Supplementary Material for Dive into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition

Jiahui She*[1,2] Yibo Hu*[2] Hailin Shi[2] Jun Wang[2] Qiu Shen†[1] Tao Mei[2]
[1]Nanjing University [2]JD AI Research
jh.she@foxmail.com, huyibo871079699@gmail.com, {wangjun492, shihailin}@jd.com
shenqiu@nju.edu.cn, tmei@live.com

1. The Generalized Similarity Preserving Loss

In this section, we give the generalized formulation of L_{sp}, which is defined as the following in the main text:

$$L_{sp} = MSP(G^1_{aux}, \cdots, G^C_{aux}, G_{tar}), \tag{1}$$

where $G^i_{aux} \in \mathbb{R}^{N_i \times N_i}$ and $G_{tar} \in \mathbb{R}^{N \times N}$ denote the similarity matrices calculated by the auxiliary and target branches, respectively. N is the batch size, and N_i is the number of images that are not annotated to the i-th class in the batch. For coding simplicity, given an image batch, we define $\boldsymbol{A}_{tar} \in \mathbb{R}^{N \times N}$ and $\boldsymbol{A}^i_{aux} \in \mathbb{R}^{N \times N}$ ($i \in \{1, 2, \cdots, C\}$), and the j-th row $\boldsymbol{a}_{tar_j}$ of $\boldsymbol{A}_{tar}$ and $\boldsymbol{a}^i_{aux_j}$ of $\boldsymbol{A}^i_{aux}$ are denoted as:

$$\boldsymbol{a}_{tar_j} = \frac{\boldsymbol{f}_{tar_j} \cdot \boldsymbol{f}^T_{tar}}{\left\| \boldsymbol{f}_{tar_j} \cdot \boldsymbol{f}^T_{tar} \right\|_2}, \tag{2}$$

$$\boldsymbol{a}^i_{aux_j} = \frac{\boldsymbol{f}^i_{aux_j} \cdot {\boldsymbol{f}^i_{aux}}^T}{\left\| \boldsymbol{f}^i_{aux_j} \cdot {\boldsymbol{f}^i_{aux}}^T \right\|_2}, \tag{3}$$

where $\boldsymbol{f}_{tar} \in \mathbb{R}^{N \times d}$ and $\boldsymbol{f}^i_{aux} \in \mathbb{R}^{N \times d}$ are the semantic features in the target branch and the auxiliary branch i, respectively, $\boldsymbol{f}_{tar_j} \in \mathbb{R}^{1 \times d}$ and $\boldsymbol{f}^i_{aux_j} \in \mathbb{R}^{1 \times d}$ are the j-th row of $\boldsymbol{f}_{tar}$ and $\boldsymbol{f}^i_{aux}$, respectively, and d is the feature dimension. Then, we implement L_{sp} by masking $\boldsymbol{A}_{tar}$ and $\boldsymbol{A}^i_{aux}$, which L_{sp} can be rewritten as:

$$L_{sp} = \frac{1}{C} \sum_{j=1}^{C} \frac{1}{N_i^2} \left\| \boldsymbol{M}^i * \boldsymbol{A}_{tar} - \boldsymbol{M}^i * \boldsymbol{A}^i_{aux} \right\|_F^2, \tag{4}$$

where $*$ is the element-wise product. The q-th row and p-th column element $m^i_{q,p}$ of $\boldsymbol{M}^i \in \mathbb{R}^{N \times N}$ is defined as:

$$m^i_{q,p} = \begin{cases} 0 & y_p = i \ or \ y_q = i \\ 1 & Others \end{cases}, \tag{5}$$

