

# Entropy-Calibrated Label Distribution Learning

## A Proof of Theorems

**Theorem 3.1.** Let  $\{\omega_m\}_{m=1}^M$  denote a group of cohesive anchor vectors,  $\mathbf{x}$  denote the feature vector of a sample, and  $\mathbf{y} = \text{softmax}([\langle \omega_m, \mathbf{x} \rangle]_{m=1}^M)$  denote the corresponding output. Without loss of generality, we assume that the anchors are all unit vectors. Then Equation (1) holds if  $\forall i \neq j, \angle(\omega_i, \omega_j) < \tau < \pi$ , where  $\angle(\omega_i, \omega_j)$  denotes the angle between anchors  $\omega_i$  and  $\omega_j$ .

$$\mathcal{H}(\mathbf{y}) \geq \frac{M\lambda^\dagger - 1}{\lambda^\circ - \lambda^\dagger} \cdot \lambda^\circ \log(\lambda^\circ) + \frac{M\lambda^\circ - 1}{\lambda^\circ - \lambda^\dagger} \lambda^\dagger \log(\lambda^\dagger), \quad (1)$$

where  $\lambda^\dagger = Z^{-1} \exp(\cos(\tau^\circ + \tau) \|\mathbf{x}\|)$ ,  $\lambda^\circ = Z^{-1} \exp(\cos(\tau^\circ) \|\mathbf{x}\|)$ ,  $\mathcal{H}(\mathbf{y})$  denotes the entropy of the label distribution  $\mathbf{y}$ ,  $\tau^\circ = \min_m \angle(\omega_m, \mathbf{x})$  is the minimum angle between  $\mathbf{x}$  and anchors,  $Z = \sum_{m=1}^M \exp(\langle \omega_m, \mathbf{x} \rangle)$  denotes the normalization factor, and  $\|\mathbf{x}\|$  denotes the  $L_2$  norm of the feature vector  $\mathbf{x}$ . Equation (1) achieves equality when  $M\lambda^\circ - 1 = k(\lambda^\circ - \lambda^\dagger)$  and  $k$  is a positive integer.

*Proof.* Let  $\lambda^\circ = \frac{1}{Z} \exp(\cos(\tau^\circ) \|\mathbf{x}\|)$  denote the maximum value in  $\mathbf{y}$ . Then the lower bound of the value in  $\mathbf{y}$  is  $\lambda^\dagger = \frac{1}{Z} \exp(\cos(\tau^\circ + \tau) \|\mathbf{x}\|)$ . According to the definition of entropy, the label distribution exhibiting lowest entropy corresponds to the configuration where the maximal number of labels take values  $\lambda^\dagger$  or  $\lambda^\circ$ . Therefore, we define a label distribution with the following configuration:  $k$  labels of value  $\lambda^\circ$ ,  $t$  labels of value  $\lambda^\dagger$ , and one label of value  $c$ , where  $\lambda^\dagger \leq c \leq \lambda^\circ$ . Then we have Equation (2).

$$\begin{cases} c + k \cdot \lambda^\circ + t \cdot \lambda^\dagger &= 1 \\ 1 + k + t &= M \end{cases} \quad (2)$$

Solving Equation (2) yields Equation (3).

$$\begin{cases} k &= \frac{M\lambda^\dagger - \lambda^\dagger + c - 1}{\lambda^\dagger - \lambda^\circ} \\ t &= \frac{-M\lambda^\circ + \lambda^\circ - c + 1}{\lambda^\dagger - \lambda^\circ} \end{cases} \quad (3)$$

Then the entropy of the label distribution  $\mathbf{y}$  can be transformed as Equation (4).

$$\begin{aligned} \mathcal{H}(\mathbf{y}) &= -k\lambda^\circ \log(\lambda^\circ) - t\lambda^\dagger \log(\lambda^\dagger) - c \log(c) \\ &= \frac{M\lambda^\dagger - \lambda^\dagger + c - 1}{\lambda^\dagger - \lambda^\circ} \lambda^\circ \log(\lambda^\circ) - \frac{-M\lambda^\circ + \lambda^\circ - c + 1}{\lambda^\dagger - \lambda^\circ} \lambda^\dagger \log(\lambda^\dagger) - c \log(c) \end{aligned} \quad (4)$$

