GReFEL: Geometry-Aware Reliable Facial Expression Learning under Bias and Imbalanced Data Distribution (Supplementary Material)

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

In this document, we provide additional ablation results as well as effect of our reliability balancing module.

1 Ablation Studies

1.1 Study of Different Values of λ

The λ values were chosen by our grid search on Aff-Wild2 dataset. Table 1 shows the results. Interestingly, setting all λ values to 1.0, which is our default setting, achieves the best performance.

Table 1: Experimental results with varying λ. Only the selected λ is modified per experiment, with others set to their optimal values.

λ_{cls}	Accuracy (↑)	λ_a	Accuracy (↑)	λ_c	Accuracy (↑)
0.1	35.67%	0.1	68.18%	0.1	69.07%
0.5	57.45%	0.5	69.85%	0.5	71.02%
1.0	**72.48%**	1.0	**72.48%**	1.0	**72.48%**

1.2 Study of Different Loss Functions

Fig. 1 demonstrates the effects of different loss function setups in the training stage of our experiment using AffWild2 [1] dataset. Anchor loss dominance causes the model to drop its performance after some initial good epochs, conveying that the model starts over-fitting on anchors, ignoring true labels. Relying more on similarities than the actual prediction performance, this setup fails to fulfill the criteria. The other setups are quite stable and close. The ideal combination used in the study helps the model train faster and better.

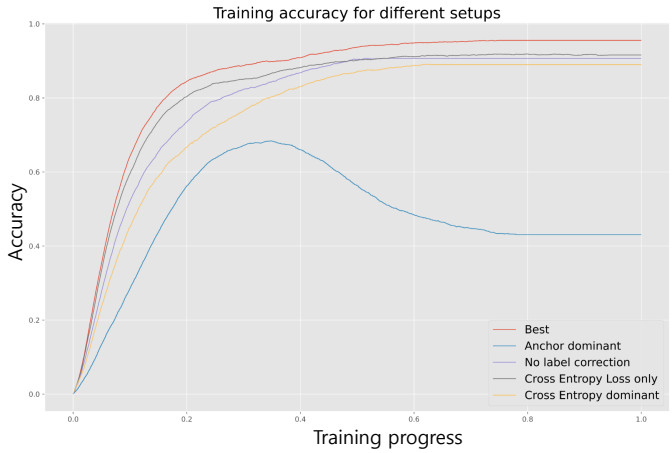

Fig. 1: Study of training progress on different setups using Accuracy (%) score. The red line shows the optimal model with perfect loss combination, blue line shows anchor loss dominant model, indigo colored line shows the model with no label correction with anchors, the gray line shows the model with Cross-Entropy Loss only and the yellow line shows where Cross-Entropy Loss is dominant.

1.3 Effects of Data Augmentation

Table 2 shows that without data augmentation, GReFEL sill obtains competitive performance and outperforms POSTER++ in challenging Aff-Wild2 dataset.

Table 2: Accuracy (↑) with and w/o augmentations and noise.

Model	with Aug.	w/o Aug.	Model	10% Noise	0% Noise
POSTER++	69.18%	66.45%	EAC	63.54%	64.92%
GReFEL	**72.48%**	70.34%	GReFEL	70.55%	**72.48%**

1.4 Study of Different Number of Anchors K

Table 3 demonstrates that optimal recognition accuracy is achieved with 8–10 anchors. Accuracy gradually increases until it reaches this range, beyond which it sharply declines. Few anchors fail to model expression similarities effectively, while excessive anchors introduce redundancy and noise, leading to decreased performance.

1.5 Study of Noise and Label Smoothing

***K* for Different Noise vs. Accuracy.** Table 4 illustrates that increasing noise levels decrease model accuracy due to data clarity and complexity issues in

Table 3: Number of Anchors K vs. Accuracy (%) (↑) means increase in accuracy.

K	0	1	4	6	8	10	20
Accuracy (%)	68.92	+1.82 (↑)	+2.19 (↑)	+2.26 (↑)	**+2.29** (↑)	**+2.29** (↑)	+0.51 (↑)

AffWild2 [1] dataset. However, increasing the value of K improves performance by considering more neighboring points, reducing the impact of noise. Modest yet consistent accuracy improvements are observed with higher K values, but balancing computational complexity is crucial. Over-smoothing from excessively high K values should also be avoided to maintain classification detail.

