GReFEL: Geometry-Aware Reliable Facial Expression Learning under Bias and Imbalanced Data Distribution

Azmine Toushik Wasi[1*], Taki Hasan Rafi[2*], Raima Islam[3], Karlo Šerbetar[4], and Dong-Kyu Chae[2†]

[1] Shahjalal University of Science and Technology, Bangladesh
[2] Hanyang University, South Korea
[3] Harvard University, USA
[4] University of Cambridge, United Kingdom
*Co-first authors. †Correspondence to: dongkyu@hanyang.ac.kr

Abstract. Reliable facial expression learning (FEL) involves the effective learning of distinctive facial expression characteristics for more reliable, unbiased and accurate predictions in real-life settings. However, current systems struggle with FEL tasks because of the variance in people's facial expressions due to their unique facial structures, movements, tones, and demographics. Biased and imbalanced datasets compound this challenge, leading to wrong and biased prediction labels. To tackle these, we introduce *GReFEL*, leveraging Vision Transformers and a facial geometry-aware anchor-based reliability balancing module to combat imbalanced data distributions, bias, and uncertainty in facial expression learning. Integrating local and global data with anchors that learn different facial data points and structural features, our approach adjusts biased and mislabeled emotions caused by intra-class disparity, inter-class similarity, and scale sensitivity, resulting in comprehensive, accurate, and reliable facial expression predictions. Our model outperforms current state-of-the-art methodologies, as demonstrated by extensive experiments on various datasets.

Keywords: Facial expression learning · Reliability balancing · Bias and uncertainty · Imbalanced Class Distribution · Bias and Uncertainty

1 Introduction

One of the most universal and significant ways that people communicate their emotions and intentions is through the medium of their facial expressions [33]. In recent years, facial expression learning (FEL) has garnered growing interest within the area of computer vision due to the fundamental importance of enabling computers to recognize interactions with humans and their emotional affect states. While FEL is a thriving and prominent research domain in human-computer interaction systems, its applications are also prevalent in healthcare, education, virtual reality, smart robotic systems, etc [28, 29].

Fig. 1: Complexities of Human Emotions (Green-colored labels are true labels).

Despite recent strides in facial expression recognition technology, the task remains daunting for several reasons. One major hurdle lies in the diverse and complex nature of human facial expressions (as presented in Figure 1). People's facial structures, movements, tones, and demographics contribute to a wide variance in expressions, making it challenging for current systems to accurately interpret and classify them. For instance, telling the difference between a happy smile and a mischievous smirk can be tricky because their lip movements can look quite similar. Also, people express emotions in different ways - some might smile broadly, while others might give a more subtle grin. This variation makes it even harder for computers to accurately pick up on the meaning behind facial expressions. Additionally, consider the challenge of differentiating between a surprised expression and a confused one. Both might involve raised eyebrows and widened eyes, but the context and subtle cues can make a big difference in interpreting the emotion accurately. The complexity of emotions such as anger is often amplified by factors like skin tone and contextual cues, leading to a multitude of potential interpretations that current FEL systems struggle to navigate. These issues, named by intra-class disparity and inter-class similarity, present persistent challenges in facial expression understanding systems [25, 36, 40, 46]. Within-class variations, such as subtle differences in expression intensity or style, pose difficulties in accurately categorizing similar expressions. For instance, a slight change in eyebrow positioning or mouth curvature can drastically alter the perceived emotion, making classification more ambiguous. Conversely, inter-class similarity adds another layer of complexity, as distinct expressions may share common features or gestures, leading to misclassification. Addressing these nuances is crucial for enhancing the reliability and robustness of FEL frameworks, yet current approaches often fall short in effectively mitigating these challenges.

Another significant obstacle in FEL stems from biased and imbalanced datasets used for training. These datasets often fail to adequately represent the diversity of facial expressions across different demographics, leading to skewed and inaccurate predictions. For example, datasets may over-represent certain facial expressions commonly exhibited by a particular demographic while under-representing those of others. This imbalance not only undermines the generalizability of FEL models but also perpetuates biases, resulting in erroneous predictions and reinforcing existing societal disparities.

Researchers use several strategies like unsupervised partitioning, leveraging unlabeled data [23, 40], using loss functions [6, 15], ViTs [19, 25, 46], attention-based models [38] and semi-supervised learning [14]. However, these unsupervised or semi-supervised approaches require extensive additional resources, like large amounts of unlabeled data [4]. Dedicated loss functions for class imbalance may produce harsh results on common labels when prioritizing low-resource classes [9]. ViTs and attention-based models excel in feature extraction, but may cause poor results in complex emotions with subtle changes [44]. This led us to explore methods tailored to effectively handle diverse facial data on a given dataset. As we know, different facial features can be represented as points in a geometric space [24], capturing the diverse connections between facial expressions such as lip, nose, eye, and eyebrow movements. These geometric features serve as descriptors for modeling the complexity of facial expressions.

Based on this perspective, we propose a geometry-based reliability balancing system. By placing learnable anchors with center loss to adapt to different facial landmarks and leveraging anchor loss to utilize geometric connections effectively, we aim to capture complex and interconnected emotions effectively. We also employ window-based cross-attention ViTs for robust feature learning across facial regions, leveraging their strong capability in feature extraction using both local and global information [25]. Combining these methods, we introduce a new reliability balancing approach using facial geometry and an attention mechanism. We place anchor points in the embedding space to measure similarity based on facial geometry features and further use multi-head self-attention to identify important features, enhancing the model's reliability and robustness. This results in improved label distribution and stable confidence scores, mitigating biases and mislabeling caused by various factors. By integrating local and global data using the cross-attention ViT, our approach adjusts for intra-class disparity, inter-class similarity, and scale sensitivity, leading to comprehensive, accurate, and reliable facial expression predictions.

