Perceive Emotions from Gazes

Qinyue Zheng (325450), Shijian Xu (327448), Sruti Bhattacharjee (346910)
CS-503 Final Project Report

***Abstract*—Facial expressions convey a lot of valuable information for emotion recognition, which can be utilized in many areas, like human-machine interaction. In the post-epidemic era, it is difficult to recognize emotions when facial expressions are usually covered by masks. However, psychological experiments suggest that gaze direction may also play an important role in emotion perception. In this project, we aim to explore the relationship between gaze direction and perceived emotions. We first train a gaze estimation model on ETH-XGaze, and then apply this model to the RAF-DB dataset to generate the pseudo-gaze labels. After that, we analyze the correlations between gazes and emotions from many different perspectives. The results suggest that the correlations between all the emotions and gaze orientations are not that strong. But if we only consider the general types of emotions and gazes, there seems to be some patterns between them.**

I. Introduction

Facial expression conveys valuable information about people's emotions and it plays an important role in daily social interactions. Recognizing emotion from facial expression not only facilitates social communications but also has the potential to help computers and machines understand human behavior and interact with them. However, in the post-pandemic era, the whole world is now covering faces with masks, which gives rise to challenges for facial expression recognition and emotion perception. As it been said, finding a more efficient way to perceive and recognize the emotions from limited facial area becomes more and more important under these circumstances.

Psychological studies show that gaze direction and facial expression are combined in the processing of emotionally relevant facial information. More importantly, a particular direction of gaze can convey multiple social meanings. According to the Shared Signal Hypothesis (SSH) [1], for different facial expressions, gaze direction changes which reflect an individual's approach and intent. Emotion perception in general conveys multiple social meanings and affects communication. In particular, evidence shows that signals of approach or avoidance [2], perception of sadness [3] are closely associated with gaze directions. Now we see that gaze direction can be a critical clue that influences emotion recognition [3] [4]. The question remained is that whether the gaze direction can be an independent cue to recognizing emotions. If so, can we train machines to perceive emotions from gaze directions automatically?

In this project, we will explore the problem stated above in the following ways. Firstly, we start with training a good gaze estimation model. We utilize the transformer model and train it on a large scale gaze estimation dataset, the ETH-XGaze [5]. Then we apply it to another facial expression recognition dataset, the RAF-DB [6], which contains the labels of 7 basic emotions, to generate the pseudo-gaze labels. Finally, we analyze the correlations between the pseudo-gaze labels and emotion labels.

II. Related Work

Gaze Estimation. Gaze estimation is a regression process to predict pitch-yaw gaze direction vectors from the face images. We have found plenty of works on model performance improvement for appearance-based gaze estimation using deep learning techniques [7], [8], [9], [10]. Among all the available datasets, we found ETH-XGaze [5], Gaze360 [11] and MPIIGaze [12] are most relevant for our project.

However, gaze estimation is not only an interesting problem in engineering (human computer interaction and robotics) but also an interesting cognitive problem. Some psychological researches suggest that there exist some associations between the perception of emotion and gaze, but the exact mechanism underlying the influence of gaze direction on facial expression perception is still unclear [3].

Facial Expression Recognition. In computer vision and deep learning research, facial expression recognition has become a popular topic in the last few years. Numerous works have been done in this realm already. Facial expression is one of the most powerful signals that allow people to convey their emotional states and intentions. Automatically and successfully recognizing the emotions is important for human-computer interaction, as well as for human-to-human communication. Many works have delved into this area and achieve great performance [13], [14], [15].

However, under the COVID19 epidemic, effective human communications are obstructed by the masks. Without facial expressions, it is difficult to perceive the emotions. The good thing is, gaze also conveys a lot of information about the perception of emotions and we might be able to utilize the gaze estimation to detect and recognize emotions.

Gaze and Emotion Recognition. In psychology, for a long time, many theorists argued that each basic emotion is associated with prototypical facial expressions [16], and in this framework, gaze directions or gaze shifts are not considered as a differentiating feature. However, some theories based on appraisal processes argued that gaze direction represents a critical cue for emotion recognition [4].