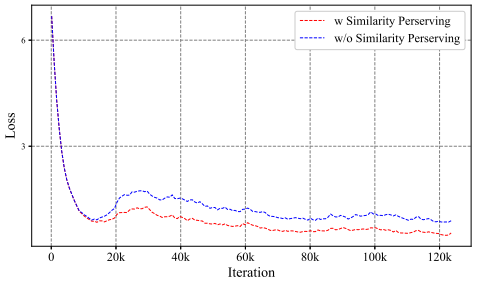

Figure 1: The loss curves of training with L_{sp} (the red curve) and without L_{sp} (the blue curve) along iterations. Experiments are conducted on AffectNet with ResNet-18 as the backbone architecture.

where y_p and y_q denote the annotations of the p-th and q-th images in the batch, respectively ($y_p, y_q \in \{1, \cdots, C\}$). L_{sp} is easy to be implemented by a few lines of code[1].

Benefits of L_{sp}. The non-zero elements in $\boldsymbol{A}_{tar}$ and $\boldsymbol{A}^i_{aux}$ represent the predicted similarity values of image pairs. With the constraint of L_{sp}, all the branches are regularized to predict consistent similarity value for an image pair. As shown in Fig. 1, similarity preserving makes the training more stable. The loss value rises up a little around the $20k$-th iteration step because the ramp functions gradually assign larger weight for L^{target}_{wce} to train the target branch.

2. Evaluation of Synthetic Ambiguity

To better demonstrate the superiority of latent distribution mining, we conduct qualitative analyses on synthetic mislabelled samples. Specifically, a portion of training samples are randomly chosen of which the labels are flipped to other categories. Then, we use DMUE to train the network on the synthetic mislabelled data and visualize the mined

[1]Our source code and pre-trained models will be released.

Table 1: Accuracy (%) on RAF-DB and AffectNet with pre-training on MS-Celeb-1M.

Backbone Architecture	DMUE	AffectNet	RAF-DB
ShuffleNetV1 (group=3;2.0×)	-	56.51	86.20
ShuffleNetV1 (group=3;2.0×)	✓	**60.87**	**88.73**
MobileNetV2	-	57.65	86.01
MobileNetV2	✓	**62.34**	**87.97**
ResNet-18	-	58.85	86.33
ResNet-18	✓	**62.84**	**88.76**
ResNet50-IBN	-	58.94	86.57
ResNet50-IBN	✓	**63.11**	**89.51**

Table 2: Accuracy (%) on RAF-DB and AffectNet without pre-training on MS-Celeb-1M.

Backbone Architecture	DMUE	AffectNet	RAF-DB
ShuffleNetV1 (group=3;2.0×)	-	55.02	85.65
ShuffleNetV1 (group=3;2.0×)	✓	**59.67**	**88.10**
MobileNetV2	-	54.94	85.44
MobileNetV2	✓	**60.43**	**88.15**
ResNet-18	-	55.22	86.01
ResNet-18	✓	**61.22**	**88.33**
ResNet50-IBN	-	55.52	85.67
ResNet50-IBN	✓	**60.54**	**88.94**

latent distributions in Fig. 3. We can observe that the mined latent distribution is able to correct the noisy annotation well. Taking the last image in Fig. 3 as an example, although its true label *Anger* is flipped to *Happy*, the latent distribution well reflects its true class *Anger*. Moreover, the latent distribution also has the capacity to reflect the second possible class for compound expressions. For the first image, the latent distribution reflects its second possible class *Anger*, which is in line with the subjective perception. By imposing the latent distribution as the additional supervision, DMUE effectively utilizes the semantic features of samples.

3. Ablation Study

Different Backbone Networks. As described in the main text, DMUE is independent to the backbone architectures. We further apply DMUE to ShuffleNetV1 [2] and MobileNetV2 [1] to demonstrate the universality of DMUE, where the last stage of ShuffleNetV1 and the last two stages of MobileNetV2 are separated for latent distribution mining. The results on AffectNet and RAF-DB are presented in Table 1. We observe that DMUE can stably improve the performance of all the architectures, including ShuffleNetV1, MobileNetV2, ResNet-18 and ResNet-50IBN, by an average of 4.30% and 2.47% on AffectNet and RAF-DB, respectively. In addition, the ResNet50-IBN achieves the best record among these architectures because of the large number of parameters and the IBN module. Furthermore, we report the results of training DMUE from scratch on AffectNet and RAF-DB in Table 2. Similar observations can also be found without using the pre-trained model on MS-Celeb-1M.