Table 4: K for Different Noise vs. Accuracy (%) (↑)

K	Noise									
	0	5	10	15	20	25	30	35	40	50
0	68.92	68.41	67.7	63.69	54.99	50.36	41.18	36.94	34.12	30.92
1	70.74	70.52	70.02	66.51	59.81	51.18	44.00	37.76	35.94	33.79
2	70.95	70.61	70.23	66.72	60.02	51.39	44.21	37.97	36.15	34.00
3	71.03	70.64	70.31	66.80	60.10	51.47	44.29	38.05	36.23	34.08
4	71.11	70.65	70.39	66.88	60.18	51.55	44.37	38.13	36.31	34.16
5	71.16	70.70	70.44	66.93	60.23	51.60	44.42	38.18	36.36	34.21
6	71.18	70.71	70.46	66.95	60.25	51.62	44.44	38.20	36.38	34.23
7	71.21	70.71	70.49	66.98	60.28	51.65	44.47	38.23	36.41	34.26
8	71.24	70.72	70.52	67.01	60.31	51.66	44.49	38.26	36.43	34.29
9	71.24	70.73	70.52	67.01	60.32	51.68	44.50	38.26	36.44	34.29
10	**71.25**	70.73	70.53	67.02	60.33	51.69	44.51	38.27	36.45	34.3

***K* for Different Label Smoothing Terms vs. Accuracy.** Table 5 illustrates the impact of label smoothing on model accuracy across various K settings in AffWild2 [1] dataset. Accuracy generally improves with higher K values, with smoothing terms affecting the degree of improvement. For instance, at K=10, maximum accuracy is 71.89% with smoothing term = 5, declining to 51.20% at smoothing term = 40. Smoothing terms between 5 and 20 yield similar accuracy values, making 10 and 11 viable options to balance overconfidence and pattern discovery. A smoothing term of 11 is determined as the optimal choice considering all aspects.

1.6 Study of Primary Mislabeled Predictions

Figure 2 illustrates the proportion of mislabeled images among all mislabeled instances using the AffWild2 dataset. Notably, happiness, sadness, and fear exhibit the highest mis-prediction rates, followed by other and neutral emotions. These trends can be attributed to the intricate nature of certain emotions discussed in

Table 5: K for Different Label Smoothing Terms vs. Accuracy (%) (↑)

K	Label smoothing Terms											
	0	5	10	11	15	18	20	25	30	35	40	50
0	68.92	69.16	69.18	69.18	69.03	68.64	67.50	64.27	61.75	59.34	55.83	50.88
1	70.74	71.38	71.94	71.97	71.46	70.68	70.62	67.59	63.07	60.66	56.15	51.20
2	70.95	71.59	72.15	72.18	71.67	70.89	70.83	67.80	63.28	60.87	56.36	51.41
3	71.03	71.67	72.23	72.26	71.75	70.97	70.91	67.88	63.36	60.95	56.44	51.49
4	71.11	71.75	72.31	72.34	71.83	71.05	70.99	67.96	63.44	61.03	56.52	51.57
5	71.16	71.80	72.36	72.39	71.88	71.10	71.04	67.01	63.49	61.08	56.57	51.62
6	71.18	71.82	72.38	72.41	71.90	71.12	71.06	67.03	63.51	61.10	56.59	51.64
7	71.21	71.85	72.41	72.44	71.93	71.15	71.09	67.06	63.54	61.13	56.62	51.67
8	71.24	71.86	72.44	72.47	71.96	71.18	71.12	67.09	63.57	61.16	56.65	51.70
9	71.24	71.88	72.44	72.47	71.96	71.18	71.12	67.09	63.57	61.16	56.65	51.70
10	71.25	71.89	72.45	**72.48**	71.97	71.19	71.13	67.1	63.58	61.17	56.66	51.71

introduction section of the main paper. Distinguishing subtle variations between happiness and surprise, or between sadness and neutral states, poses challenges for accurate prediction; and our model effectively solves the issue.

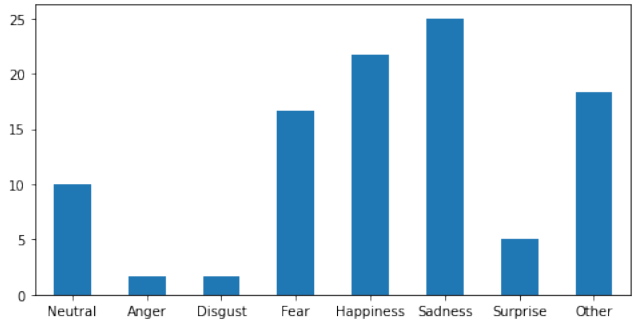

Fig. 2: Percentage of incorrect labels among all incorrect labels in the AffWild2 dataset for GReFEL

We have compared label correction of ours with SCN on the AffWild2 dataset. Figure 3 shows the result of SCN. For SCN, the errors are higher for *Surprise*, *Anger* and *Disgust* more than GReFEL, indicating a more robust feature extraction of GReFEL. Additionally, GReFEL performs better with complex and ambiguous emotions such as *Anger*, *Disgust*, and *Fear* when compared to SCN.