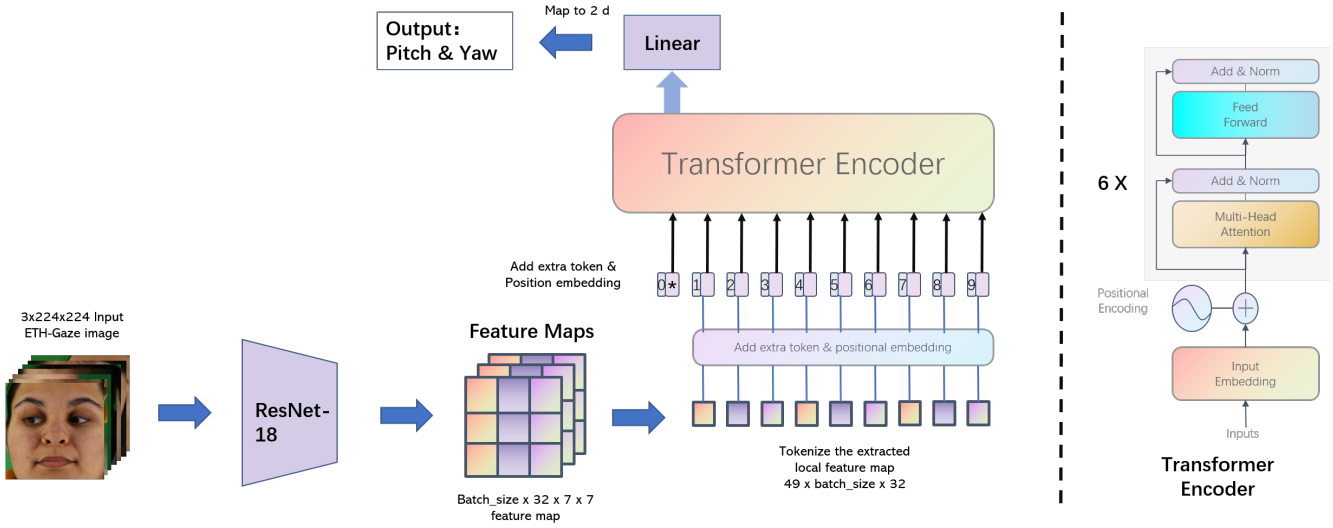

Figure 1: Gaze detection model

Studies say that gaze direction is potentially related to an individual's behavioral intent (approach-avoidance) communicated by an emotional expression, and the perception of that emotion is enhanced or facilitated. Milders et al. [17] found that fearful faces were detected more frequently with averted gaze than with direct gaze, whereas happy and angry faces were detected more frequently with direct gaze.

In this project, we aim to explore the relationship between gaze and emotion perception. More specifically, we want to verify whether human or machine can recognize different emotions according to gaze estimation.

III. Method

This section describes the models and analysis methods we propose. We start with a brief description of the transformer architecture and attention mechanism and then proceed with presenting the analysis pipelines for the correlations between gazes and emotions.

A. Transformer for Gaze Estimation

We utilize the transformer model for gaze estimation. The core of the transformer is the self-attention mechanism. In details, given a feature matrix $\mathbf{X} \in R^{n \times d}$, the feature is projected into queries $\mathbf{Q} \in R^{n \times d_k}$, keys $\mathbf{K} \in R^{n \times d_k}$ and values $\mathbf{V} \in R^{n \times d_v}$ with multi-layer perceptrons (MLP), where n is the batch size, d, d_k and d_v is the dimension of each feature. The self-attention is computed as:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \quad (1)$$

In our project, we use a hybrid model, which consists of a ResNet-18 as spacial feature extractor, and a transformer. More specifically, as in most of the computer vision tasks, only the encoder part of the transformer is used.

B. Implementation Details

We use $3 \times 224 \times 224$ face images for gaze direction estimation. The estimated gaze is a 2D pitch-yaw vector. We employ L1-loss and Adam for training and optimization.