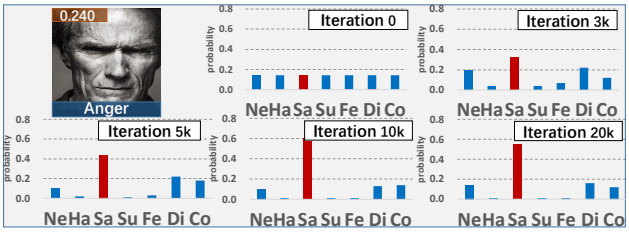

Figure 2: The iteratively updated latent distribution of the sample from Fig. 4 in the main text. Best viewed in color. Zoom in for better view.

Mining latent distribution. In the main text Table 4, we describe the quantitative comparison aiming at investigating which way to mine latent distribution is better. In the quantitative comparison, we train the auxiliary branches with the whole image batch, and their predictions for each image are averaged, denoted as LD-A. As shown in Fig. 4, LD-A reflects the visual feature of images to some extent. But the second and the third possible class of a compound expression is not discriminative in LD-A.

In Fig 2, we provide a toy example of latent label space some intermediate iterations. The latent distribution is completely random at the beginning. It gradually reflects the visual feature of the sample during iterations.

4. More Results and Mathematical Reason for the Uncertainty Estimation

Assume a class center feature $\boldsymbol{c}_i$ of the i-th class, the angle between a i-th class sample's feature $\boldsymbol{x}$ and $\boldsymbol{c}_i$ is θ, that $\theta = acos(\langle \boldsymbol{x}, \boldsymbol{c}_i \rangle)$. Without losing generality, assume $\theta \sim \mathcal{N}(0, \sigma^2)$ in $[-\pi, \pi]$ (as the ambiguity increases, the number of samples decreases). Given a sample $\boldsymbol{a}$ with semantic feature $\boldsymbol{f}$, that $\langle \boldsymbol{f}, \boldsymbol{c}_i \rangle = cos\alpha$. We have:

$$S_{a,i} = \underset{\theta \sim \mathcal{N}(0,\sigma^2), \theta \in [-\pi,\pi]}{E} \{< \boldsymbol{x}, \boldsymbol{f} >\}, \tag{6}$$

where $\|\boldsymbol{c}_i\| = \|\boldsymbol{x}\| = \|\boldsymbol{f}\| = 1$. We first study the equation in a special case where $\alpha = \frac{\pi}{2}$. We denote $\boldsymbol{x}^{\perp}$ as the projection from $\boldsymbol{x}$ to the linear subspace $\boldsymbol{W}$:

$$\boldsymbol{x} = \boldsymbol{c}_i + \boldsymbol{x}^{\perp}, \tag{7}$$

$\mathbb{R}^n = span\{\boldsymbol{c}_i\} \oplus W$, $\oplus$ is the direct sum. Let $\{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{n-1}\}$ is a set of basis of $\boldsymbol{W}$. We have $\boldsymbol{x}^{\perp}$ under uniform distribution as prior for Softmax or other angle-based loss, that is $E\{\langle \boldsymbol{x}^{\perp}, \boldsymbol{z}_k \rangle\} = 0$, with $1 \leq k \leq n-1$. As $\alpha = \frac{\pi}{2}$ in this special case, $\boldsymbol{f}$ can be rewritten as:

$$\boldsymbol{f} = \sum_{k}^{n-1} \omega_k \boldsymbol{z}_k. \tag{8}$$

We have $\boldsymbol{f} \perp \boldsymbol{c}_i$, and have:

$$
\begin{aligned}
E\{< \boldsymbol{x}^{\perp}, \boldsymbol{f} >\} &= E\{\boldsymbol{x}^{\perp} \cdot \sum\nolimits_{k}^{n-1} \omega_k \boldsymbol{z}_k\} \\
&= \sum\nolimits_{k}^{n-1} \omega_k E\{\boldsymbol{x}^{\perp} \cdot \boldsymbol{z}_k\} \\
&= 0, \\
E\{< \boldsymbol{x}, \boldsymbol{f} >\} &= 0.
\end{aligned} \tag{9}
$$

