \times W$ is the spatial resolution. After flatten along the spatial dimension and transformed by using three 1×1 convolutions, Q_s, K_s, and V_s tensors are obtained. The dimensions of them are $C \times N$, where N represents HW. We model the spatial relationships by matrix multiplication of Q_s and K_s followed by softmax operation over locations to obtain the global spatial attention map A_s^g:

$$A_s^g = softmax(Q_s^T K_s) \tag{1}$$

The global spatial attention map is multiplied with V_s. Finally, the refined feature map $F^g \in \mathbb{R}^{C \times H \times W}$ is obtained by using 1×1 convolution and the flatten operation.

After obtaining the enhanced feature maps F^l an F^g from local attention modules and global attention modules respectively, the final feature map F^{GL} is obtained through the input feature map F combined with F^l and F^g. The summation operation is adopted for the feature combination.

3.3 Feature Interpolation Regularization Module

In order to optimize the model capability to produce better feature embedding for predicting head poses, we propose a feature interpolation regularization module in our siamese structure during the training process. The two feature embeddings generated from head pose estimation module with the siamese structure become the inputs of the feature interpolation regularization module, as shown in Figure 1 highlighed by red rectangles. They are originally used to predict the head poses through the one FC layer. In the feature interpolation regularization module, two feature embeddings $\hat{f}_1$ and $\hat{f}_2$ are utilized to generate the one

feature embedding $\hat{f}_3$ first by:

$$\hat{f}_3 = (\hat{f}_1 + \hat{f}_2)/(||(\hat{f}_1 + \hat{f}_2)||_2) \tag{2}$$

where $\hat{f}_1$ and $\hat{f}_2$ represent the unit feature vectors after flatten and L2 normalization operations. The corresponding ground-truth rotation matrix label R_3 is generated by two ground-truth rotation matrices R_1 and R_2 of those feature embedding. Finally, we can use the generated rotation matrix label R_3 to supervise the rotation prediction generated from the feature embedding $\hat{f}_3$ through the shared FC layer in the head pose estimation module by rotation loss function. The detail of generating rotation matrix label is shown in the supplementary material.

3.4 Loss Function

For the prediction of rotation matrix, we adopt the geodesic distance loss, which is formulated as follows:

$$Geo(R_i^p, R_i^{gt}) = cos^{-1}(\frac{tr(R_i^p R_i^{gt}) - 1}{2}) \tag{3}$$

where $Geo(\cdot)$ is the geodesic distance loss. R_i^p and R_i^{gt} denote the prediction and ground truth of the rotation matrix from i^{th} image, respectively. $tr(\cdot)$ is the trace of the matrix $R_i^p R_i^{gt}$.

In order to learn a more abundantly representation, our model will predict the landmarks $P \in \mathbb{R}^{68\times 2}$ as an auxiliary task during training. The landmarks loss is defined as Eq. 4:

$$MSE(P_i^p, P_i^{gt}) = \sum_{k=1}^{L} (P_{ik}^{gt} - P_{ik}^p)^2 \tag{4}$$

where $MSE(\cdot)$ is the MSE loss calculate from all the landmarks. P_{ik}^p and P_{ik}^{gt} denote the k^{th} landmark of the prediction and ground truth from i^{th} image, respectively.

The overall loss of our model combines the geodesic distance loss and the landmark loss. By the feature interpolation regularization module we mentioned above, an additional geodesic distance loss will be considered. The overall loss function is formulated as follows:

$$\begin{aligned} L_{total} = & \frac{1}{N}\sum_{i=1}^{N}(Geo(R_i^p, R_i^{gt}) + \alpha \times MSE(P_i^p, P_i^{gt})) \\ & + \frac{1}{N}\sum_{j=1}^{N}(Geo(R_j^p, R_j^{gt}) + \alpha \times MSE(P_j^p, P_j^{gt})) \\ & + \frac{1}{N}\sum_{r=1}^{N}\beta \times Geo(R_r^p, R_r^{gt}) \end{aligned} \tag{5}$$

where N is equal to the batch size. i, j, and r denote the paired data from the siamese network and the data from the feature interpolating operation of the paired data, respectively. α is a hyper-parameter to balances the geodesic distance loss and landmark prediction loss. In this paper, α is empirically set to 0.1 and β is set to 0.5.

4 Experiments

This section illuminates the implementation details, the dataset and evaluation for training and testing, the comparison with state-of-the-art methods, and the ablation studies.