Then we calculate the first-order and second-order derivatives of  $\mathcal{H}(\mathbf{y})$  w.r.t.  $c$ :

$$\frac{d\mathcal{H}(\mathbf{y})}{dc} = \frac{-(\lambda^\dagger - \lambda^\circ)(\log(c) + 1) + \log(\lambda^\dagger \lambda^\circ - \lambda^\circ)}{\lambda^\dagger - \lambda^\circ}, \quad \frac{d^2\mathcal{H}(\mathbf{y})}{dc^2} = -\frac{1}{c}. \quad (5)$$

Therefore, we have  $\mathcal{H}(\mathbf{y}) \geq \min\{\mathcal{H}(\mathbf{y})|_{c=\lambda^\dagger}, \mathcal{H}(\mathbf{y})|_{c=\lambda^\circ}\}$ . By substituting  $c = \lambda^\dagger$  and  $c = \lambda^\circ$  into Equation (4) respectively, we ultimately obtain the lower bound of the entropy of  $\mathbf{y}$ :

$$\mathcal{H}(\mathbf{y}) \geq \frac{M\lambda^\dagger - 1}{\lambda^\circ - \lambda^\dagger} \cdot \lambda^\circ \log(\lambda^\circ) + \frac{M\lambda^\circ - 1}{\lambda^\circ - \lambda^\dagger} \lambda^\dagger \log(\lambda^\dagger). \quad (6)$$

□

Table 1: Dataset Information.

Dataset	# of Samples	# of Features	# of Labels	$\tilde{\mathcal{H}}$
Jaffe	213	243	6	$0.96 \pm 0.03$
BU-3DFE	2,500	243	6	$0.95 \pm 0.04$
Movie	7,755	1,869	5	$0.88 \pm 0.06$
Music Mood	360	5,992	9	$0.94 \pm 0.03$
Natural Scene	2,000	294	9	$0.47 \pm 0.27$
Emotion6	1,980	168	7	$0.64 \pm 0.16$
Art Painting	280	142	8	$0.72 \pm 0.13$
M2B	1240	250	5	$0.41 \pm 0.12$

## B Experiments

### B.1 Datasets

We select eight datasets: Jaffe, BU-3DFE, Movie, Music Mood, Natural Scene, Emotion6, Art Painting, M2B. Their quantitative information is presented in Table 1. Jaffe dataset [2] consists of 213 grayscale facial expression images ( $256 \times 256$  pixels) from 10 Japanese female expressors, with each providing 3-4 examples of the six basic emotions (happiness, sadness, surprise, anger, disgust, fear) and a neutral expression. The images were captured under controlled lighting conditions using a semi-reflective plastic sheet to maintain consistent positioning and later digitized via a flatbed scanner. In terms of the feature extraction, we followed the processing method in [1]. The labeling process involved 92 Japanese female undergraduates who rated the intensity of each expression on a 1-4 scale (1 = none, 4 = strongest). The raters were divided into four groups, each assessing different subsets of images (some excluded fear expressions). The final labels were 6-dimensional vectors representing the average intensity scores for each expression across all relevant raters, providing continuous emotion measures for facial expression recognition tasks.

BU-3DFE dataset [4] is a comprehensive collection of 3D facial expression data designed to advance research in facial behavior analysis. It contains 2,500 samples from 100 subjects, each performing seven prototypical facial expressions: neutral, happiness, surprise, fear, sadness, disgust, and anger. These expressions are captured at four intensity levels (low, middle, high, and highest), resulting in 25 distinct facial models per subject. The data was collected using a 3D face digitizing system, which includes multiple synchronized digital cameras and light projectors. The system captures both the 3D geometry of the face and the 2D texture images from two angles ( $+45^\circ$  and  $-45^\circ$ ). The raw data consists of 2,500 3D face shape models, which capture the detailed geometry of the face. In addition, 2,500 corresponding 2D texture images are captured, providing surface color and texture information. The raw data is processed to generate several key outputs. These include 2,500 cropped face shape models, which isolate the face region from the rest of the scan, and 2,500 frontal textures of the facial regions, which are used to standardize texture data for analysis. Additionally, 2,500 datasets of facial feature points and the original facial poses are included, helping researchers assess the positioning and movements of the face during expression. In terms of the feature extraction and label distribution generation, we followed the treatment in [1].