Now, for the general case, we construct:

$$
\begin{cases}
\boldsymbol{f}_1 =< \boldsymbol{f}, \boldsymbol{c}_i > \boldsymbol{c_i}, \\
\boldsymbol{f}_2 = \boldsymbol{f} - \boldsymbol{f}_1,
\end{cases} \tag{10}
$$

$$
< \boldsymbol{x}, \boldsymbol{f} >=< \boldsymbol{x}, \boldsymbol{f}_1 > + < \boldsymbol{x}, \boldsymbol{f}_2 > . \tag{11}
$$

We notice that $\boldsymbol{f}_2 \perp \boldsymbol{c}_i$:

$$
\begin{aligned}
< \boldsymbol{f}_2, \boldsymbol{c}_i > &=< \boldsymbol{f}, \boldsymbol{c}_i > - < \boldsymbol{f}_1, \boldsymbol{c}_i > \\
&=< \boldsymbol{f}, \boldsymbol{c}_i > - << \boldsymbol{f}, \boldsymbol{c}_i > \boldsymbol{c_i}, \boldsymbol{c}_i > \\
&=< \boldsymbol{f}, \boldsymbol{c}_i > - < \boldsymbol{f}, \boldsymbol{c}_i > \|\boldsymbol{c}_i\| \\
&= 0.
\end{aligned} \tag{12}
$$

From Eq. 9, we have $E\{\langle \boldsymbol{x}, \boldsymbol{f}_2 \rangle\} = 0$, so we have:

$$
\begin{aligned}
S_{a,i} &= \underset{\theta \sim \mathcal{N}(0,\sigma^2), \theta \in [-\pi,\pi]}{E} \{< \boldsymbol{x}, \boldsymbol{f} >\} \\
&= \underset{\theta \sim \mathcal{N}(0,\sigma^2), \theta \in [-\pi,\pi]}{E} \{< \boldsymbol{x}, \boldsymbol{f}_1 >\} \\
&= \underset{\theta \sim \mathcal{N}(0,\sigma^2), \theta \in [-\pi,\pi]}{E} \{< \boldsymbol{x}, < \boldsymbol{f}, \boldsymbol{c}_i > \boldsymbol{c_i} >\} \\
&=< \boldsymbol{f}, \boldsymbol{c}_i > \underset{\theta \sim \mathcal{N}(0,\sigma^2), \theta \in [-\pi,\pi]}{E} \{< \boldsymbol{x}, \boldsymbol{c_i} >\} \\
&= \cos\alpha \underset{\theta \sim \mathcal{N}(0,\sigma^2), \theta \in [-\pi,\pi]}{E} \{\cos\theta\}.
\end{aligned} \tag{13}
$$

Obviously, if $\boldsymbol{a}$ is the j-th class sample mislabelled to the i-th class, then $|\alpha|$ is large, $S_{a,i}$ becomes small and $S_{a,j}$ becomes large, which is contrary to the concatenated label. Thus, we can estimate the uncertainty from $\boldsymbol{SV}_a$ as it carries ambiguity information.

We present more visualization results of the estimated uncertainty score in Fig. 6, where lower scores mean more ambiguous images. It is obvious that the estimated uncertainty level is in line with the subjective perception. With the uncertainty estimation module, DMUE is able to suppress the adverse influence from ambiguous data, encouraging the network to utilize the semantic features and learn the latent distribution for the ambiguous image.