As shown in Figure 1, in first step of feature extraction, we use ResNet-18 get $32 \times 7 \times 7$ feature maps. The feature maps are further flatten and tokenized, and then fed into a stack of 6-layer transformer encoder. The number of heads of attention-mechanism is set as 8. At the time of positional encoding, we add an extra learnable token, whose state at the output of the transformer encoder (pos 0) serves for further regression as the gaze representation. The output of this token is then transformed into a fully connected layer and mapped into 2d.

C. Transfer to Emotion Dataset and Correlation Analysis

We first train the model on the gaze estimation dataset ETH-XGaze. After that, we directly apply this pre-trained model to the facial expression recognition dataset, which in our case is RAF-DB. This dataset contains the labels for 7 basic emotions but does not have any gaze labels. We use the pre-trained model to predict/generate the pseudo-gaze labels. Based on the assumption that the gaze estimation is good enough on this dataset, we then analyze the correlations between the pseudo-gazes and the ground-truth emotions.

IV. Experiments

This section describes our experimental setup including pre-training methods, datasets and transferring results. All these experiments were conducted using PyTorch [18].

A. Setup

Dataset for pre-training. We use ETH-XGaze [19] for pre-training. ETH-XGaze consists of over one million high-resolution images of varying gaze under extreme head poses.

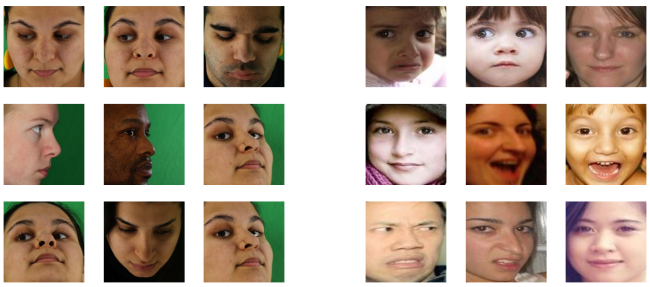

Figure 2: Sample images from ETH-XGaze dataset(left) and RAF-DB dataset(right).

It was collected from 110 participants with a custom hardware setup.

Dataset for emotion recognition. We use the Real-world Affective Faces Database (RAF-DB) for facial expression recognition. This is a large-scale facial expression database with around 30K great-diverse facial images. It covers 7 basic human emotions, which are surprise, fear, disgust, happiness, sadness, anger and neutral. We apply the pre-trained gaze estimation model to this dataset to generate the pseudo-gaze labels for the images. After that, we investigate the relationships between the predicted gazes and the ground-truth emotion labels. Fig. 2 shows some sample images from ETH-XGaze and RAF-DB.

Training. We train the model on ETH-XGaze with 512 batch size and 100 epochs. The learning rate is set as 0.0005. We use Adam optimizer to train the model with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We also use the linear learning rate warmup strategy, which is set as 5 epochs. After that, a standard StepLR learing rate schdeduler is usd, with step_size = 60 and $\gamma = 0.5$. The model is implemented using PyTorch and trained on a single NVIDIA V100 GPU.

B. Gaze Estimation Results

We directly apply the pre-trained gaze estimation model to RAF-DB dataset. Due to the lack of ground-truth gaze labels for this dataset, we only show some visual qualitative results in Fig. 3.

While we can not quantitatively measure the gaze prediction performance, the visual checks suggest that the gaze predictions are not too bad. Actually, most of the images show visually consistent gaze predictions. The following evaluations and analyses are based on the assumption that the gaze predictions on RAF-DB are reliable.

C. Correlation Analysis for Gazes and Emotions

Dataset observation. To analyze the relationships between gazes and emotions, we first do some dataset observations to the RAF-DB dataset. Fig. 4 shows different emotion distributions on this dataset. From this figure, we can see that the emotion distributions are not balanced and happiness has the most samples while fear has the least samples.

Figure 3: Sample visual results of the gaze prediction on RAF-DB.