Our contributions are summarized in three folds:

- We propose a novel approach, **GReFEL**, a novel framework consisting of multi-level attention-based feature extraction with a reliability balancing module for robust FEL with extensive data preprocessing and refinement methods to fight against biased data and poor class distributions.
- We introduce geometry-aware adaptive anchors in the embedding space to learn and differentiate between different facial landmarks to increase the reliability and robustness of the model by correcting erroneous labels, stabilizing class distributions for poor predictions, and mitigating the issues of similarity in different classes effectively, addressing intra-class disparity, inter-class similarity, and scale sensitivity.
- Empirically, our **GReFEL** method is rigorously evaluated on diverse in-the-wild FEL databases. Experimental outcomes exhibit that our method consistently surpasses most of the state-of-the-art FEL systems.

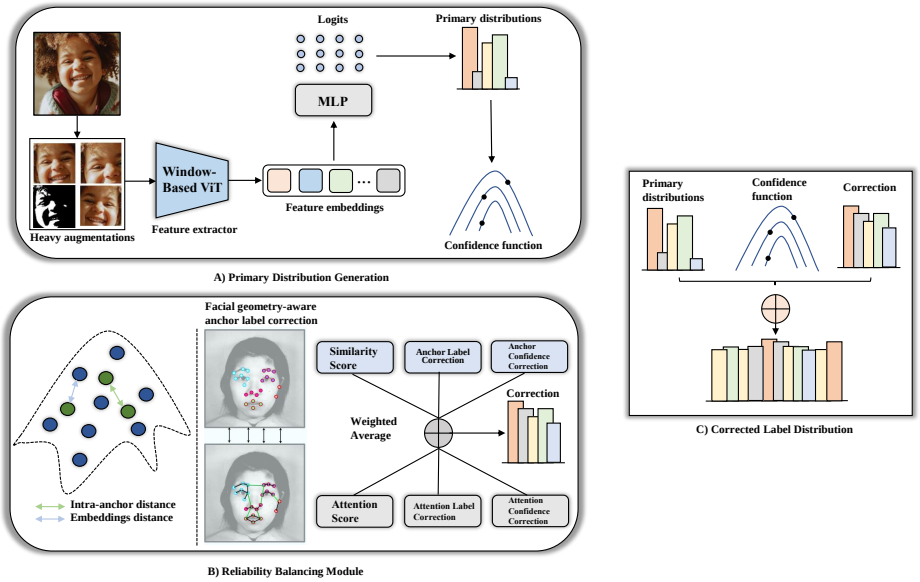


Fig. 2: Pipeline of **GReFEL**. Heavy Augmentation enhances input images, while Data Refinement selects properly distributed class batches per epoch. Window-Based Cross-Attention ViT provides multi-level feature embeddings. MLP predicts primary labels, Confidence is derived from primary label distribution. Reliability balancing utilizes trainable anchors for similarity search and Multi-head self-attention for label correction and confidence calculation. A weighted average of these determines final label correction, resulting in a more reliable model.

2 Related Works

Facial Expression Learning. Facial expression learning involves labeling expressions from facial images, comprising facial detection, feature extraction, and expression recognition phases [33]. Deep learning algorithms, such as self-supervised feature extraction [39], have optimized FEL systems. Recent advancements include multi-branch networks [36], uncertainty estimation [29] and relation-aware local-patch representations [38]. Attention networks based on regions have shown effectiveness for robust FEL [20, 32].

Vision Transformers in FEL. Recent works demonstrate the resilience of Vision Transformers (ViT) against disruption and occlusion [27]. Mask Vision Transformer (MVT) addresses FEL challenges by removing complicated backdrops and occlusion, and adapting labels [19]. Expression Snippet Transformer (EST) effectively models intra/inter snippet changes for video expression recognition [21]. Resilient lightweight multimodal facial expression vision Transformer (MFEViT) handles multimodal FEL data [18]. Neural Resizer balances noise and imbalance in Transformers [10]. Transformer-based multimodal fusion architecture leverages emotional knowledge from diverse viewpoints [43]. POSTER

[46] employs a two-stream pyramid cross-fusion transformer network with a transformer-based cross-fusion method and pyramid structure, while POSTER++ [25] simplifies architecture and enhances performance through improved cross-fusion, two-stream design, and multi-scale feature extraction, combining multi-scale features of landmarks with images.

Other Perspectives on FEL. Researchers address various challenges in FEL through distinct approaches. INV-REG [23] and Meta-Face2Exp [40] reduce data bias using unsupervised partitioning and unlabeled data, respectively. ArcFace [6] and IvReg [15] boost discriminative power and dynamic recognition via novel loss functions and attention mechanisms. EAC [42] and Ada-CM [14] handle noisy labels and semi-supervised learning through advanced training strategies. LatentOFER [13] and LA-Net [37] tackle occlusion and landmark use for improving accuracy and mitigating label noise. M3DFEL [31] introduces temporal modeling, while DAN [35] captures subtle class differences using feature clustering and attention. Each method contributes unique strategies, reflecting the broad spectrum of challenges and innovations in facial expression recognition.