5. More Results of User Study

As described in the main text, we pick 20 images from FER datasets and have them labelled by 50 volunteers. We provide more visualization results of the mined latent distribution and the perception from volunteers in Fig. 5. Accordingly, we draw the following conclusions: (**1**) One inherent property of facial expression is that compound facial expressions may exist. It is easy for volunteers to have disagreements with the exact type of images whose annotations in dataset are *Fear*, *Disgust*, *Sad* and *Anger*. One reason may be that folds in the region of eyebrow are often involved in those easily confused expressions. Thus, they may share some common visual features, making it hard to define the exact expression type from a static image. (**2**) The main goal of latent distribution is to provide reasonable guidance to the target branch, rather than finding the exact label distribution of a facial expression image. Thus, we utilize the L_2 loss to minimize the deviation because it is bounded and less sensitive to the incorrect prediction.

Why auxiliary branches work? Based on the conclusions above, we find the one-hot label is hard to represent the visual features of expressions. The annotation of face expression is subjective and difficult, because different expressions naturally entangle each other in the visual space. Auxiliary branches are proposed to disentangle such connections. Each auxiliary branch is a classifier that maps images to their latent classes. By doing so, we disentangle the ambiguity in the label space.

References

[1] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 2

[2] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, 2018. 2

Figure 3: DMUE yields the latent truth for noisy samples. The bottom of each image is tagged by its original annotation. The top of each image is tagged by manually flipped noisy label. DMUE is adopted to train network on synthetic noisy datasets. We visualize the mined latent distribution for synthetic noisy samples at the right of each image. The mined latent distribution is in line with the human subjective perception, where the most possible class reflected by latent distribution is corresponding to the original annotation. (Neu=Neutral, Hap=Happy, Sad=Sad, Sur=Surprise, Fea=Fear, Dis=Disgust, Ang=Anger, Con=Contempt)

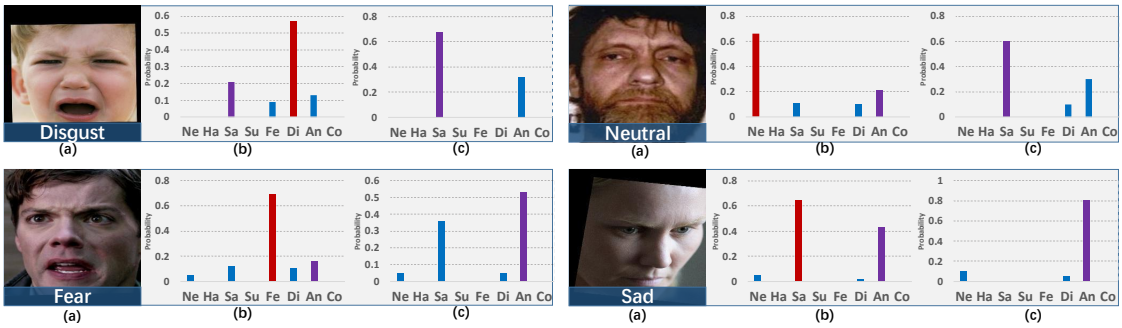

Figure 4: Qualitative comparison between LD-A and LD-N. The red bar denotes the positive class predicted in LD-A. The purple bar denotes the second possible class predicted in LD-A and LD-N for ambiguous images. (a) Images tagged by their original annotation. (b) The LD-A can reflect the visual feature to a certain extent, yet the images' possibility distribution among its negative classes is not discriminative. (c) The LD-N that we used in DMUE, describes an image on its negative classes discriminatively. (Ne=Neutral, Ha=Happy, Sa=Sad, Su=Surprise, Fe=Fear, Di=Disgust, An=Anger, Co=Contempt)