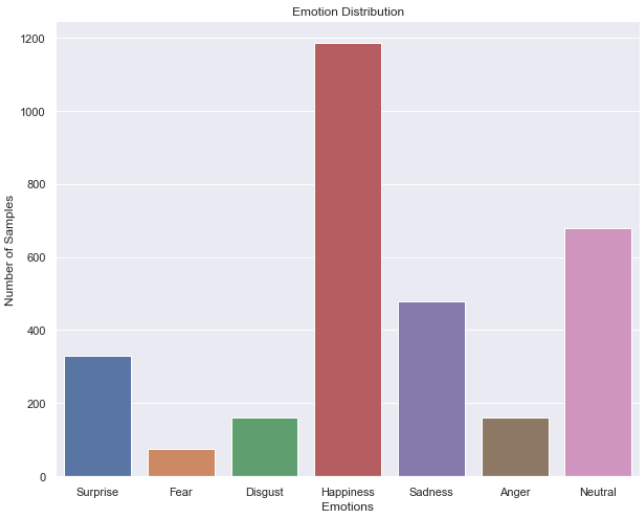

Figure 4: Emotion distributions on RAF-DB.

Relationship between 2d gazes and emotions. To see if there are any potential relationships between the pseudo-gazes (we call it gazes in the following), we draw the scatter plots for both 2d gazes (pitch and yaw) and 3d gazes (x, y, z) in Fig. 5 and Fig. 6, where the number 1-7 represent the 7 basic emotions (surprise, fear, disgust, happiness, sadness, anger and neutral) respectively. Unfortunately, from these two figures, it seems there are no obvious patterns/correlations between the 2d/3d gazes and 7 emotions.

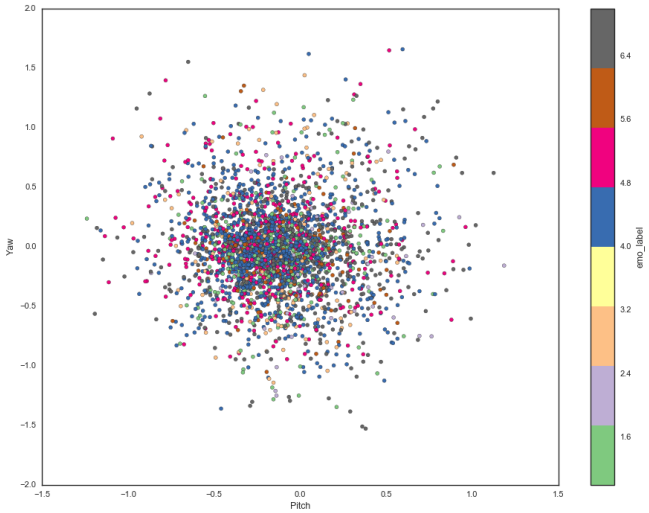

Figure 5: Scatter plot for 2d gazes.

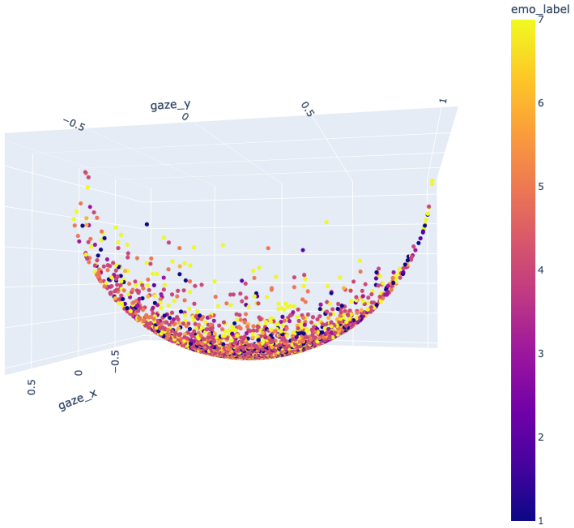

Figure 6: Scatter plot for 3d gazes.