Movie dataset [1] is a movie rating dataset which includes 7,755 movies and 54,242,292 ratings from 478,656 Netflix users. The ratings are on a scale from 1 to 5 (5 labels). The rating distribution for each movie is quantified by the proportion of each rating score. The features of the movie are extracted as a 1,869-dimensional vector from the movie metadata such as director, actor, country, budget, etc.

Music Mood dataset comprises 360 pop songs (120 each from Brazil, South Korea, and the US), with audio features extracted using Spotify’s Web API, including danceability (rhythmic quality), energy (perceived intensity), valence (emotional positivity), and acousticness (non-electronic instrumentation). These features were chosen for their relevance to mood perception in music information retrieval (MIR). Human annotations were collected from 166 participants across the three countries, who rated each song on nine mood attributes (sad, cheerful, energetic, calm, dreamy, love, tense, danceable, electronic) using 4-point scales, alongside familiarity (4-point scale) and preference (5-point scale). The mood terms were rigorously validated across English, Portuguese, and Korean

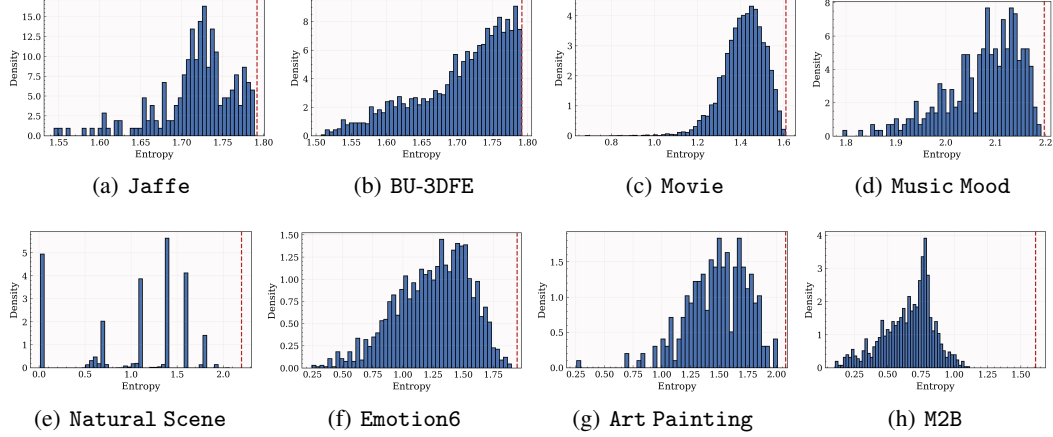


Figure 1: Entropy distribution. The red dashed line denote the maximum entropy given corresponding number of labels, i.e.,  $\log(M)$ .

via image-based semantic alignment to ensure cross-cultural consistency. In total, 11,500 ratings were collected.

**Natural Scene** dataset [1] results from the multilabel rankings of 2,000 natural scene images. Each image is associated with nine possible labels, i.e., plant, sky, cloud, snow, building, desert, mountain, water, and sun. For each image, ten annotators first select the labels that are relevant to the image, and then rank the relevant labels in descending order of relevance to the image. Then, the rankings for each image are transformed into a label distribution by a nonlinear programming process. Finally, for each image, a 294-dimensional feature vector is extracted.

**Emotion6** dataset was systematically constructed through a rigorous pipeline involving the collection of 1,980 images, each annotated with Ekman’s six basic emotions (anger, disgust, joy, fear, sadness, surprise) and a neutral label. Images were sourced from Flickr using emotion-related keywords as search queries, followed by manual filtering to exclude samples containing strong facial expressions or emotion-associated text, thereby focusing on the influence of low-level visual features. In terms of the feature extraction, we followed the treatment in [3]. Annotation was crowdsourced via Amazon Mechanical Turk, where participants performed dual labeling tasks: (1) selecting one or more keywords from seven options (six basic emotions plus neutral) to describe evoked emotions, generating probability distributions of emotional responses, and (2) rating valence and arousal using the 9-point Self-Assessment Manikin (SAM) scale.

**Art Painting** contains 280 ambiguous artworks, with each piece receiving multiple rounds of emotional annotations from approximately 230 online crowdworkers, averaging 14 votes per painting; the highest-voted emotion was assigned as the label after excluding samples with significant annotation discrepancies. In terms of the feature extraction, we followed the treatment in [3].

