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

Hsueh−Wei Chen[1], Yi Chen[1], Pei−Yung Hsiao[2], Li−Chen Fu[1] , ZI−RONG DING[3]

[1]Department of Computer Science and Information Engineering National Taiwan University
[2]Department of Electrical Engineering National University of Kaohsiung
[3]Automotive Research & Testing Center, Taiwan

BMVC 2022

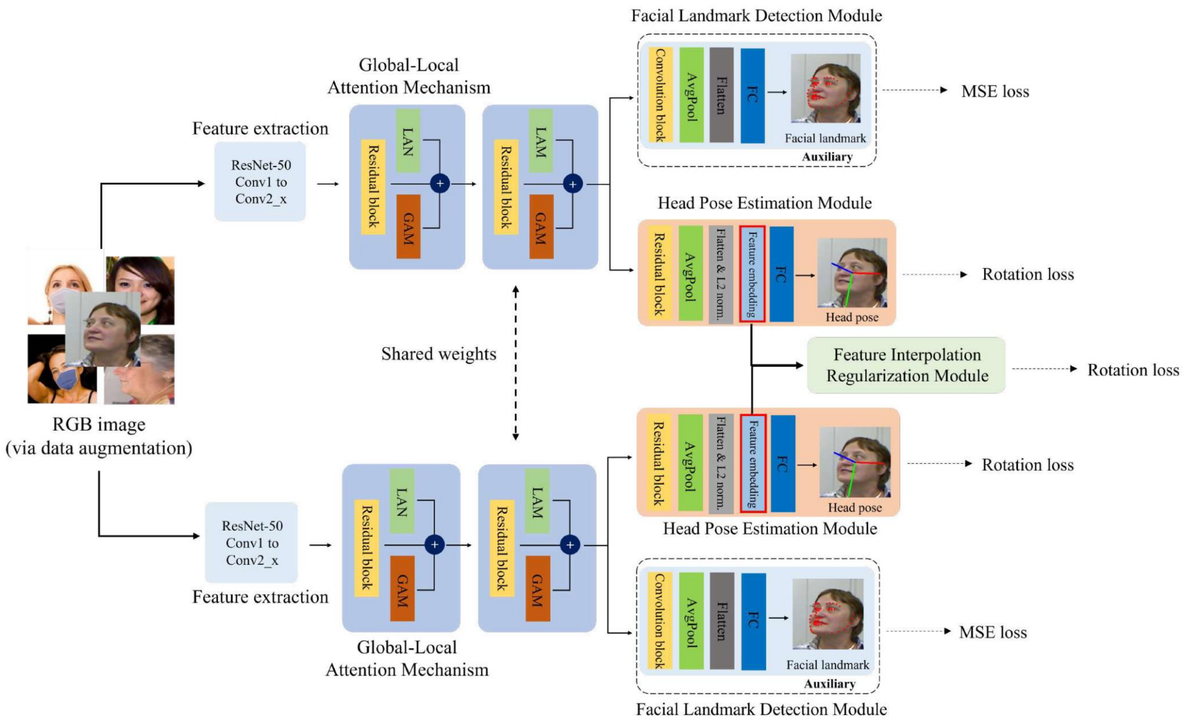
Motivation and Objective

- The public datasets for head pose estimation have fewer cases of facial mask occlusion.
  - Generate the synthesized facial masks on original datasets as a way of the data augmentation.
- Hardly to discriminate the valuable information from feature maps, especially under facial mask situations.
  - Develop a deep learning model with attention mechanisms and auxiliary task supervision to make features more discriminative to estimate head pose.
- The performance for head pose estimation might degrade when the head orientations are not seen before.
  - Design the regularization module to optimize the feature space for head pose estimation to increase the robustness.

Methodology

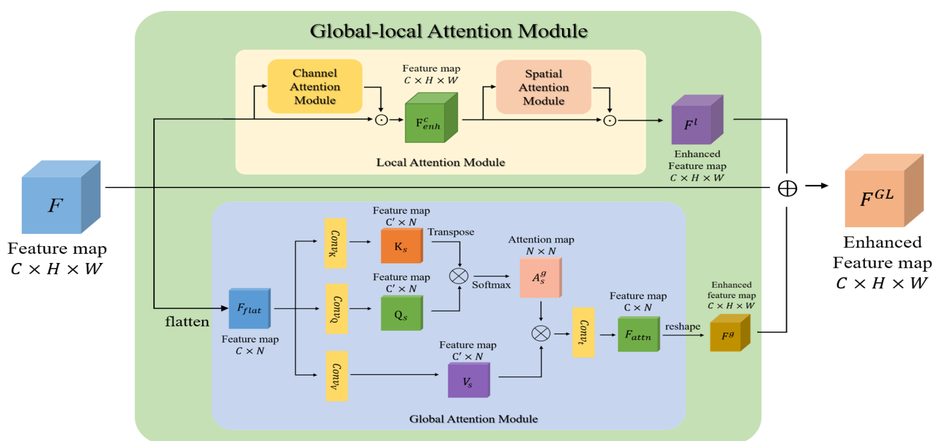
Architecture Overview

- The proposed GLPose integrates the global-local attention mechanisms in the backbone network. 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. The feature embeddings from two HPE branches are fed into the feature interpolation regularization module (FIRM).
- The FIRM and the facial landmark detection module can be removed in the reference stage and thus will not affect the inference time.



