Supplementary Materials for Uncertainty-aware Label Distribution Learning for Facial Expression Recognition

Nhat Le*[1, 2, 3], Khanh Nguyen*[1, 2, 5], Quang Tran[3], Erman Tjiputra[3], Bac Le[1, 2], and Anh Nguyen[4]

[1]Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]AIOZ, Singapore
[4]Department of Computer Science, University of Liverpool, Liverpool, UK
[5]FPT Software AI Center, Vietnam

1. Experiments on Original Datasets

Table 1: Test accuracy of different backbone architectures on original RAF-DB and AffectNet datasets.

Backbone	LDLVA	Accuracy (%)	
		RAF-DB	AffectNet
MobileNetV2 [4]	-	85.45	61.32
MobileNetV2 [4]	✓	88.62	65.06
ResNet-18 [1]	-	86.27	62.24
ResNet-18 [1]	✓	89.57	65.74
ResNet-50 [1]	-	87.06	63.19
ResNet-50 [1]	✓	90.51	66.23
ResNet-101 [1]	-	88.28	63.50
ResNet-101 [1]	✓	91.26	66.48

To demonstrate that our method can be easily integrated into existing networks to enhance the robustness, we provide the results corresponding to different backbone architectures. Specifically, we apply LDLVA to the MobileNetV2 [4], ResNet-18 [1], ResNet-50 [1], and ResNet-101 [1] and report the results on RAF-DB [2] and AffectNet [3] datasets in Table. 1. The results show that LDLVA consistently improves the performance of all architectures by a large margin. In addition, we observe that there is a trade-off between the architecture complexity and the computational costs. More complicated architecture can give better performance but requires more memory and computations. This suggests that choosing a suitable backbone architecture is also important for various FER applications.

2. Visualization of Learned Features

In Figure 1, we visualize the t-SNE [6] embedding of the learned features of (a) total loss in the main paper (cross-entropy loss with our discriminative loss), (b) cross-entropy loss only, and (c) center loss on randomly chosen samples from the RAF-DB dataset. Although the center loss (c) makes the features belonging to the same class closer to each other, there still exists overlappings between clusters. This problem is even more obvious on the dataset with noisy labels, since the features on mislabelled examples are pulled toward their wrong labels. In contrast, incorporating our proposed discriminative loss results in more compact and better separated clusters. It can also help mitigate the effect of the noisily learned features. This indicates that learning to discriminate ambiguous facial features is also important for robust facial expression recognition.

3. User Survey for Real-world Ambiguity

As described in the main paper, we randomly picked 21 images from FER test sets and asked 50 volunteers to select the most clearly expressed emotion on each image. We then compare the results of our proposed method and the current state-of-the-art method DMUE [5] on our user survey data. Quantitatively, we use Jeffrey's divergence to measure the difference between the emotion distributions voted by human and predicted by each model. The formula is as