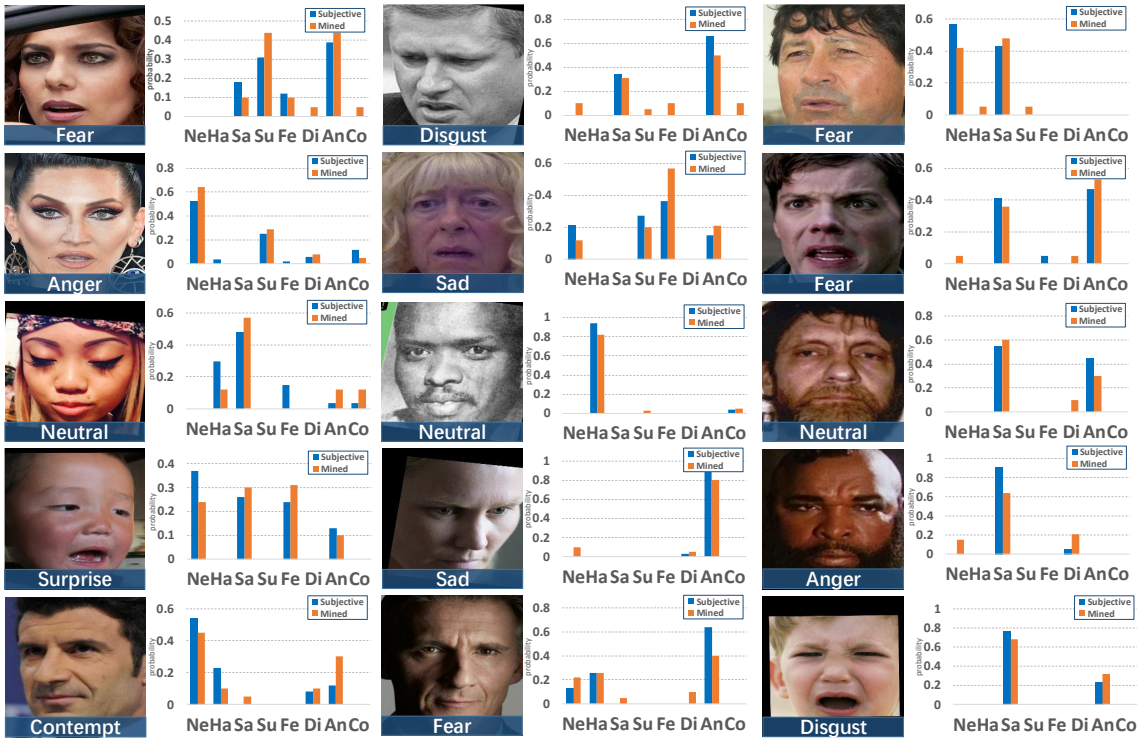

Figure 5: More visualizations of the mined latent distributions and subjective survey results. The age of 50 volunteers range from 17 to 51. Each image is tagged with its annotation. The orange bar denotes the mined latent distribution and the blue bar denotes the subjective survey results. To process the votes from volunteers, we set the number of votes on each sample's positive class as zero. Then we normalize the results, which reflect the probability that image belonging to each negative class. As we can see, the mined latent distribution is consistent with human intuition in general. (Ne=Neutral, Ha=Happy, Sa=Sad, Su=Surprise, Fe=Fear, Di=Disgust, An=Anger, Co=Contempt)

Figure 6: More visualization results of the estimated confidence score. From top to bottom, each row presents images from the same batch. The bottom of each image is tagged with its annotation from the dataset. The upper left of each image is tagged with its estimated uncertainty score. Lower scores are assigned to those more ambiguous images. The upper right of each image is tagged with its confidence rank in the batch. From left to right, we present images with their confidence scores in an ascending order. Images near the right side of the figure are less ambiguous, while images near the left side are in the opposite. In general, we observe that the estimated uncertainty score is in line with the subjective perception. Moreover, we insert an anchor image which is annotated to *Fear* in two different batches (the green bounding box in the 5-th and 7-th row). The uncertainty estimation module predicts consistent confidence score for this anchor image, which indicates the stability of our uncertainty estimation module.