Project Report: Quality Adaptive Margin for POSTER backbone in Face Recognition

Thi-Trang Nguyen
HICR Lab, Computer science , SeoulTech
trangnguyen.hust117@gmail.com

Hamza
HICR Lab, Computer science , SeoulTech
hamza.seoultech@gmail.com

Abstract

Facial Expression Recognition (FER) has received increasing interest in the computer vision community. As a challenging task, there are three key issues especially prevalent in FER: inter-class similarity, intra-class discrepancy, and scale sensitivity. Pyramid Cross-Fusion TransformER network (POSTER) [Zheng et al.2022] is proposed that aims to holistically solve these issues. Moreover, advances in margin-based loss functions have resulted in enhanced discriminability of faces in the embedding space. Previous studies have studied the effect of adaptive losses to assign more importance to misclassified (hard) examples. [Kim et al.2023] introduced another aspect of adaptiveness in the loss function - AdaFace, namely the image quality. They argue that the strategy to emphasize misclassified samples should be adjusted according to their image quality. In this project, aiming to understanding deeper in FER, we explore the POSTER model which has experimental results outperforms SOTA methods on RAF-DB. Moreover, we adapt Quality Adaptive Margin into POSTER and investigate the effect of this approach on backbone POSTER for FER.

1 Introduction

Facial expression recognition functions as a universal language, allowing us to express a wide range of emotions without saying a single word. These small variations in our facial characteristics, from the happy brilliance of a grin to the furrowed brow of concern, allow us to connect, sympathize, and comprehend one another on a profound level. [Ekman and Rosenberg2005]

Recent advances in computer vision, machine learning, and artificial intelligence have resulted in tremendous development in the field of facial expression detection, predicting a future in which robots can interpret and respond to human emotions [Li and Deng2022]

The study and interpretation of facial muscle movements known as face action units (AUs) is part of facial expression identification. We get vital insights into an individual's emotional state and intentions by integrating these AUs with contextual indicators such as head position, eye gazing, and voice [Adyapady and Basava2022]. However, automated recognition of facial expressions poses its own set of obstacles, such as variations in illumination, occlusions, position discrepancies, and individual differences. To overcome these obstacles, advanced algorithms, robust feature representations, and large datasets must be developed. The primary purpose of this study is to make a contribution to the field of facial expression recognition by creating a novel system capable of detecting and classifying facial expressions.

In summary, the purpose of this work is to make important advances to the field of facial expression identification by developing a robust and accurate system capable of decoding the secret language of human emotions contained within facial expressions. We combine the two architectures of POSTER [Zheng et al.2022] and AdaFace [Kim et al.2023] where POSTER act as a backbone for the

AdaFace: Quality Adaptive Margin for Face Recognition where the loss function to emphasize the samples according to the image's quality.

2 Related Work

Deep learning in FER: With the rapid progress of deep learning techniques in computer vision tasks, deep learning solutions has increasingly been implemented to handle challenging FER task and achieving promising performance such as a Region Attention Network (RAN) [Wang et al.2019a], Deep Attentive Center Loss (DACL) [Farzaneh and Qi2021], a Self-Cure Network (SCN) [Wang et al.2020], and so on

Facial landmarks in FER: Facial landmark detection aims to estimate the location of predefined keypoints on the human face. The detected facial landmark are used in many face analysis tasks such as face recognition [Taigman et al.2014,Liu et al.2018] face tracking [Khan et al.2017], and emotion recognition [Jung et al.2015, Hasani and Mahoor2017, Ruan et al.2021]. Existing methods that utilize facial landmarks ignore the correlations of landmark features and image features. Among SOTA methods [She et al.2021, Shi et al.2021, Xue et al.2021] in the FER task, none of them use facial landmarks

Vision transformer: The breakthrough of transformer networks in Natural Language Processing (NLP) has sparked great interest in the computer vision domain. [Zheng et al.2022] employs a two stream pyramid cross-fusion transformer network POSTER to explore the correlation of image features and landmark features to tackle inter-class similarity, intra-class discrepancy, and scale sensitivity issues in FER.

Margin Based Loss Function. The margin based softmax loss function is widely used for training face recognition (FR) models [Deng et al.2022, Huang et al.2020, Liu et al.2018, Wang et al.2018]. Margin is added to the softmax loss because without the margin, learned features are not sufficiently discriminative. SphereFace [Liu et al.2018], CosFace [Wang et al.2018] and ArcFace [Deng et al.2022] introduce different forms of margin functions.