Direct / averted gazes, approach / avoidance-oriented emotions. It seems that the direct relationship analysis between the raw gazes and the 7 basic emotions are not very informative and no obvious patterns can be observed. The shared signal hypothesis (SSH) [2] suggests that when gaze direction is combined with the intention conveyed by a particular expression, it enhances the perception of that emotion. Expressions of happiness and anger are classified as "approach-oriented emotions" and are therefore usually accompanied with direct gaze. Fear and sadness expressions are categorized as "avoidance-oriented emotions" and are

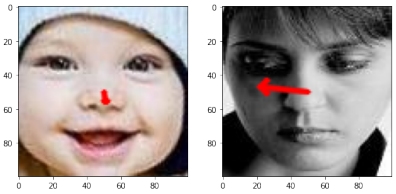

Figure 7: Sample images of direct gaze and averted gaze. The left image shows the happy emotion, accompanied with direct gaze. The right image shows the sadness emotion, accompanied with averted gaze.

therefore more likely to be categorized when accompanied with averted gaze. Based on this observation, we turn to investigate the relationships between the coarse classes of gazes and coarse classes of emotions. Given an image with gaze estimation, we compute the length of the projected gazes in the image and consider the gazes shorter than a threshold as direct gazes, and the others as averted gazes. More specifically, for an image with 2d gaze value $(pitch, yaw)$, we compute the projected dx and dy of the gaze on the image: $dx = -length * sin(yaw), dy = -length * sin(pitch)$, where $length$ is the default gaze length (we directly set is to be 40) without any distortion. Then, the length of the projected gaze is computed as $\sqrt{dx^2 + dy^2}$. The threshold is empirically set as $10\sqrt{5}$. For emotion, we simply classify fear, sadness and disgust as the avoidance-oriented emotions and the other four as approach-oriented emotions. Fig. 7 shows the examples of direct gaze and averted gaze, which are related to happy and sadness respectively.

Select Face-Forward RAF images. RAF-DB contains many images which have faces turning to other directions. While gaze estimation can predict reasonable gaze directions for these faces, but they are not suitable for our analysis due to the difficulty in distinguishing the direct or averted gazes. To alleviate this problem, we do some pre-processing to the RAF-DB dataset and select the face-forward images. To do so, we compute the distance between the central of two eyes and the tips of the nose, based on the landmark location labels provided in the dataset. If the distances of the two eyes to the nose differ too much, then we discard this images. Fig. 8 illustrate the two cases.

Correlation Analysis. After data pre-processing, we compute the correlations between the original 7 emotions/2 emotion types and the 2 gaze types. The results are shown in Table I. We first compute the correlations on the original RAF-DB dataset, without removing the non-face-forward images. From the table we can see that both the correlation between the raw emotions and gaze types and the correlation between the emotion types and gaze types are very low. After selecting the face-forward images, the both correlations increase, but still quite low. We hypothesize that too many

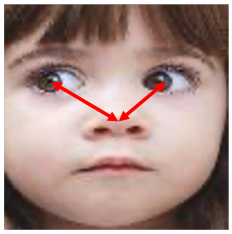

Figure 8: Select face-forward RAF-DB images. The left image is considered to be face-forward while the right image will be discarded.

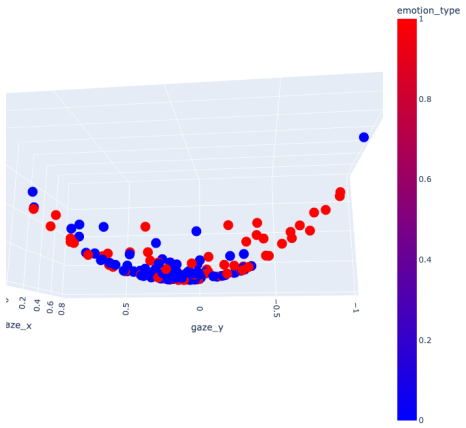

Figure 9: The scatter plot for the 3d gazes of anger and disgust emotions. Red dots represent anger while blue dots represent disgust.

emotions compounded together maybe hinder the correlation analysis. So we turn to examine the patterns of two specific emotions. Specifically, we choose anger and disgust as the representatives of the approach-oriented emotions and the avoidance-oriented emotions, and then compute their correlations between the two types of gazes. It turns out that there is a relatively high correlation between them. A visualized plot is shown in Fig. 9. We can see there is a not so vague pattern between the gazes and emotions.