4.1 Implementation details

The model is end-to-end trained on a single RTX 3090 GPU with a mini-batch size of 64 for 30 epochs. Adam optimizer [16] is adopted with an initial learning rate starting from 0.0001. During the training period, the learning rate decay at epochs 10 and 20 by a factor of 5. The data augmentation is applied on the 224×224 input image with randomly scaling and flipping. Additionally, we utilize the synthetic facial mask operation with the tool of MaskTheFace[1] as a way of data augmentation. In each training iteration, we randomly sample 30% data to synthesize facial masks with different textures and colors.

4.2 Datasets and Evaluation

Dataset: 300W-LP [31], AFLW2000 [31] and BIWI [7] datasets are commonly used for the head pose estimation task. 300W-LP dataset is synthesised by using face profiling technique with 3D meshing to generate 61,225 images in total and further expands to 122,450 images with flipping. AFLW2000 dataset contains the first 2,000 images of the AFLW [18] dataset with large variations, various illumination and occlusion conditions. BIWI dataset was collected in a lab environment. It contains 24 videos of 20 subject about 15,000 images. However, those datasets have fewer cases with mask occlusion. MAFA [9] dataset provides the head pose classification label such as frontal, left side, right side labels under facial masks. After data cleaning in the MAFA testing dataset, there are about 6,000 face images for testing.

Evaluation: For training and evaluating the proposed network compared to other research, we follow the following two protocols for head pose estimation task. In Protocol 1 used in [23, 27], the network is trained on 300W-LP dataset, and AFLW2000 and BIWI datasets are used for testing. Mean absolute error (MAE) of the Euler angles is adopted as the evaluation metric. In Protocol 2, the network is trained on 300W-LP dataset and evaluated on MAFA dataset to validate the performance under facial mask situations. Similar to the evaluation setting in [19] which focuses on the two categories, the front face and side face, we further divide the side face into left and right sides. Specifically, we take $[-20^\circ, 20^\circ]$ of yaw angle as the range for front face category, and others for two other categories accordingly. The evaluation metric is the classification accuracy.

4.3 Competing Methods

We compare GLPose with the state-of-the-art methods in Table 1 on Protocol 1. Our proposed method outperforms the state-of-the-art works in each angle on the AFLW2000 dataset. Specifically, our method has obtained 3.35, 4.58 and 3.11 angle errors in yaw, pitch, roll, respectively. On the BIWI dataset, the performance is promising compared to the other works. Furthermore, the average errors over three major pose angles are the lowest in both datasets, as shown in the last column of each dataset in Tabel 1. In Protocol 2, we select two competitive works [12, 27] for comparison. There are two reasons why we take two head pose models: 1). they provide source codes so that we can implement their works for our purpose. 2). the one is widely used for analysis and the other is the latest work. As shown in Table 2, we can observe that our method has obtained promising performance compared to other competitive methods, which obtains 86.04% classification accuracy over three classes. Furthermore, the qualitative results on three datasets are shown in the supplementary material.

[1] https://github.com/aqeelanwar/MaskTheFace

Table 1: Comparison with state-of-the-art on BIWI and AFLW2000 for Protocol 1.

	BIWI				AFLW2000			
Method	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
Dlib (68 points) [15]	16.8	13.8	6.19	12.2	23.1	13.6	10.5	15.8
FAN (12 points) [1]	8.53	7.48	7.63	7.89	6.36	12.3	8.71	9.12
HopeNet (α=1) [23]	4.81	6.61	3.27	4.90	6.92	6.64	5.67	6.41
HopeNet (α=2) [23]	5.17	6.98	3.39	5.18	6.47	6.56	5.44	6.16
QuatNet [13]	**2.94**	5.49	4.01	4.15	3.97	5.62	3.92	4.50
FSA-Net [27]	4.27	4.96	2.76	4.00	4.50	6.08	4.64	5.07
WHENet-V [30]	3.60	4.10	2.73	3.48	4.44	5.75	4.31	4.83
WHENet [30]	3.99	4.39	3.06	3.81	5.11	6.24	4.92	5.42
FDN [28]	4.52	4.70	**2.56**	3.93	3.78	5.61	3.88	4.42
TriNet [2]	3.04	4.76	4.11	3.97	4.20	5.77	4.04	4.67
6DRepNet [12]	3.24	4.48	2.68	3.47	3.63	4.91	3.37	3.97
MFDNet [20]	3.40	4.68	2.77	3.62	4.30	5.16	3.69	4.38
Ours	4.18	**3.45**	2.67	**3.43**	**3.35**	**4.58**	**3.11**	**3.68**

Table 2: Comparison with competitive works on MAFA for Protocol 2.