Adaptive Loss Functions. Many studies have introduced an element of adaptiveness in the training objective for either hard sample mining [Lin et al.2018, Wang et al.2019b], scheduling difficulty during training [Huang et al.2020, Shrivastava et al.2016], or finding optimal hyperparameters [Zhang et al.2019]

Quality Adaptive Margin for Face Recognition AdaFace [Kim et al.2023] introduce another aspect of adaptiveness in the loss function, namely the image quality. they argue that the strategy to emphasize misclassified samples should be adjusted according to their image quality.

3 Our Proposed

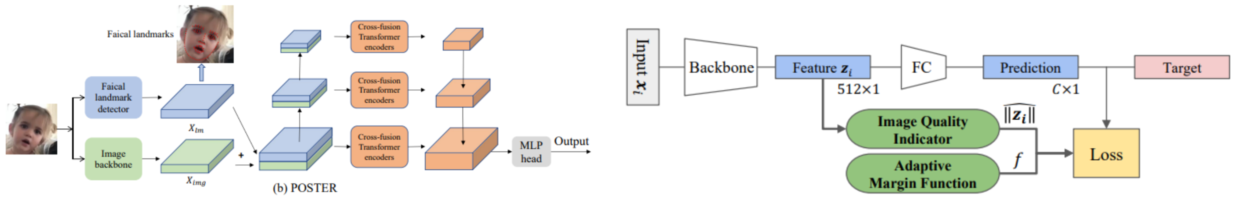
In this section, we present our proposed method for Facial expression problem. We intend to improve POSTER [Zheng et al.2022] model by combining Image-quality feature to loss function as AdaFace [Kim et al.2023]. Therefore we describe the briefly architecture of POSTER and Adaptive margin approach as in Figure 1

3.1 Quality Adaptive Margin for Face Recognition

Advances in margin-based loss functions have resulted in enhanced discriminability of faces in the embedding space. Further, previous studies have studied the effect of adaptive losses to assign more importance to misclassified (hard) examples. [Kim et al.2023] proposed a new loss function that emphasizes samples of different difficulties based on their image quality.

Adaptive Margin based on Norm: Image quality is a comprehensive term that covers characteristics such as brightness, contrast and sharpness. Image quality assessment (IQA) is widely studied in computervision [Zhai and Min2020]. SER-FIQ [Terhörst et al.2020] is an unsupervised DL method for face IQA. BRISQUE [Mittal et al.2012] is a popular algorithm for blind/no-reference IQA. However, such methods are computationally expensive to use during training. In this AdaFace,the feature norm is used as a proxy for the image quality(IQ). They show that the high correlation between the feature norm and the IQ score supports them use of feature norm as the proxy of image quality as in Figure 2

Image Quality Indicator: As the feature norm, $\|z_i\|$ is a model dependent quantity, it is normalized using



(a) POSTER (b) AdaFace

Figure 1: (a) The architecture of POSTER backbone, (b) Proposed adaptive margin function (AdaFace) that is adjusted based on the image quality indicator. If the image quality is indicated to be low, the loss function emphasizes easy samples (thereby avoiding unidentifiable images). Otherwise, the loss emphasizes hard samples.

batch statistics μ_z and σ_z as follow:

$$\widehat{\|z_i\|} = \left\lfloor \frac{\|z_i\| - \mu_z}{\sigma_z/h} \right\rceil_{-1}^{1} \tag{1}$$

where μ_z and σ_z are the mean and standard deviation of all $\|z_i\|$ within a batch. And $\lfloor . \rceil$ refers to clipping the value between -1 and 1 and stopping the gradient from flowing. Since $\frac{\|z_i\|-\mu_z}{\sigma_z/h}$ makes the batch distribution of $\widehat{\|z_i\|}$ as approximately unit Gaussian, they clip the value to be within -1 and 1 for better handling. And h is the hyperparameter to control the concentration.

Adaptive Margin Function: the margin function is designed such that 1) if image quality is high, we emphasize hard samples, and 2) if image quality is low, we de-emphasize hard samples. i achieve this with two adaptive terms g_{angle} and g_{add}, referring to angular and additive margins, respectively.

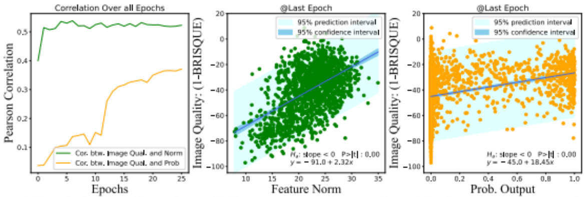
$$f(\theta_j, m)_{AdaFace} = \begin{cases} s(cos(\theta_j + g_{angle}) - g_{add}, j = y_i \\ scos\theta_j, j \neq y_i \end{cases} \tag{2}$$

where g_{angle} and g_{add} are the functions of $\widehat{\|z_i\|}$ as defined:

$$g_{angle} = -m \cdot \widehat{\|z_i\|}, g_{add} = m \cdot \widehat{\|z_i\|} + m \tag{3}$$

3.2 POSTER Backbone

There are three key issues especially prevalent in Facial Expression Recognition (FER): inter-class similarity, intra-class discrepancy, and scale sensitivity. Existing methods typically address some of these issues, but do not tackle them all in aunified framework.



a) Correlation for all epochs b) Feature norm vs img. qual. c) Prob. output vs img. qual.