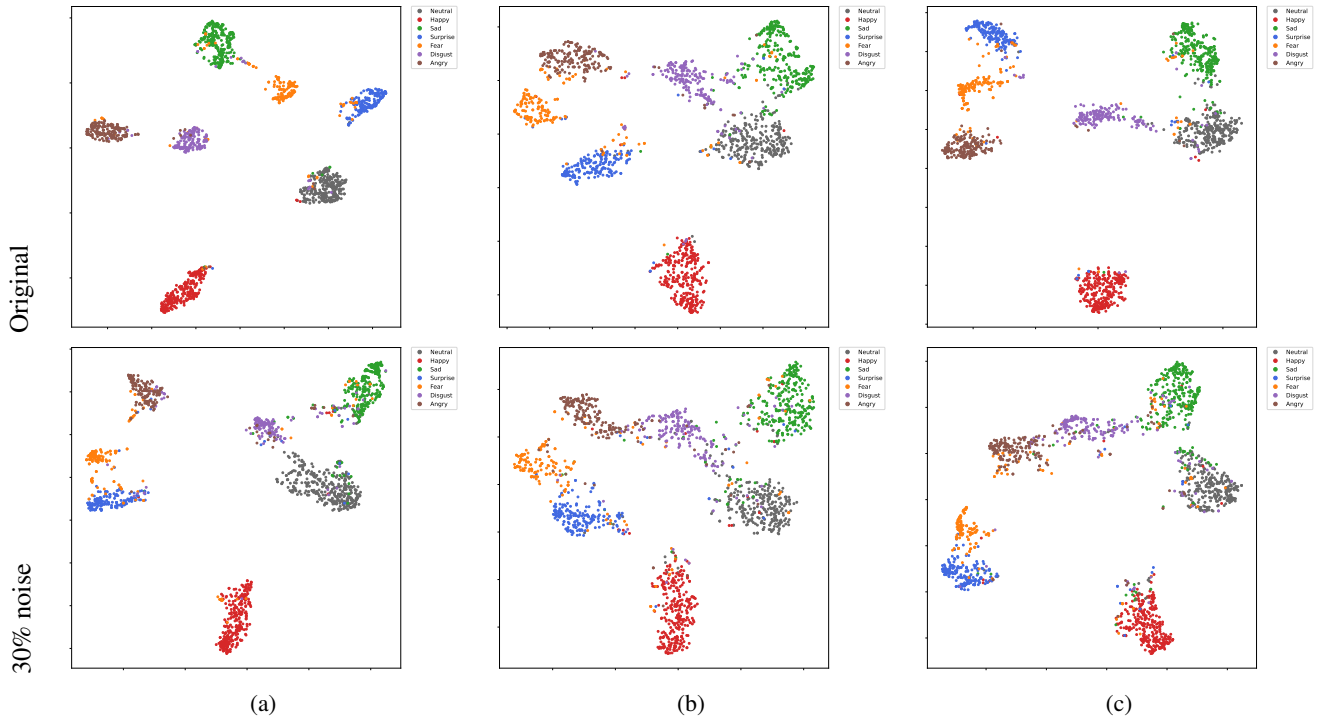

Figure 1: Visualization of the learned features using (a) total loss (cross-entropy with discriminative loss), (b) cross-entropy loss only, and (c) center loss of random samples from RAF-DB dataset

Table 2: Average Jeffrey's divergence of our method and DMUE [5] on survey data. (lower is better)

Method	Jeffrey's divergence
DMUE [5]	4.80
LDLVA (ours)	1.74

follows:

$$D = \frac{1}{n}\sum_{i=1}^{n} D_J\left(f(\boldsymbol{x}^i;\theta), \boldsymbol{d}^i\right) \tag{1}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(d^i_{y_j} - f_j(\boldsymbol{x}^i;\theta)\right)\log\frac{d^i_{y_j}}{f_j(\boldsymbol{x}^i;\theta)} \tag{2}$$

As shown in Table 2, it is clear that the predictions of our method are significantly more similar to human perceptions than those of DMUE [5], demonstrated by a much lower Jeffrey's divergence value.

Furthermore, we also compare the two models qualitatively and present the results in Figure 2. Generally, although the most confident emotion of both models matches with the survey, we can see that the distributions predicted by our model follow human results more closely than the previous method DMUE [5]. While our model tends to distribute scores for different emotions, especially in ambiguous cases, DMUE [5] model gives very high scores for only 1 or 2 classes. This can lead to the case when DMUE [5] fails to cover the expressed emotions (the middle image in the fourth row, for example) but our LDLVA can give the correct prediction.

These results confirm that training models directly with label distributions as in our proposed approach can effectively help to cover multiple emotions that are often expressed in real-world scenarios.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[2] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, 2017.

[3] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2019.