Our approach, GReFEL, extracts features using a cross-window-based ViT to get both local and global information, then collects facial landmark geometry data utilizing geometry-aware anchor points and attention mechanisms to learn about distinctive facial data for different emotions effectively, avoiding bias, imbalance, and uncertainties and producing accurate facial expression predictions in real-world scenarios.

3 Approach

In our approach, we propose a robust feature extraction strategy using ViT and a reliability balancing mechanism to address challenges in FEL. We scale input photos and apply augmentation techniques like rotation and color enhancement for better augmentation. Our pipeline mitigates biases and overfitting by randomly selecting images and expressions during training. Cross-attention ViT is employed for feature extraction, addressing scale sensitivity and intra-class discrepancy. Landmark extraction locates facial landmarks, and a pre-trained image backbone model extracts features. Multiple feature extractors detect low to high-level features, integrated using a cross-attention mechanism for feature vector embedding. Then, primary label distributions are generated using MLPs. Confidence is evaluated using Normalized Entropy. We introduce a reliability balancing method to improve model predictions, addressing limitations in predicting similar classes. Learnable anchors and multi-head self-attention mechanism stabilize label distribution, enhancing reliability. Dropout layers provide additional regularization for robustness against noise and inadequate data. The resulting model, integrating extensive feature extraction and reliability balancing, offers precise and credible predictions even in ambiguous contexts.

Problem Formulation. Let x^i be the i-th instance variable in the input space $\mathcal{X}$ and $y^i \in \mathcal{Y}$ be the label of the i-th instance with $\mathcal{Y} = \{y_1, y_2 \dots y_{N_{cls}}\}$ being the label set. Let $\mathcal{P}^n$ be the set of all probability vectors of size n. Furthermore,

let $l^i \in \mathcal{P}^{N_{cls}}$ be the discrete label distribution of i-th instance. Additionally, let $e = p(x; \theta_p)$ be the embedding output of the Window-Based Cross-Attention ViT (explained in 3.1) network p with parameters θ_p and let $f(e; \theta_f)$ be the logit output of the MLP classification head network f_{CH} with parameters θ_f.

3.1 Feature Extraction

We use a complex image encoder by integrating a window-based cross-attention mechanism, to capture patterns from input images. We extract features by the image backbone and facial landmark detectors. We use IR50 [34] as image backbone and MobileFaceNet [5] as facial landmark detector, both pre-trained models. For each level, firstly, division of image features $X_{img} \in \mathcal{R}^{N_p \times D}$ is performed, where N_p represents the number of patches and D denotes the feature dimensions. The number of patches dictates how the image is fragmented into smaller pieces (e.g., 9 patches would result in 9 small pieces in 3×3 formation). These patches are then transformed into many non-overlapping windows, $z_{img} \in \mathbb{R}^{M \times D}$, where z_{img} contains M tokens. We use 28×28 patches for low-level (local) feature extraction, 14×14 for mid-level, and 7×7 for high-level (global) feature extraction, as described in Section 4.1.

After $z_{img} \in \mathcal{R}^{M \times D}$, down-sampling of the landmark feature $X_{lm} \in \mathcal{R}^{A_c \times H \times W}$ takes place, where A_c is the number of channels in the attention network, H and W are the height and width of the image. The down-sampled features are converted into the window size, where the smaller representation of the image is taken and it is represented by $z_{lm} \in \mathcal{R}^{c \times h \times w}$ where $c = D, h \times w =$ M. The features are reshaped in accordance with z_{img}'s shape. The cross-attention with I heads in a local window can be formulated as follows at this point:

$$q = z_{lm} w_q, k = z_{img} w_k, v = z_{img} w_v \tag{1}$$

$$o^{(i)} = softmax(q^{(i)} k^{(i)T} / \sqrt{d} + b) v^{(i)}, i = 1, \ldots, I \tag{2}$$

$$o = [o^{(1)}, \ldots, o^{(I)}] w_o \tag{3}$$

where w_q, w_k, w_v and w_o are the matrices used for mapping the landmark-to-image features, and q, k, v denote the query matrix for landmark stream, and key, and value matrices for the image stream, respectively from different windows used in the window-based attention mechanism. $[\cdot]$ represents the merge operation where the images patches are combined to identify the correlations between them and lastly, the relative position bias is expressed as $b \in \mathcal{R}^{I \times I}$ which aids in predicting the placement between landmarks and image sectors.

We use the equations above to calculate the cross-attention for all the windows, named by **O**verall **C**ross **A**ttention (OCA), as shown in Figure 3. The transformer encoder for the cross-fusion can be calculated as follows:

$$X'_{img} = OCA_{(img)} + X_{img} \tag{4}$$

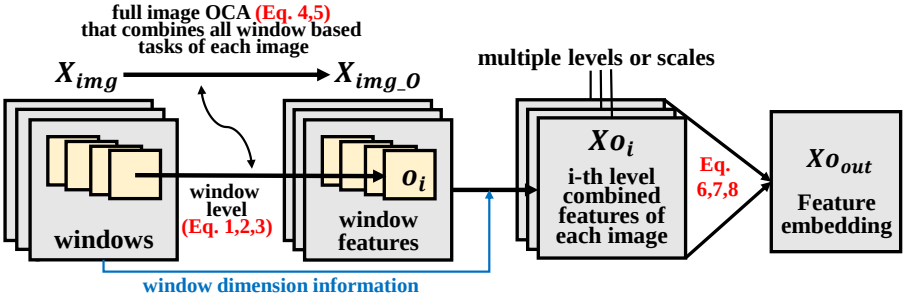


Fig. 3: Data flow in the Window-Based Cross-Attention ViT network

$$X_{img\_O} = MLP(Norm(X'_{img})) + X'_{img} \tag{5}$$

where X'_{img} is the combined image feature using OCA, X_{img_O} the output of the Transformer encoder, and $Norm(\cdot)$ represents a normalization operation for the full image of all windows combined. Using window information and dimensions $(z_{img}, M, D, C, H, W, etc.)$, we extract and combine window based feature information to Xo_i (i-th level window-based combined features of each image) from X_{img_O} (extracted features of all windows of each image together).

We introduce a vision transformer to integrate the obtained features at multiple scales $Xo_1, ..., Xo_i$. Our attention mechanism is able to capture long-range dependencies as it combines information tokens of all scale feature maps:

$$Xo = [Xo_1, ..., Xo_i] \tag{6}$$

$$Xo' = MHSA(Xo) + Xo \tag{7}$$

$$Xo_{out} = MLP(Norm(Xo)) + Xo' \tag{8}$$

where $[\cdot]$ denotes concatenation and $MHSA(\cdot)$ stands for the multi-head self-attention mechanism. Output of the multi-scale feature combination module Xo_{out}, which is equal to feature embedding e, is the final output of the encoder network denoted by $p(x; \theta_p)$.

3.2 Reliability Balancing

Majority of Facial Expression Learning datasets are labeled using only one label for each sample. Inspired by [7,12], we provide an alternative approach, in which, we learn and improve label distributions utilizing a label correction approach. We calculate a label distribution primarily that uses the embedding e directly into the MLP network. Subsequently, the reliability balancing section employs label correction techniques to stabilize the primary distribution. This results in improved predictive performance through more accurate and reliable labeling.

Primary Label Distribution. From sample x, using the p network we can generate the corresponding embedding $e = p(x; \theta_p)$ and using the f-network consisting MLP, we can generate the corresponding primary label distribution:

$$l = softmax(f(e; \theta_f)). \tag{9}$$

We use the information contained in the label distribution with label corrections during training to improve the model performance.

Confidence Function. To evaluate the credibility of predicted probabilities, a confidence function is designed. Let $C_f : \mathcal{P}^{N_{cls}} \rightarrow [0, 1]$, be the confidence function. C_f measures the certainty of a prediction made by the classifier using normalized entropy function H(l). The functions are defined as:

$$C_f() = 1 - H(l) \tag{10}$$

$$H(l) = -\frac{\sum_i l^i \log(l^i)}{N_{cls}}. \tag{11}$$

For a distribution where all probabilities are equal, the normalized entropy is 1, indicating maximum uncertainty, and the confidence value is 0. Conversely, if the probability of one class is 1 and all others are 0, the normalized entropy is 0, indicating no uncertainty, and the confidence value is 1.

3.3 Label Correction

The conundrum of label accuracy, distribution stability, and reliability has been a mainstream problem in FEL. The novel approach we propose to resolve this is a combination of two distinct measures of label correction: anchor label correction (geometric) and attentive correction.

Anchor Label (Geometric) Correction. We define anchor a^{ij} ($i \in \{1, \ldots, N_{cls}\}$, $j \in \{1, 2 \ldots K\}$) to be a point in the embedding space. Let $\mathcal{A}$ be a set of all anchors. During training, we use K trainable anchors for each label, with K being a hyperparameter ($k^{th} \in K$). We assign another label distribution $m^{ij} \in \mathcal{P}^{N_{cls}}$ to anchor a^{ij}, where m^{ij} is defined as:

$$m_k^{ij} = \begin{cases} 1, \text{ if } k = i \quad (k^{th}\text{anchor}) \\ 0, \text{ otherwise} \end{cases} . \tag{12}$$

Intuitively, here it means anchors $a^{1,1}, a^{1,2} \ldots a^{1,K}$ are labeled as belonging to class 1, anchors $a^{2,1}, a^{2,2} \ldots a^{2,K}$ are labeled as belonging to class 2 and so on. To correct the final label and stabilize the distribution, we use geometric information about similarity between the embeddings and anchors. The similarity score is $s^{ij}(e)$ is a normalized measure of similarity between an embedding e and an anchor $a^{ij} \in \mathcal{A}$. The distance between e and a for each batch and class is:

$$d(e, a) = \sqrt{\sum |a - e|^2}. \tag{13}$$

Here, dim_e is the dimension of embedding e. Distances $|a-e|^2$ are reduced over the last dimension dim_e and element–wise square root is taken for stabilizing values. The similarity score s^{ij} is then obtained by normalizing distances:

$$s^{ij}(e) = \frac{\exp(-\frac{d(e,a^{ij})}{\delta})}{\sum_i^N \sum_j^K \exp(-\frac{d(e,a^{ij})}{\delta})} \tag{14}$$

where δ is a hyperparameter used in the computation of Softmax to control the steepness of the function. The default value used for δ is 1.0. From similarity scores we can calculate the anchor label correction term as follows:

$$t_g(e) = \sum_i^N \sum_j^K s^{ij}(e) m^{ij}. \tag{15}$$

Attentive Correction. For multi-head attention [30], Let a query with query embeddings $q \in \mathcal{R}^{d_Q}$, key embeddings $k \in \mathcal{R}^{d_K}$, and value embeddings $v \in \mathcal{R}^{d_V}$ is given. With the aid of independently learned projections, they can be modified with h, which is the attention head. These parameters are then supplied to attention pooling. Finally, these outputs are altered and integrated using another linear projection. The process is described as follows:

$$h_i = f(W_i^{(q)} q, W_i^{(k)} k, W_i^{(v)} v) \in \mathcal{R}^{p_V}, W_{out} = W_o \left[h_1 \dots h_{n_{heads}}\right] \tag{16}$$

where $W_i^{(Q)} \in \mathcal{R}^{d_{model} \times d_Q}$, $W_i^{(K)} \in \mathcal{R}^{d_{model} \times d_K}$, $W_i^{(V)} \in \mathcal{R}^{d_{model} \times d_V}$, and $W_o \in \mathcal{R}^{n_{heads} d_V \times d_{model}}$ are trainable parameters [30], f is the attentive pooling, and each $h_i (i = 1, 2, ..., n_{heads})$ is an attention head. Also, $d_Q = d_K = d_V = d_{model}/n_{heads}$ following [30].

As we are using self-attention, all inputs (q, k, v denoting query, key and value parameters respectively) are equal to the embedding e [30]. Self-attention is applied to individual visual embeddings, not across the entire batch. e is passed through the multi-head self-attention layer to obtain the attentive correction term t_a. t_a is calculated based on the output W_{out} from Eq. (16):

$$t_a = softmax(W_{out}). \tag{17}$$

Multi-head self attention (MHSA) [30] is designed to focus on the crucial parts relevant to a particular class. Self-attention offers context-aware representations for each sequence element, while multi-head self-attention enhances this by learning various aspects of element relationships, resulting in a more robust understanding [3, 30]. In this work, MHSA can identify important facial areas for each class, thereby improving classification accuracy.

Final Label correction. To combine the correction terms, we use weighted sum, with weighting being controlled by the confidence of label corrections:

$$t = \frac{c_g}{c_g + c_a} t_g + \frac{c_a}{c_g + c_a} t_a \tag{18}$$

where $c_g = C_f(t_g)$ and $c_a = C_f(t_a)$. t_a is the attentive correction term, achieved from h by normalizing. $C_f()$ stands for the confidence function, calculates confidence of each class predictions.

Finally, to obtain the final label distribution L_{final}, we use a weighted sum of label distribution l and label correction t, as follows:

$$L_{final} = \frac{c_l}{c_l + c_t} l + \frac{c_t}{c_l + c_t} t \tag{19}$$

where $c_l = C_f(l)$ and $c_t = C_f(t)$. The label with the maximum value in the final corrected label distribution L_{final} is provided as a corrected label or a final predicted label.

3.4 Loss Function

The loss function used to train the model consists of three terms such as class distribution loss, anchor loss, and center loss.

Class Distribution Loss ($\mathcal{L}_{cls}$): To make sure each example is classified correctly, we use the negative log-likelihood loss between the corrected label distribution ${L_{final}}_j^i$ and label y^i:

$$\mathcal{L}_{cls} = -\sum_{i}^{m}\sum_{j}^{N} y_j^i \log {L_{final}}_j^i. \tag{20}$$

Anchor Loss ($\mathcal{L}_a$): In order to amplify the discriminatory capacity of the model, we want to make margins between anchors large so that we add an additional loss term:

$$\mathcal{L}_a = -\sum_i \sum_j \sum_k \sum_l |a^{ij} - a^{kl}|_2^2. \tag{21}$$

We include the negative term in front because we want to maximize this loss. The loss is also normalized for standard uses.

Center Loss ($\mathcal{L}_c$): To make anchors good representation of their class, we want to make sure anchors and embeddings of the same class stay close in the embedding space. To ensure that, we add an additional error term:

$$\mathcal{L}_c = \min_k |x^i - a^{y^i k}|_2^2. \tag{22}$$

Total Loss ($\mathcal{L}_{total}$): Our final loss function can be defined as:

$$\mathcal{L}_{total} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_a \mathcal{L}_a + \lambda_c \mathcal{L}_c \tag{23}$$

where λ_{cls}, λ_a, and λ_c are hyperparameters used to keep the loss functions on the same scale.

4 Experiments

4.1 Experimental Setup

Datasets. We use **AffectNet** [26] (420,299 samples; 8 classes), **Aff-Wild2** [11] (1,413,000 samples), **RAF-DB** [16, 17] (68,718 samples),**FERG-DB** [1]

Table 1: Comparison of Accuracy (%) (↑) with SOTAs.