**M2B** dataset is a designed to comprehensively analyze the attractiveness of a person through three modalities: facial appearance, attire, and voice. The dataset comprises samples from 620 Eastern and 620 Western subjects, each including a  $128 \times 128$ -pixel facial image, a full-body attire image, and a 5-second voice clip, primarily collected from YouTube TV programs and academic lecture videos. For feature extraction, facial images are characterized using Local Binary Patterns (LBP), Gabor filter responses, color moments, and other features (reduced to 250 dimensions) to capture texture and color information; attire images are described by HOG, LBP, color histograms, and other descriptors to represent local details of upper and lower garments (reduced to 300 dimensions); while voice clips are processed to extract pitch, spectral, rhythm features, and others (reduced to 50 dimensions). The label distribution is normalized from the attractiveness scores (1–10 points) which are obtained through a 10-wise ranking tool to derive global preferences across six tasks (single-modality: facial F, attire D, voice V; multi-modality combinations: facial+attire FD, facial+voice FV, trimodality FDV), with each task annotated by at least 15 participants from the same cultural background.

Finally, we visualize the entropy distribution of each dataset in Figure 1.

Table 2: Performance of Comparison Algorithms with  $L_2$  Regularization.

	KL ( $\downarrow$ )			Cosine ( $\uparrow$ )			
	ECA	LEA	HEA	ECA	LEA	HEA	
LDLIAR w/ $L_2$	<b>0.040</b> $\pm 0.005$	<b>0.051</b> $\pm 0.018$	0.044 $\pm 0.002$	<b>0.963</b> $\pm 0.005$	<b>0.952</b> $\pm 0.017$	0.958 $\pm 0.002$	Jaffe
LDL-LDM	0.045 $\pm 0.007$	0.071 $\pm 0.016$	0.041 $\pm 0.002$	0.957 $\pm 0.007$	0.932 $\pm 0.016$	0.961 $\pm 0.002$	
LDL-DPA	0.074 $\pm 0.012$	0.097 $\pm 0.025$	0.074 $\pm 0.008$	0.938 $\pm 0.010$	0.919 $\pm 0.025$	0.935 $\pm 0.007$	
LDL-FCC	0.041 $\pm 0.005$	0.063 $\pm 0.018$	<b>0.041</b> $\pm 0.002$	0.961 $\pm 0.005$	0.940 $\pm 0.017$	<i>0.961</i> $\pm 0.002$	
LDL-LRR	<i>0.041</i> $\pm 0.005$	<i>0.062</i> $\pm 0.018$	<i>0.041</i> $\pm 0.002$	<i>0.961</i> $\pm 0.004$	<i>0.940</i> $\pm 0.017$	<b>0.961</b> $\pm 0.002$	
LDLIAR w/ $L_2$	<b>0.054</b> $\pm 0.002$	<b>0.067</b> $\pm 0.003$	0.052 $\pm 0.002$	<b>0.949</b> $\pm 0.002$	<b>0.937</b> $\pm 0.003$	0.948 $\pm 0.002$	BU-3DPE
LDLLDM	0.060 $\pm 0.005$	0.079 $\pm 0.008$	<i>0.049</i> $\pm 0.002$	0.943 $\pm 0.005$	0.925 $\pm 0.009$	<i>0.951</i> $\pm 0.002$	
LDLDPA	<i>0.055</i> $\pm 0.003$	<i>0.069</i> $\pm 0.003$	0.051 $\pm 0.002$	<i>0.948</i> $\pm 0.003$	<i>0.936</i> $\pm 0.003$	0.949 $\pm 0.002$	