[4] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2:

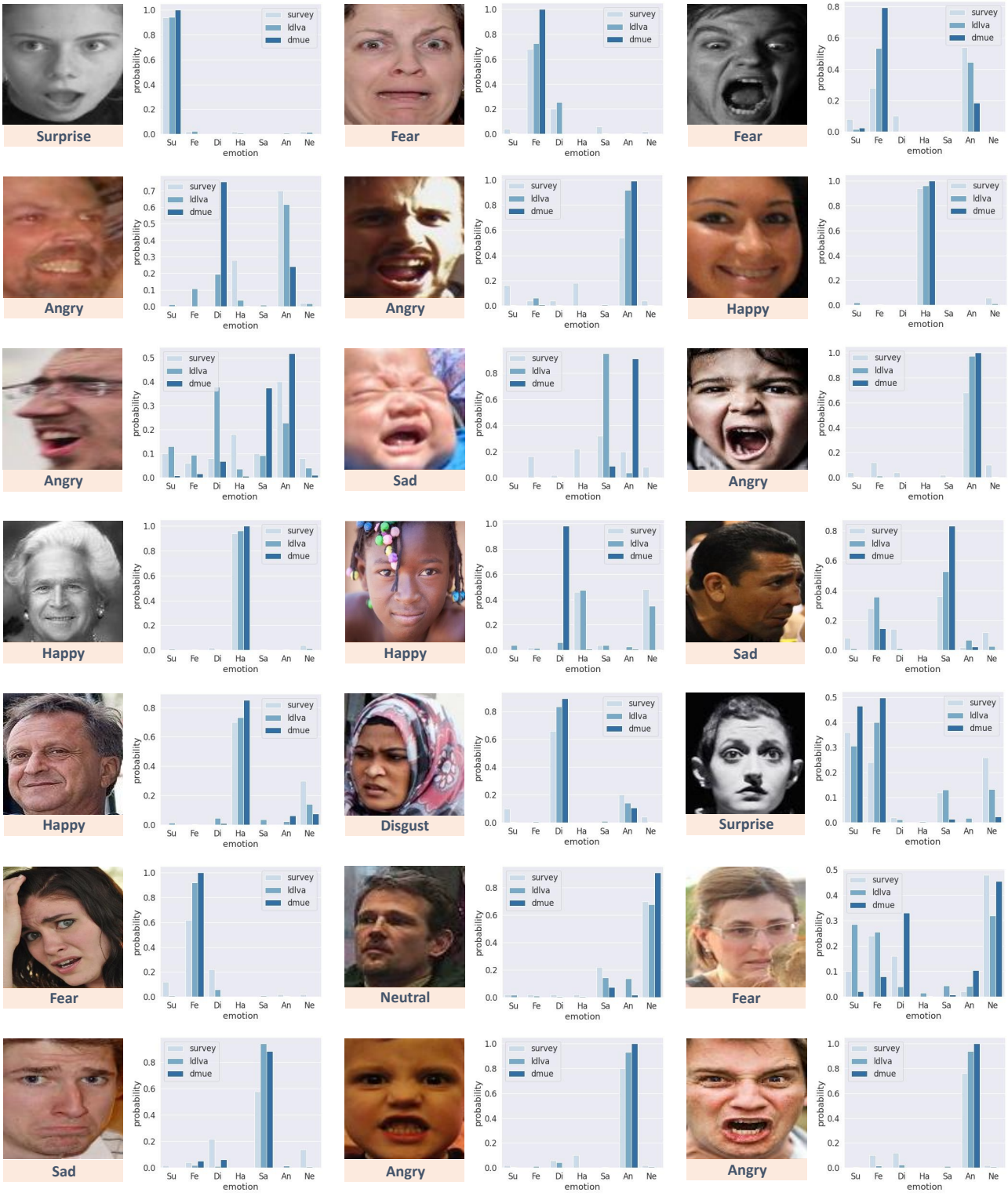

Figure 2: Comparison of the emotion distributions of the user survey, our LDLVA model and DMUE [5] method.

Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.

[5] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *CVPR*, 2021.

[6] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.