Models	Datasets					
	AffectNet*†	AffWild2*†	RAF-DB*†	JAFFE*	FER+*†	FERG
SCN [33] (CVPR'20)	56.35	60.55	87.03	86.33	85.97	90.46
RAN [32] (TIP'20)	52.97	59.81	86.90	88.67	83.63	90.22
DMUE [29] (CVPR'21)	61.21	63.64	83.19	-	-	-
Tr.FER [38] (ICCV'21)	66.23	68.92	90.91	-	-	-
RUL [41] (NIPS'21)	60.65	62.37	88.98	92.33	-	92.35
Eff.Face [45] (AAAI'21)	59.89	62.21	88.36	92.33	-	92.16
F2Exp [40] (CVPR'22)	64.23	66.34	88.54	-	-	-
POSTER [46] (IC-W'22)	63.34	67.74	92.05	94.57	91.62	95.82
EAC [42] (ECCV'22)	61.11	63.54	88.02	-	87.03	-
L.OFER [13] (ICCV'23)	63.90	66.02	89.60	-	-	-
LA-Net [37] (ICCV'23)	64.54	66.76	91.56	-	91.78	-
DAN [35] (Bioinf.'23)	62.09	65.82	89.70	-	-	-
POSTER^{++} [25] ('23)	63.76	69.18	92.21	96.67	92.28	96.36
GReFEL (Ours)	**68.02**	**72.48**	**92.47**	**96.67**	**93.09**	**98.18**

†in-the-wild datasets *class-imbalanced

(55,769 samples), **JAFFE** [22] (213 samples), and **FER+** [2] (35,801 samples) datasets, having 6-8 classes. Among them, AffectNet, Aff-Wild2, FER+, and RAF-DB datasets exhibit class imbalances and are collected in real-world settings.

Data Distribution Adjustments. We use sample augmentation to expand the training set in class-imbalanced cases, aiding feature identification. Common FEL pre-processing steps include resizing, scaling, rotating, flipping, cropping, color augmentation, and normalization. Uneven class distributions can cause bias and over-fitting. To counter this, equally distributing information from all classes improves model accuracy. Refining datasets ensures balanced training data, mitigating biases. During training, N_{pg} images are randomly selected from each video or face group. From these, B images per expression are chosen for training, creating a batch of ($B \times N_{cls}$ (number of classes)) images per epoch, reducing biases and overfitting.

Baselines. We utilized the following baselines in our experiments: SCN [33], RAN [32], TransFER (T.FER) [38], DMUE [29], RUL [41], EfficientFace [45], Face2Exp (F2Exp) [40], POSTER [46], EAC [42], Latent-OFER (L. OFER) [13], LA-Net [37], DAN [35], and POSTER^{++} [25].

Implementation Details. For each dataset, we exclusively use cropped and aligned images. These images are resized to 256×256 and then randomly cropped to 224×224 to address overfitting and data imbalance. Heavy augmentation methods are applied during pre-processing as described in Section 4.1. For data refinement, we consider 512 images per video or face group (N_{pg}), combining them to create an unbiased set. During training, we select 500 images per class category (B) from this set. IR50 backbone is trained on Ms-Celeb-1M [8] dataset, MobileFaceNet backbone is trained on Web260M [47] dataset, provided

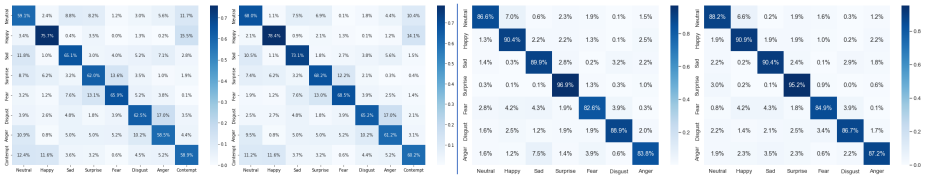


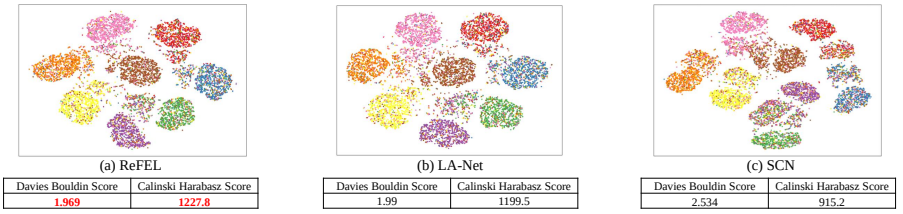
Fig. 4: Confusion Matrix.

via *face.evoLVe* library. Image embeddings are obtained using the Cross Attention ViT network. In feature extraction, we use 28×28 patches for low-level (local) feature extraction, 14×14 for mid-level and 7×7 for high-level (global) feature extraction. In Eq. 16, $d_Q = d_K = d_V = d_{model}/n_{heads} = 64$. Three loss functions are combined for training: Anchor loss maintains distance between anchors, center loss minimizes distance between embeddings and anchors, and class distribution loss ensures correct classification. Our training lasts for 1000 epochs. We employ the ADAM optimizer with an initial learning rate of 0.0003, utilizing exponential decay with γ of 0.995 to optimize the model. Primary prediction is done using an MLP with 2 hidden layers of size 64, each followed by ReLU activation, dropout, and batch normalization, except for the last layer. Dropout layers have a drop probability of 0.5 for regularization. For other models, we use default settings as mentioned in their respective papers.

4.2 Comparison with State-of-the-Art Methods

The table shows the comparison of the accuracy of multiple State-of-the-Art facial expression learning methods. Upon investigation of the results, it is apparent that GReFEL outperforms all other models across all datasets, attaining the highest accuracy scores for each dataset. Specifically, GReFEL earns an accuracy score of 68.02% in AffecteNet, 72.48% on the AffWild2 and 92.47% on RAF-DB dataset (large in-the-wild dataset), which is significantly higher than POSTER++ and the third best TransFER [29] (CVPR'21). Among the compared methods, we think POSTER^{++} (AffectNet 63.76%, AffWild2 69.18%) is the most suitable baseline of our work. Compared to this baseline, our ReFEL achieves 68.02% on AffectNet 72.4% on AffWild2 (3-5% better accuracy than POSTER^{++} on these in-the-wild benchmarks). GReFEL also outperforms every other model in the study, with accuracy scores on the FER+, FERG-DB and JAFFE datasets of 93.09%, 98.18% and 96.67%, respectively, outperforming every other model tested. Our novel reliability balancing section reduces all kinds of biases, resulting in exceptional performance in all circumstances.