Method	Accuracy (%)
FSA-Net [27]	64.16
6DRepNet [12]	81.52
Ours	**86.04**

4.4 Ablation studies

In this section, we conduct extensive experiments to verify and understand the effectiveness of design components. It is noted that the network with the siamese structure will only be utilized when the feature interpolation regularization module is used. As shown in Table 3, we first compare each individual component. The feature interpolation regularization module shows the better performance because it focuses on optimizing the embedded features. Combining the global-local attention mechanism with multi-task learning is more useful to learn the facial information. Moreover, the final performance is the best when all the modules are fully utilized.

Table 3: Ablation analysis on different modules. GLAM, MTL and FIRM represent global-local attention mechanism, multi-task learning and feature interpolation regularization module respectively.

			BIWI				AFLW2000			
GLAM	MTL	FIRM	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
			4.34	4.06	2.77	3.72	3.76	4.98	3.23	3.99
✓			4.17	3.64	2.77	3.52	3.63	4.87	3.22	3.90
	✓		4.19	3.63	2.74	3.52	3.75	4.95	3.24	3.98
		✓	4.25	3.53	2.74	3.50	3.41	4.83	3.21	3.81
✓	✓		**4.05**	3.60	**2.60**	**3.41**	3.50	4.86	3.23	3.86
✓	✓	✓	4.18	**3.45**	2.67	3.43	**3.35**	**4.58**	**3.11**	**3.68**

5 Conclusion

In this paper, we propose a novel network called GLPose for head pose estimation. In the proposed GLPose, the global-local attention mechanism integrated into the backbone network extracts local and global information, while the feature interpolation regularization module optimizes the network modeling ability to produce better feature embeddings and the facial landmark detection module learns the additional information. Furthermore, to address the performance degradation of the previous head pose estimation models under the situation where humans wear facial masks and problem of lacking enough data with mask occlusion in existing datasets, we utilize the synthetic facial mask operation as a way of data augmentation for model training. Extensive experiments validate the effectiveness of our proposed model, which shows competitive performance compared to other existing methods, even when the faces are occluded by facial masks.

6 Acknowledgement

This research was supported by the Ministry of Science and Technology of Taiwan, and Center for Artificial Intelligence & Advanced Robotics, National Taiwan University, under the grant numbers MOST 108-2221-E-390-019-MY3, MOST 110-2634-F-002-049 and MOST 110-2221-E-002-166-MY3.

References

[1] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1021–1030, 2017.

[2] Zhiwen Cao, Zongcheng Chu, Dongfang Liu, and Yingjie Chen. A vector-based representation to enhance head pose estimation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1187–1196, 2021.

[3] Donggen Dai, Wangkit Wong, and Zhuojun Chen. Rankpose: Learning generalised feature with rank supervision for head pose estimation. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[4] 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, 2021.

[5] Naina Dhingra. Lwposr: Lightweight efficient fine grained head pose estimation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1204–1214, 2022.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning*

Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[7] 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, 2012.

[8] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981. ISSN 0001-0782.

[9] Shiming Ge, Jia Li, Qiting Ye, and Zhao Luo. Detecting masked faces in the wild with lle-cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2682–2690, 2017.

[10] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. *ArXiv*, abs/2202.12555, 2022.

[13] 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, 2019.

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

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

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] 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, 2017.

[18] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2144–2151, 2011.

[19] Shuang Li, Xin Ning, Lina Yu, Liping Zhang, Xiaoli Dong, Yuan Shi, and Wei He. Multi-angle head pose classification when wearing the mask for face recognition under the covid-19 coronavirus epidemic. In *HPBD&IS*, pages 1–5, 2020.

[20] 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, 2022.

[21] Shentong Mo and Xin Miao. Osgg-net: One-step graph generation network for unbiased head pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 2465–2473, New York, NY, USA, 2021. Association for Computing Machinery.

[22] Prajval Kumar Murali, Mohsen Kaboli, and Ravinder Dahiya. Intelligent in-vehicle interaction technologies. *Advanced Intelligent Systems*, 4(2):2100122, 2022.

[23] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2155–215509, 2018.

[24] Chull Hwan Song, Hye Joo Han, and Yannis Avrithis. All the attention you need: Global-local, spatial-channel attention for image retrieval. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 439–448, 2022.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[26] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 3–19, Cham, 2018. Springer International Publishing.

[27] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1087–1096, 2019.

[28] Hao Zhang, Mengmeng Wang, Yong Liu, and Yi Yuan. Fdn: Feature decoupling network for head pose estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12789–12796, Apr. 2020.

[29] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746, 2019.

[30] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

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