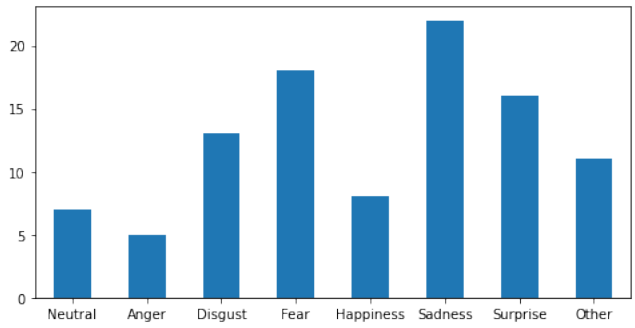

Fig. 3: Percentage of incorrect labels among all incorrect labels in the AffWild2 dataset for SCN

2 Explaining Reliability Balancing

The reliability balancing module plays a crucial role in enhancing the accuracy and reliability of predictions by stabilizing probability distributions in our framework. This strategy increases probability confidence values for appropriate labels while decreasing confidence in incorrect predictions, as Fig. 4 clearly indicates. For instance, Labels 2, 5, and 7 experience a noticeable rise in their maximum confidence values after applying reliability balancing, ensuring more accurate predictions. Conversely, the method reduces the confidence levels of incorrect predictions, as seen in Labels 0, 1, and 3, where the incorrect maximum values decrease to a range of 0.15-0.25. Notably, even in these cases, the correct labels maintain a probability range of 0.2–0.3, enabling the model to make the right predictions. After implementing the corrective measures, the maximum and minimum probabilities across the sample increased to 0.5429 and 0.0059, respectively, resulting in a more stable and balanced distribution. A key observation is that the standard deviation of the corrected predictions (0.0881) was found to be lower than that of the primary predictions (0.1316), providing strong evidence for enhanced stability and balance.

Furthermore, the reliability balancing strategy proves invaluable in scenarios where the primary model struggles with label ambiguity, intra-class similarity, or disparity issues within the images. As evident from Fig. 4, even when the maximum primary probability exceeds 0.4, the associated labels may be erroneous, rendering the model unreliable. Thus, the reliability balancing method supports the model in both extremely uncertain conditions and extremely confident scenarios where the primary model makes poor conclusions.

References

1. Kollias *et al.*, D.: Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. arXiv:2303.01498 (2023) 1, 3

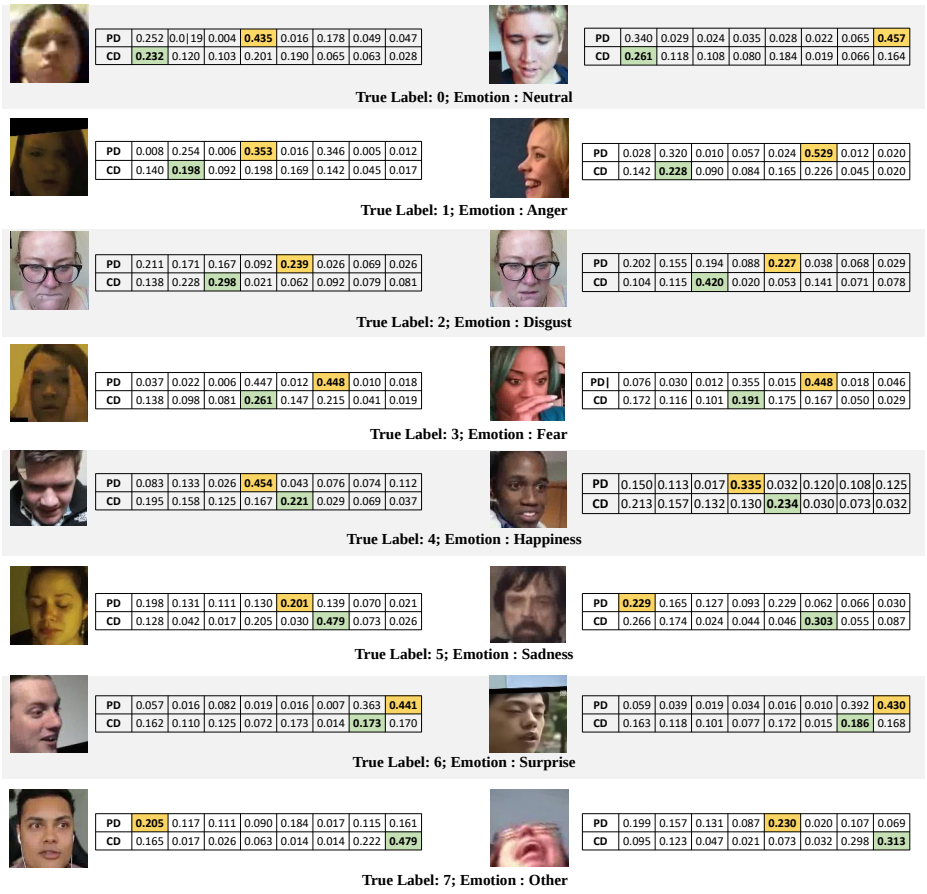

Fig. 4: Observation of **confidence probability distributions in GReFEL** using *Aff-Wild2* dataset. Eight different emotions—**Neutral, Anger, Fear, Disgust, Happiness, Sadness, Surprise,** and **Other**—are represented by columns under each image sequentially. **Primary Distribution (PD)** is the initial prediction, while **Corrected Distribution (CD)** is the accurate prediction after **Reliability Balancing**. The correct label after reliability balancing is marked as green, and the inaccurate primary prediction label is marked as yellow.