Confusion Matrix. Figure 4 shows confusion matrices from the AffectNet and RAF-DB datasets, with and without reliability balancing, and reveals several key insights. In AffectNet, reliability balancing notably enhances true positive rates for most emotions, except for neutral and contempt expressions. Without



(a) ReFEL

Davies Bouldin Score	Calinski Harabasz Score
1.969	1227.8

(b) LA-Net

Davies Bouldin Score	Calinski Harabasz Score
1.99	1199.5

(c) SCN

Davies Bouldin Score	Calinski Harabasz Score
2.534	915.2

Fig. 5: t-SNE visualization of Embeddings with **Davies Bouldin Score (↓) and Calinski Harabasz Score (↑)** of our model **GReFEL** comparing with LA-Net and SCN using *Aff-Wild2* dataset containing 8 classes.

balancing, the classifier struggles with neutral, contempt, and anger distinctions. RAF-DB's performance sees minor improvements with balancing, showcasing a better overall classification compared to AffectNet. Despite this, neutral, contempt, and anger remain challenging to classify accurately. Both datasets show higher true positive rates for surprise and happy expressions, with intriguing confusions between certain emotion pairs like fear and surprise. It indicates that reliability balancing functions effectively, reducing disparities between classes.

Feature Extraction and Clustering. In Fig. 5, the t-SNE plot visually illustrates class differences in the embedding space, with each color representing a distinct class using AffWild2 dataset. The Davies-Bouldin score (↓) evaluates cluster resemblance, while the Calinski-Harabasz score (↑) measures cluster variance. Observations reveal uniformly spaced groups with reliable classifications and noisy areas indicating inter-class similarity and disparity issues. GReFEL outperforms LA-Net and SCN in both Davies-Bouldin (1.969 vs. 1.990 and 2.534) and Calinski-Harabasz scores (1227.8 vs. 1199.5 and 915.2). GReFEL exhibits well-dispersed and discriminating embeddings compared to other models, as evident from the plots and scores.

4.3 Ablation Study

Here we explore the impact of different reliability balancing and loss function setups. More ablation studies are available in the supplementary materials.

Study of Different Model Setups for Reliability Balancing. The table 2 summarizes model setups, their accuracy and the F1 score for the AffWild2 dataset. Integration of the Reliability Balancing (RB) module indicates that the F1 scores significantly increase after using reliability balancing methods. We also observe that the initial ViT-based feature extraction requires 43.6M parameters to achieve an accuracy of 68.15%. However, by incorporating a few additional parameters for reliability balancing, we can significantly enhance the performance, achieving an accuracy of 72.48% in the model. Also, the increment in computational complexity is minimal.

Study of Different Loss Setups. Table 3 summarizes different loss setups and their associated accuracy and F1 score using AffWild2 dataset. Combining classification, anchor, and center losses achieves the highest accuracy of 72.48%,

Table 2: Reliability Balancing Setups.

Model Setup	Accuracy	F1 Score	Params.	FLOPs
Without reliability balancing	68.15%	0.632	43.6M	8.32G
Reliability balancing with only Anchors	70.36%	0.678	43.74M	8.36G
Reliability balancing with only MHSA	69.05%	0.657	43.8M	8.43G
Reliability balancing with both strategies	72.48%	0.731	43.84M	8.51G

Table 3: Loss Setups.

Loss	Accuracy	F1 Score
Class Distribution Loss ($\mathcal{L}_{cls}$)	68.15%	0.682
Anchor Loss ($\mathcal{L}_a$)	67.96%	0.680
Center Los ($\mathcal{L}_c$)	Not Converge	Not Converge
Class Distribution and Anchor Loss ($\mathcal{L}_{cls} + \mathcal{L}_a$)	68.87%	0.692
All Loss Function ($\mathcal{L}_{cls} + \mathcal{L}_a + \mathcal{L}_c$)	72.48%	0.731

indicating enhanced model performance through multi-loss integration. More ablations results can be found in the supplementary material.

5 Conclusion

Our paper introduces **GReFEL**, a novel FEL approach addressing biased and unbalanced data. GReFEL combines attentive feature extraction with reliability balancing using heavy augmentation and data refinement alongside a Vision Transformer (ViT). Our method effectively handles inter-class similarity, intra-class disparity, and label ambiguity. By incorporating trainable anchor points in embedding space to learn and differentiate between different facial expression landmarks, we stabilize distributions and enhance performance. Experimental analysis across datasets demonstrates GReFEL's superiority over state-of-the-art models, highlighting its potential to advance facial expression learning.

Acknowledgements This work was partly supported by (1) the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00345398) and (2) the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2020-II201373, Artificial Intelligence Graduate School Program (Hanyang University)).

