LA-Net: Landmark-Aware Learning for Reliable Facial Expression Recognition under Label Noise

Zhiyu Wu, Jinshi Cui*
School of Intelligence Science and Technology, Peking University
wuzhiyu@pku.edu.cn, cjs@cis.pku.edu.cn

Asymmetric Noise	Method	RAF-DB	FERPlus	AffectNet
10%	Baseline	80.93±0.49	82.98±0.09	57.05±0.26
	SCN	81.55±0.28	83.02±0.16	58.49±0.33
	RUL	85.53±0.32	85.12±0.29	60.09±0.10
	EAC	87.44±0.18	86.11±0.14	61.03±0.15
	LA-Net	**88.53±0.20**	**87.21±0.12**	**62.45±0.17**
20%	Baseline	75.62±0.39	80.97±0.27	55.91±0.19
	SCN	79.53±0.42	82.03±0.09	57.03±0.09
	RUL	83.55±0.29	83.76±0.14	58.45±0.21
	EAC	85.09±0.31	84.61±0.28	59.85±0.29
	LA-Net	**86.31±0.21**	**85.44±0.17**	**61.58±0.18**
30%	Baseline	70.38±0.62	76.09±0.29	50.84±0.30
	SCN	74.29±0.39	80.27±0.17	54.56±0.21
	RUL	78.78±0.44	80.99±0.14	56.08±0.11
	EAC	79.62±0.18	82.09±0.11	58.50±0.19
	LA-Net	**81.03±0.24**	**83.62±0.16**	**60.19±0.22**

Table 1. Evaluation (%) on asymmetric noise (anger → disgust, disgust → anger, fear → surprise, happiness → neutral, sadness → neutral, surprise → anger, neutral → sadness, contempt → neutral). We reproduce SCN, RUL, and EAC and report the performance, as they do not consider asymmetric noise in their studies.

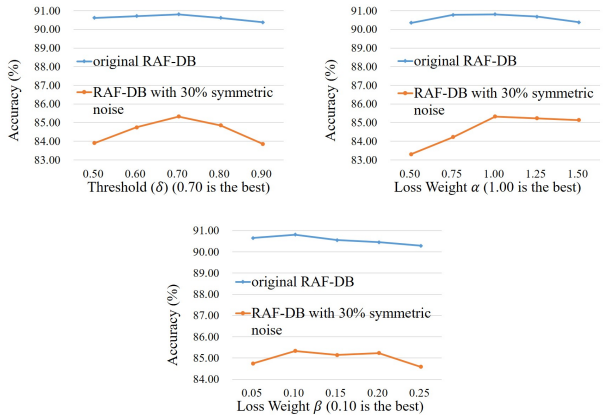

Figure 1. Performance with varying parameters and noise levels.

A. Performance with Asymmetric Noise

Tab. 1 follows prior studies and generates noisy labels by randomly flipping the label to other classes uniformly, referred to as symmetric noise. Given that symmetric noise may not reflect real-world ambiguity, we test asymmetric noise, where the label is flipped to its most similar class based on the confusion matrix. As shown in Tab. 1, LA-Net consistently outperforms previous models when dealing with asymmetric noise. Specifically, compared to EAC, the approach achieves an average promotion of 1.24%, 1.15%, and 1.61% on the three datasets. Overall, the impressive robustness against asymmetric noise demonstrates the potential of LA-Net to be deployed in real-world applications.

B. More Grid Search Results.

We test various temperature values τ (0.05, 0.07, 0.1, 0.15, 0.2) and find 0.1 is optimal. We test various momentum decay ω (0.8, 0.85, 0.9, 0.95, 0.99) and find 0.9 is optimal. We present the grid search for threshold δ and weights α, β in Fig. 1.

C. More User Study Results

We conduct a user study using 50 images selected from AffectNet [3] and compare the one-hot labels, human perception, and generated label distributions to evaluate the effectiveness of our LA-Net. To present our findings more succinctly, we plot all of the mistakenly annotated and ambiguous images, as well as a part of the correctly labeled images (24 in total) in Fig. 2.

The first two lines present several images that are mistakenly annotated in AffectNet (highlighted in red). LA-Net identifies these errors and reveals the latent truth. Moreover, images plotted in rows 3-6 are found to be ambiguous according to user study results (highlighted in blue). Fortunately, the proposed model generates label distributions consistent with human perception, reducing the uncertainty of these ambiguous samples. Additionally, as shown in the last two lines of Fig. 2, LA-Net produces targets that align well with the one-hot labels for correctly annotated samples (highlighted in green). Overall, the model demonstrates some level of agreement with human perception and effectively mitigates label noise.