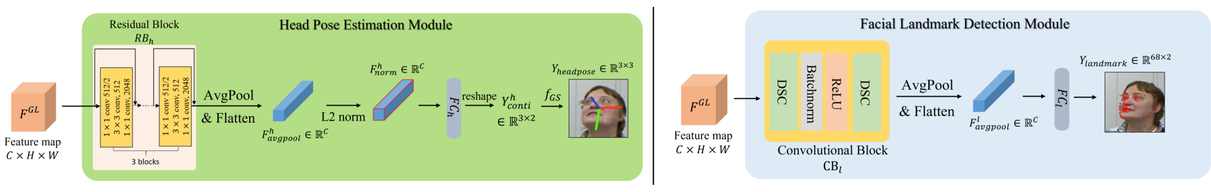
Global-Local Attention Module

- To discriminate the valuable features effectively, we design global-local attention module, which consists of the global attention module and local attention module.
- We apply CBAM[1] as our local attention module, which introduce channel attention and spatial attention into the model. And inspired by the Vision Transformer[2], self-attention is applied as our global attention module to extract the features of larger areas over the feature maps to obtain global information.



Facial Landmark Detection & Head Pose Estimation Module

- To help the model learn additional and discriminative features, we design a landmark detection branch as the auxiliary task supervision. The output of this module is landmark locations $Y_{landmark} \in \mathbb{R}^{68\times2}$, where DSC refer to depthwise separable convolution. The HPE branch predicts the rotation matrix $Y_{headpose} \in \mathbb{R}^{3\times3}$, which is transformed from Y^h_{conti} using the transformation function f_{GS}.



Feature Interpolation Regularization Module

- In order to optimize the feature embedding for the head pose estimation model for further improvement, we design the feature interpolation regularization module during training.
  1. Take $F^h_{norm1}, F^h_{norm2} \in \mathbb{R}^{C\times1\times1}$ from two HPE branches as input
  2. $F^h_3 = (F^h_{norm1} + F^h_{norm2})/2$
  3. $F^h_{norm3} = F^h_3 / ||F^h_3||_2$
  4. $Y_{headpose3} = f_{Gs}(FC_h(F^h_{norm3}))$

Loss Function

- Geodesic distance loss: $Geo(\mathrm{R}^p_i, \mathrm{R}^g_i) = \cos^{-1}\left(\frac{tr(\mathrm{R}^p_i \mathrm{R}^g_i)-1}{2}\right)$
- Facial landmark prediction loss: $MSE(P^p_i, P^g_i) = \sum_{k=1}^{L} \left|\left|p^p_{ik} - p^g_{ik}\right|\right|_2$

For i^{th} input image, where
R^p_i: the model prediction of the rotation matrix
R^p_i: the ground truth of the rotation matrix
$tr(\cdot)$: the trace of the rotation matrix

For i^{th} input image, where
P^p_i: the model prediction of landmark locations
P^g_i: the ground truth of landmark locations
L : Number of landmarks

- Overall loss: $L_{total} = \frac{1}{N}\left(\sum_{i=1}^{N}\left[Geo(\mathrm{R}^{p1}_i, \mathrm{R}^{g1}_i) + \alpha \cdot MSE(P^{p1}_i, P^{g1}_i)\right]\right) + \frac{1}{N}\left(\sum_{j=1}^{N}\left[Geo\left(\mathrm{R}^{p2}_j, \mathrm{R}^{g2}_j\right) + \alpha \cdot MSE\left(P^{p2}_j, P^{g2}_j\right)\right]\right) + \frac{1}{N}\sum_{k=1}^{N}\beta \cdot Geo(R^{p3}_k, R^{g3}_k)$

Experiments

Comparisons on AFLW2000 and BIWI datasets

Method	AFLW2000				BIWI			
	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
HopeNet[1]	6.47	6.56	5.44	6.16	5.17	6.98	3.39	5.18
FSA-Net[2]	4.50	6.08	4.64	5.07	4.27	4.96	2.76	4.00
TriNet[3]	4.20	5.77	4.04	4.67	3.04	4.76	4.11	3.97
OsGG-Net[4]	3.96	5.71	3.51	4.39	**3.26**	4.85	3.38	3.83
WHENet[5]	5.11	6.24	4.92	5.42	3.99	4.39	3.06	3.81
6DRepNet[6]	3.63	4.91	3.37	3.97	**3.24**	4.48	**2.68**	**3.47**
Our	**3.33**	**4.71**	**3.16**	**3.73**	4.23	**3.54**	2.78	3.51
Our*	**3.35**	**4.58**	**3.11**	**3.68**	4.18	**3.45**	**2.67**	**3.43**

* indicates using the synthesis data during training

Ablation Study

Module			BIWI				AFLW2000			
GLAM	Multi-Task	Feature Interpolation	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
		√	4.24	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**

[1] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, “CBAM: Convolutional block attention module,” *ECCV*, Sep. 2018, Munich, Germany.
[2] 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,” *ICLR*, 2021.
[3] N. Ruiz, E. Chong and J. M. Rehg, "Fine-Grained Head Pose Estimation Without Keypoints," *CVPRW*, 2018.
[4] T. Yang, Y. Chen, Y. Lin and Y. Chuang, "FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation From a Single Image," *CVPR*, 2019.
[5] Zhiwen Cao, Zongcheng Chu, Dongfang Liu, and Yingjie Chen. A vector-based representation to enhance head pose estimation. *WACV*, 2021.
[6] Mo, Shentong, and Xin Miao. "OsGG-Net: One-step Graph Generation Network for Unbiased Head Pose Estimation." *ICM*, 2021.
[7] Yijun Zhou and James Gregson. Whenet: Real-time finegrained estimation for wide range head pose. arXiv preprint arXiv:2005.10353, 2020.
[8] Thorsten Hempel and Ahmed A. Abdelrahman and Ayoub Al-Hamadi, "6D Rotation Representation for Unconstrained Head Pose Estimation", ICIP 2022.