References

1. Aneja *et al.*, D.: Modeling stylized character expressions via deep learning. In: ACCV. pp. 136–153 (2016)

2. Barsoum *et al.*, E.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: ICMI (2016)
3. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 397–406 (2021)
4. Chen, Y., Mancini, M., Zhu, X., Akata, Z.: Semi-supervised and unsupervised deep visual learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(3), 1327–1347 (2024)
5. Chen *et al.*, S.: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: CCBR. pp. 428–438 (2018)
6. Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 5962–5979 (2022)
7. Deng *et al.*, J.: Arcface: Additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699 (2019)
8. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition (2016)
9. Horna, D., Mateusz, L., Stefanowski, J.: Deep similarity learning loss functions in data transformation for class imbalance (2023)
10. Hwang *et al.*, H.: Vision transformer equipped with neural resizer on facial expression recognition task. In: ICASSP. pp. 2614–2618 (2022)
11. Kollias, D., Zafeiriou, S.: Aff-wild2: Extending the aff-wild database for affect recognition (2019)
12. Le *et al.*, N.: Uncertainty-aware label distribution learning for facial expression recognition. In: WACV. pp. 6088–6097 (2023)
13. Lee, I., Lee, E., Yoo, S.: Latent-ofer: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1536–1546 (2023)
14. Li, H., Wang, N., Yang, X., Wang, X., Gao, X.: Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4166–4175 (June 2022)
15. Li, H., Niu, H., Zhu, Z., Zhao, F.: Intensity-aware loss for dynamic facial expression recognition in the wild. In: Proceedings of the AAAI conference on artificial intelligence (2023)
16. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE TIP pp. 356–370 (2019)
17. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2852–2861 (2017)
18. Li *et al.*, H.: Mfevit: A robust lightweight transformer-based network for multimodal 2d+ 3d facial expression recognition. arXiv:2109.13086 (2021)
19. Li *et al.*, H.: Mvt: mask vision transformer for facial expression recognition in the wild. arXiv:2106.04520 (2021)
20. Li *et al.*, Y.: Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Trans. on Image Process. pp. 2439–2450 (2018)
21. Liu *et al.*, Y.: Expression snippet transformer for robust video-based facial expression recognition. PR p. 109368 (2023)
22. Lyons, M.J., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets (ivc special issue). arXiv:2009.05938 (2020)

23. Ma, J., Yue, Z., Tomoyuki, K., Tomoki, S., Jayashree, K., Pranata, S., Zhang, H.: Invariant feature regularization for fair face recognition. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20861–20870 (2023)
24. Majumder, A., Behera, L., Subramanian, V.K.: Emotion recognition from geometric facial features using self-organizing map. Pattern Recognition **47**(3), 1282–1293 (2014)
25. Mao *et al.*, J.: Poster v2: A simpler and stronger facial expression recognition network. arXiv:2301.12149 (2023)
26. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1) (2019)
27. Naseer *et al.*, M.M.: Intriguing properties of vision transformers. NIPS **34**, 23296–23308 (2021)
28. Ruan *et al.*, D.: Feature decomposition and reconstruction learning for effective facial expression recognition. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7660–7669 (2021)
29. She *et al.*, J.: Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6248–6257 (2021)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Wang, H., Li, B., Wu, S., Shen, S., Liu, F., Ding, S., Zhou, A.: Rethinking the learning paradigm for dynamic facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17958–17968 (June 2023)
32. Wang *et al.*, K.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE TIP pp. 4057–4069 (2020)
33. Wang *et al.*, K.: Suppressing uncertainties for large-scale facial expression recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6897–6906 (2020)
34. Wang *et al.*, Q.: Face. evolve: A high-performance face recognition library. arXiv:2107.08621 (2021)
35. Wen, Z., Lin, W., Wang, T., Xu, G.: Distract your attention: Multi-head cross attention network for facial expression recognition. Biomimetics **8**(2) (2023)
36. Weng *et al.*, J.: Attentive hybrid feature with two-step fusion for facial expression recognition. In: ICPR. pp. 6410–6416 (2021)
37. Wu, Z., Cui, J.: La-net: Landmark-aware learning for reliable facial expression recognition under label noise. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20698–20707 (October 2023)
38. Xue, F., Wang, Q., Guo, G.: Transfer: Learning relation-aware facial expression representations with transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3601–3610 (2021)
39. Xue *et al.*, F.: Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2412–2418 (2022)
40. Zeng, D., Lin, Z., Yan, X., Liu, Y., Wang, F., Tang, B.: Face2exp: Combating data biases for facial expression recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20259–20268 (2022)

41. Zhang, Y., Wang, C., Deng, W.: Relative uncertainty learning for facial expression recognition. NIPS pp. 17616–17627 (2021)
42. Zhang, Y., Wang, C., Ling, X., Deng, W.: Learn from all: Erasing attention consistency for noisy label facial expression recognition. In: The European Conference on Computer Vision (ECCV). pp. 418–434. Springer (2022)
43. Zhang *et al.*, W.: Transformer-based multimodal information fusion for facial expression analysis. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2428–2437 (2022)
44. Zhao, W., Yang, Z.: An emotion speech synthesis method based on vits. Applied Sciences **13**(4) (2023)
45. Zhao, Z., Liu, Q., Zhou, F.: Robust lightweight facial expression recognition network with label distribution training. In: AAAI. pp. 3510–3519 (2021)
46. Zheng, C., Mendieta, M., Chen, C.: Poster: A pyramid cross-fusion transformer network for facial expression recognition pp. 3138–3147 (2023)
47. Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Lu, J., Du, D., Zhou, J.: Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10492–10502 (2021)