In addition to visualization, we quantitatively evaluate the consistency between the label distributions and the user

*Corresponding author

study results of the selected images using Jensen-Shannon divergence (*JS* divergence). Mathematically, given probability distributions $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$, their *JS* divergence can be calculated by:

$$JS(\boldsymbol{p_1}||\boldsymbol{p_2}) = \frac{1}{2}KL(\boldsymbol{p_1}||\frac{\boldsymbol{p_1}+\boldsymbol{p_2}}{2}) + \frac{1}{2}KL(\boldsymbol{p_2}||\frac{\boldsymbol{p_1}+\boldsymbol{p_2}}{2}) \quad (1)$$

Tab. 2 reveals a significant difference between one-hot labels and user study results, indicating that FER datasets [2, 3, 1] suffer from serious label noise. In contrast, LA-Net generates label distributions that are in better agreement with human perception. This improvement leads to better FER performance, especially when dealing with label noise.

	user study
one-hot labels	0.2030
label distributions	**0.0909**

Table 2. Jensen-Shannon divergence between the user study results and two expression descriptors.

References

[1] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, pages 279–283, 2016. 2

[2] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2852–2861, 2017. 2

[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, 10(1):18–31, 2017. 1, 2

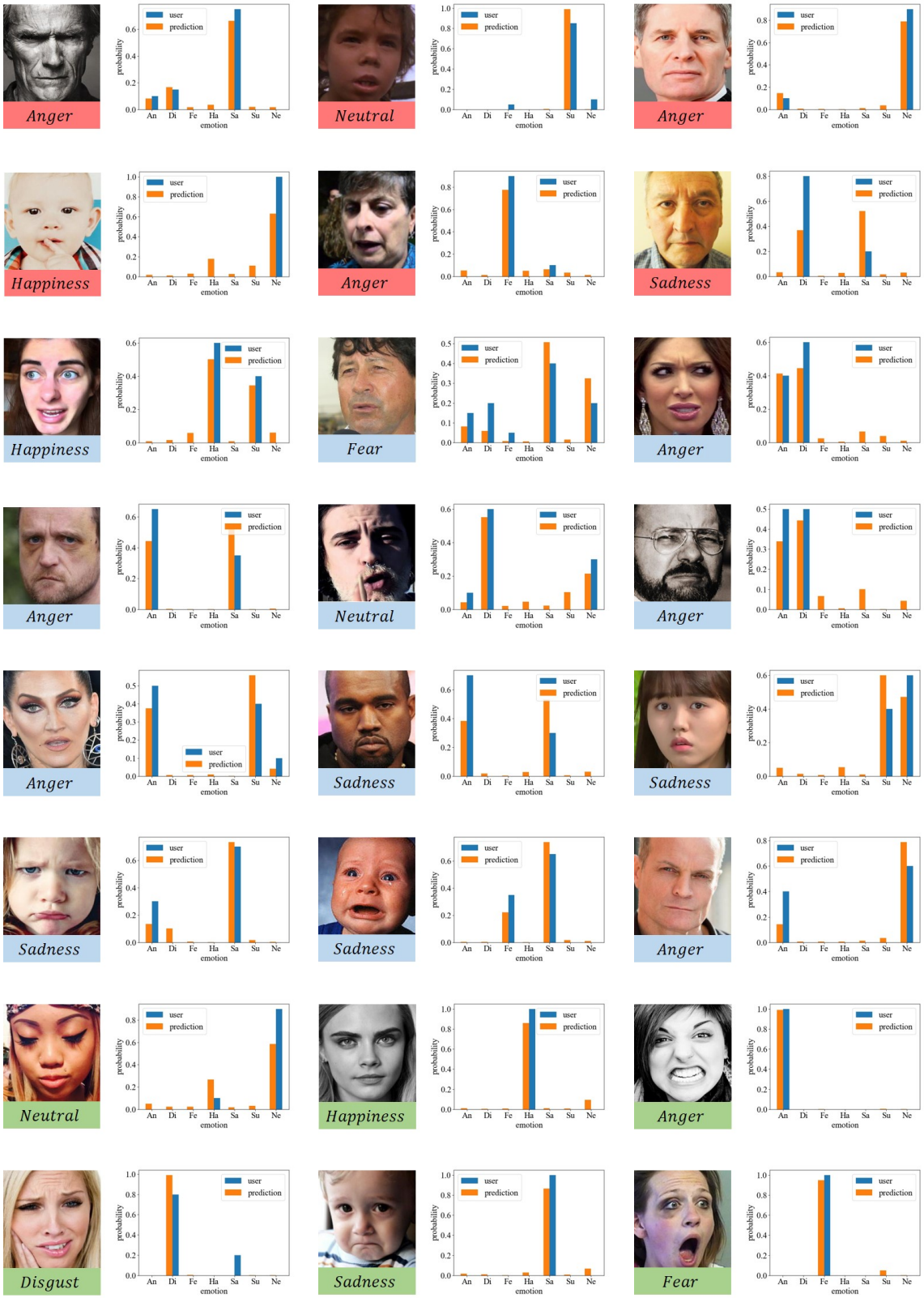

Figure 2. More user study results of the images from AffectNet. We highlight the mistakenly annotated images, ambiguous images, and images with correct labels in red, blue, and green, respectively.