Figure 2: (a) A plot of Pearson correlation with image quality score (1-BRISQUE) over training epochs. The green and orange curves correspond to the correlation plot using the feature norm $\|z_i\|$ and the probability output for the ground truth index P_{y_i} ,respectively. (b) and (c) Corresponding scatter plots for the last epoch. The blue line on the scatter plot and the corresponding equation shows the least square line fitted to the data points

Therefore, [Zheng et al.2022] proposed a two-stream Pyramid cross-fusion Transformer network (POSTER) that aims to holistically solve these issues. Specifically, the Transformer-based cross-fusion paradigm was designed that enables effective collaboration of facial landmark and direct image features to maximize proper attention to salient facial regions. Furthermore, POSTER employs a pyramid structure to promote scale invariance. The architecture of POSTER is described as in Figure 1(a)

Facial Landmarks are crucial locations on the face that aid in the identification of important regions associated with facial expressions, such as the cheeks or forehead wrinkles. Aside from landmarks, global information beyond the key locations is essential for detecting human expressions.

The Cross-Fusion Transformer Encoder as in

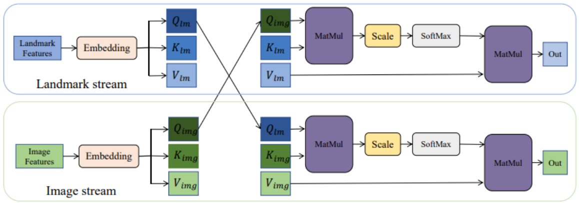


Figure 3: Cross-fusion Multi-head Self-Attention of POSTER

Figure 3 is essential in this procedure. The input image features are mapped to query Q_{img} , key K_{img} , and value V_{img} matrices in the image stream. Similarly, the landmark characteristics in the landmark stream are translated into the equivalent matrices Q_{lm}, K_{lm}, V_{lm} . The attention mechanism is then used to compute the cross-fusion, which combines salient region guidance and global information. A number of mapping functions are used to implement the cross-fusion transformer encoder. The image and landmark queries Q_{img} , Q_{lm} are switched, allowing the image features to incorporate salient regions from the landmarks while receiving global information from the image features.

Feature Pyramid Structure is used to accommodate differences in image quality and resolution. By creating different tiers of extracted features, this structure generates multi-scale features. Cross-fusion transformer encoders process large, medium, and small features separately to record specified scales. The emotion feature is formed by aggregating the outputs of these encoders. Finally, based on the aggregated information, a multilayer perceptron head predicts the emotion label.

3.3 Our combined

In AdaFace [Kim et al.2023] showed that (1) feature norm can be a good proxy for the image quality, and (2) various margin functions amount to assigning different importance to different difficulties of samples.

These two findings are combined in a unified loss function, AdaFace, that adaptively changes the margin function to assign different importance to different difficulties of samples, based on the image quality. In this works, the features are generated by backbone ResNet50 and ResNet100 [He et al.2015] through the Residual Blocks which are important component of the ResNet architecture. From this part, we discuss that whether using a different backbone with different features can improve the for adaface model. Conversely, using the adaptive margin on a backbone with a completely different structure from ResNet can bring an improvement to the model.

Pyramid Cross-Fusion Transformer Network (POSTER) which was designed as a transformer-based cross-fusion paradigm that enables effective collaboration of facial landmark and direct image features to maximize proper attention to salient facial regions.

In this project, we take the idea of using adaptive margin for the backbone as POSTER with the generated feature combined from facial landmark and direct image features to Facial Recognition problem.

As the Figure 1, we use the POSTER's output as the input of Image Quality Indicator by using Norm proxy, then z_i add into Adaptive Margin Function.