V. Conclusion and Limitations

The experiments and analysis results suggests that gaze direction is an important cue in the perceptual processing of facial displays of emotion, which has not been previously demonstrated in the research literature. They also indicate that the effects of gaze direction on the processing of emotions depend on the specific type of emotion displayed by the face, i.e., some types of approach-oriented emotions are more closely related to the direct gazes and some types of avoidance-oriented emotions are more closely related to the averted gazes (like anger and disgust).

Data	Description	Correlation
Row RAF Test Data	Raw Emotions vs. Gaze Types	0.0116
	Emotion Types vs. Gaze Types	0.1039
Filtered RAF Data	Row Emotions vs. Gaze Types	0.0207
	Emotion Types vs. Gaze Types	0.1321
Fine-grained Emotion (Anger & Disgust)	Emotion Types vs. Gaze Types	**0.3399**
Fine-grained Emotion (Happy & Sad)	Emotion Types vs. Gaze Types	0.1118

Table I: Correlation analysis. Row emotions means the 7 basic emotions. Emotion types means the approach-oriented emotions and the avoidance-oriented emotions. Gaze types means the direct gazes and the averted gazes.

However, there are also some limitations of our project. First, we do not have quantitative results for the gaze and emotion predictions on the RAF-DB due to the lack of gaze labels on this dataset. Second, there is a domain gap between ETH-XGaze and RAF-DB. In the future work, we need to minimize the domain gap between the pretrained gaze dataset and the emotion dataset. Or we can try to create a new dataset that contains both gaze and emotion labels.

References

[1] J. Liang, Y.-Q. Zou, S.-Y. Liang, Y.-W. Wu, and W.-J. Yan, "Emotional gaze: The effects of gaze direction on the perception of facial emotions," *Frontiers in Psychology*, vol. 12, p. 2796, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2021.684357

[2] R. B. Adams Jr and R. E. Kleck, "Perceived gaze direction and the processing of facial displays of emotion," *Psychological science*, vol. 14, no. 6, pp. 644–647, 2003.

[3] O. Semyonov, A. Ziv-El, E. G. Krumhuber, S. Karasik, and H. Aviezer, "Beyond shared signals: The role of downward gaze in the stereotypical representation of sad facial expressions." *Emotion*, 2019.

[4] K. N'diaye, D. Sander, and P. Vuilleumier, "Self-relevance processing in the human amygdala: gaze direction, facial expression, and emotion intensity." *Emotion*, vol. 9, no. 6, p. 798, 2009.

[5] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation," in *European Conference on Computer Vision*. Springer, 2020, pp. 365–381.

[6] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2584–2593.

[7] Y. Liu, R. Liu, H. Wang, and F. Lu, "Generalizing gaze estimation with outlier-guided collaborative adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3835–3844.

[8] X. Cai, B. Chen, J. Zeng, J. Zhang, Y. Sun, X. Wang, Z. Ji, X. Liu, X. Chen, and S. Shan, "Gaze estimation with an ensemble of four architectures," *arXiv preprint arXiv:2107.01980*, 2021.

[9] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 623–10 630.

[10] Z. Guo, Z. Yuan, C. Zhang, W. Chi, Y. Ling, and S. Zhang, "Domain adaptation gaze estimation by embedding with prediction consistency," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[11] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, , and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[12] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4511–4520.

[13] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6897–6906.

[14] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5683–5692.

[15] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2402–2411.

[16] P. Ekman and W. V. Friesen, *Facial action coding system: Investigator's guide*. Consulting Psychologists Press, 1978.

[17] M. Milders, J. K. Hietanen, J. M. Leppänen, and M. Braun, "Detection of emotional faces is modulated by the direction of eye gaze." *Emotion*, vol. 11, no. 6, p. 1456, 2011.

[18] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[19] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation," in *European Conference on Computer Vision (ECCV)*, 2020.