4 Experiments and Discussions

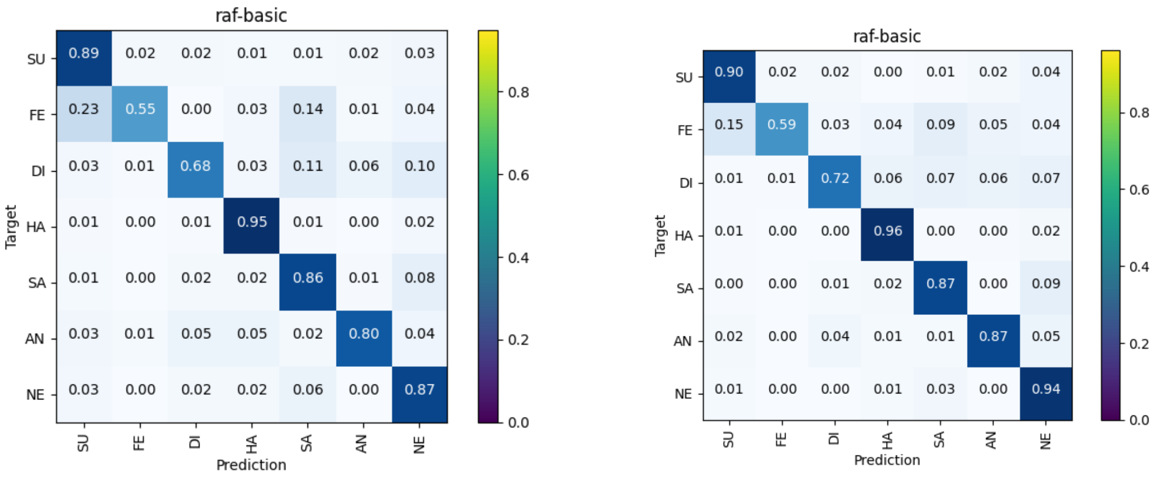
4.1 Data set

RAF-DB: RAF-DB: Real-world Affective Faces Database (RAF-DB) [Li et al.2017] is a large-scale facial expression dataset with 29,672 real-world facial images. All images are collected from the Internet with great variability in the subject's age, gender, ethnicity, lighting conditions, ollusions, etc. For FER task, there are 15,339 facial expression images utilized (12,271 images are used for training and 3,068 images are used for testing) with seven basic expressions (happiness, surprise, sadness, anger, disgust, fear, and neutral)

4.2 Implementation Details

We used POSTER backbone code with Pytorch from github[1] on NVIDIA TITAN Xp GPUs in an end-to-end manner. We utilized IR50 [Deng et al.2022] as the image backbone, which is pretrained on Ms-Celeb-1M dataset [Guo et al.2016]. The weights of the image backbone are updated during training.

[1] https://github.com/zczcwh/FER_POSTER.git



(a) POSTER-Ada-s confusion matrix

(b) POSTER-s confusion matrix

Figure 4: (a) The confusion matrix of POSTER-Ada-s model on RAF-DB dataset, (b) The confusion matrix of POSTER-s model on RAF-DB dataset

For the facial landmark detector, we select MobileFaceNet [Chen2021] with all of the weights frozen to ensure it outputs landmark features. In the feature pyramid structure, we produce $large - medium - small$ extracted features with embedding dim DH = 512, DM = 256, and DL = 128, respectively. For the cross-fusion transformer encoders, each level of encoders consists of depth = 8 transformer encoders. The mlp ratio and drop path rate in transformer encoders are 2 and 0.01, receptively. We set the batch size to 100 and Adam optimizer with a learning rate of 4×10^5.

For Quality Adaptive Margin implementation, we use the Ada head classifier [Kim et al.2023][2] instead of the MLP head classifier as in POSTER. For hyperparameter margin value $m \in \{0.4, 0.75\}$ and $h = 0.33$.

4.3 Our result

Table 1 reports the test set result of our proposed models, employing standard evaluation metrics of the accuracy for Facial classification. The results are categorized as following:

- **POSTER** [Zheng et al.2022]: we categorize into three comparison of pyramid features settings: large - *POSTER-l*, medium - *POSTER-m*, small - *POSTER-s*
- **POSTER with Margin Adaptive** *(POSTER-Ada)*: In this experiment, we use medium and small case for pyramid feature. In small feature case, we compare the performance between margin value at 0.4 and margin value at 0.75. Moreover, we explore our proposed by using extend feature: (1) Keeping MLP head as a additional feature layer - *w/MLP*, (2) using Data Augmentation as in [Kim et al.2023]-*w/DA*, (3) only using IR Feature in Poster to compute Norm for Quality Predictor Image-*IRf*

Firstly, in term of POSTER with Margin Adaptive setting, we show that the performance in value of margin $m = 0.4$ is 3% higher than the other on RAF-DB dataset. Next, in case of extend features, the POSTER-Ada with *MLP* feature reach at 87.32 % while our proposed with Data Augmentation and *IRf* features are slightly lower than *MLP*, at 86.99 % and 87.06 % respectively. It shows that combining facial landmark and direct image features to compute Norm as proxy of Image's quality is best choice in our proposed.

Secondly, the performance with small pyramid feature is 1% higher than the medium one when using adaptive margin to loss function. However, our result doesn't outperform original POSTER setting which is 2% higher than the best our result in the same pyramid feature dimension.

[2] https://github.com/mk-minchul/AdaFace.git

Model	*m*	**Acc** (%)
POSTER-l [Zheng et al.2022]	#	91.59
POSTER-m [Zheng et al.2022]	#	91.87
POSTER-s [Zheng et al.2022]	#	90.94
POSTER-Ada-m	0.4	87.26
POSTER-Ada-s	**0.4**	88.30
	0.75	85.3
POSTER-Ada-s + w/MLP	0.4	87.32
POSTER-Ada-s + w/DA	0.4	86.99
POSTER-Ada-s + w/IRf	0.4	87.06

Table 1: The Accuracy of our proposed on RAF-DB

Additionally, we provide Confusion matrix of the POSTER-Ada-s as in Figure 4(a) and the POSTER-s as in Figure 4(b). As we can see that the performances of the POSTER-Ada-s are much lower than the other in the four class: FE-Fear(-4%), DI-Disgust(-4%), AN-Anger(-7%), NE-Neutral(-7%)

We discuss this issue and give several possible reasons for our result as follow:

In the AdaFace, they used the MS1MV2 [Deng et al.2022] and MS1MV3 [Deng et al.2019] datasets for training, then they evaluated on the other datasets. Maybe RAF-DB is not enough to learn how to emphasize samples follow the quality image. And the simple MLP head is more effective to learn a classifier. Additionally, using Norm as proxy of Image's quality maybe isn't optimization. We'll explore the other proxy for improvement in the future.

5 Conclusion

Facial Expression Recognition (FER) has received increasing interest in the computer vision. community. Because of the time limit of the subject, in this team project, we focus on understanding the basics of the FER problem. We conduct experimental investigation and survey the influence of Quality Adaptive Margin into the POSTER backbone on RAF-DB dataset. Although the our results haven't improved compared to the original POSTER model, this is an interesting starting point for future research on this problem.

References

R. Adyapady and Annappa Basava. 2022. A comprehensive review of facial expression recognition techniques. *Multimedia Systems*, 29, 07.

Cunjian Chen. 2021. PyTorch Face Landmark: A fast and accurate facial landmark detector. Open-source software available at https://github.com/cunjian/pytorch$_face_landmark$.

Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. 2019. Lightweight face recognition challenge. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2638–2646.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. 2022. ArcFace: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, oct.

Paul Ekman and Erika L. Rosenberg. 2005. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 04.

Amir Hossein Farzaneh and Xiaojun Qi. 2021. Facial expression recognition in the wild via deep attentive center loss. pages 2401–2410, 01.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition.

Behzad Hasani and Mohammad H. Mahoor. 2017. Facial expression recognition using enhanced deep 3d convolutional neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jul.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. 2020. Curricularface: Adaptive curriculum learning loss for deep face recognition.

Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. 2015. Joint fine-tuning in deep neural networks for facial expression recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2983–2991.

Muhammad Haris Khan, John McDonagh, and Georgios Tzimiropoulos. 2017. Synergy between face alignment and tracking via discriminative global consensus optimization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3811–3819.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. 2023. Adaface: Quality adaptive margin for face recognition.

Shan Li and Weihong Deng. 2022. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215, jul.

Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2018. Sphereface: Deep hypersphere embedding for face recognition.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708.

Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen, and Hanzi Wang. 2021. Feature decomposition and reconstruction learning for effective facial expression recognition.

Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. 2021. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition.

Jiawei Shi, Songhao Zhu, and Zhiwei Liang. 2021. Learning to amend facial expression representation via de-albino and affinity.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708.

Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. 2020. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition.

Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. 2019a. Region attention networks for pose and occlusion robust facial expression recognition.

Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. 2019b. Mis-classified vector guided softmax loss for face recognition.

Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. 2020. Suppressing uncertainties for large-scale facial expression recognition.

Fanglei Xue, Qiangchang Wang, and Guodong Guo. 2021. Transfer: Learning relation-aware facial expression representations with transformers.

Guangtao Zhai and Xiongkuo Min. 2020. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63.

Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. 2019. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations.

Ce Zheng, Matias Mendieta, and Chen Chen. 2022. Poster: A pyramid cross-fusion transformer network for facial expression recognition.