LDLFCC	0.057 $\pm 0.002$	0.075 $\pm 0.003$	<b>0.049</b> $\pm 0.002$	0.945 $\pm 0.002$	0.929 $\pm 0.003$	<b>0.951</b> $\pm 0.002$	
LDLLRR	0.057 $\pm 0.003$	0.074 $\pm 0.005$	0.050 $\pm 0.002$	0.946 $\pm 0.003$	0.930 $\pm 0.005$	0.950 $\pm 0.002$	
LDLIAR w/ $L_2$	<b>0.255</b> $\pm 0.052$	<b>0.425</b> $\pm 0.113$	0.094 $\pm 0.004$	<b>0.854</b> $\pm 0.024$	<b>0.753</b> $\pm 0.062$	0.937 $\pm 0.002$	Movie
LDLLDM	0.262 $\pm 0.054$	0.442 $\pm 0.121$	0.092 $\pm 0.004$	0.851 $\pm 0.024$	0.743 $\pm 0.066$	0.939 $\pm 0.002$	
LDLDPA	0.263 $\pm 0.058$	0.447 $\pm 0.127$	0.102 $\pm 0.004$	0.850 $\pm 0.026$	0.743 $\pm 0.066$	0.932 $\pm 0.003$	
LDLFCC	0.261 $\pm 0.053$	0.441 $\pm 0.123$	<i>0.091</i> $\pm 0.003$	0.852 $\pm 0.024$	0.744 $\pm 0.067$	<i>0.939</i> $\pm 0.002$	
LDLLRR	<i>0.260</i> $\pm 0.054$	<i>0.438</i> $\pm 0.123$	<b>0.090</b> $\pm 0.003$	<i>0.852</i> $\pm 0.024$	<i>0.745</i> $\pm 0.067$	<b>0.940</b> $\pm 0.002$	
LDLIAR w/ $L_2$	0.129 $\pm 0.011$	<b>0.183</b> $\pm 0.020$	0.077 $\pm 0.009$	0.911 $\pm 0.006$	<b>0.880</b> $\pm 0.012$	0.937 $\pm 0.007$	Music Mood
LDLLDM	0.128 $\pm 0.013$	0.188 $\pm 0.021$	0.069 $\pm 0.007$	0.913 $\pm 0.007$	0.876 $\pm 0.012$	0.944 $\pm 0.005$	
LDLDPA	0.148 $\pm 0.016$	0.202 $\pm 0.022$	0.089 $\pm 0.008$	0.897 $\pm 0.010$	0.864 $\pm 0.014$	0.929 $\pm 0.006$	
LDLFCC	<i>0.127</i> $\pm 0.011$	<i>0.187</i> $\pm 0.019$	<i>0.068</i> $\pm 0.006$	<i>0.913</i> $\pm 0.006$	<i>0.876</i> $\pm 0.011$	<i>0.945</i> $\pm 0.005$	
LDLLRR	<b>0.127</b> $\pm 0.011$	0.187 $\pm 0.019$	<b>0.068</b> $\pm 0.006$	<b>0.913</b> $\pm 0.006$	0.876 $\pm 0.011$	<b>0.945</b> $\pm 0.005$	
LDLIAR w/ $L_2$	<b>0.710</b> $\pm 0.013$	<b>0.888</b> $\pm 0.037$	<b>0.626</b> $\pm 0.020$	<b>0.767</b> $\pm 0.005$	<b>0.757</b> $\pm 0.010$	<b>0.756</b> $\pm 0.009$	Natural Scene
LDLLDM	1.043 $\pm 0.136$	1.066 $\pm 0.078$	0.964 $\pm 0.138$	0.703 $\pm 0.014$	0.713 $\pm 0.016$	0.682 $\pm 0.026$	
LDLDPA	<i>0.791</i> $\pm 0.048$	<i>0.909</i> $\pm 0.044$	<i>0.714</i> $\pm 0.039$	<i>0.746</i> $\pm 0.009$	<i>0.750</i> $\pm 0.011$	<i>0.729</i> $\pm 0.012$	
LDLFCC	1.052 $\pm 0.080$	1.059 $\pm 0.058$	1.004 $\pm 0.061$	0.698 $\pm 0.007$	0.715 $\pm 0.009$	0.669 $\pm 0.009$	
LDLLRR	0.874 $\pm 0.065$	0.947 $\pm 0.055$	0.813 $\pm 0.040$	0.729 $\pm 0.008$	0.739 $\pm 0.012$	0.705 $\pm 0.007$	
LDLIAR w/ $L_2$	<b>0.668</b> $\pm 0.049$	<b>0.809</b> $\pm 0.036$	0.472 $\pm 0.024$	<b>0.701</b> $\pm 0.021$	<b>0.669</b> $\pm 0.017$	0.749 $\pm 0.010$	Emotion6
LDLLDM	0.688 $\pm 0.039$	0.868 $\pm 0.052$	0.469 $\pm 0.022$	0.694 $\pm 0.017$	0.642 $\pm 0.025$	0.751 $\pm 0.011$	
LDLDPA	0.704 $\pm 0.054$	<i>0.814</i> $\pm 0.038$	0.506 $\pm 0.017$	0.686 $\pm 0.019$	<i>0.664</i> $\pm 0.016$	0.735 $\pm 0.007$	
LDLFCC	0.671 $\pm 0.039$	0.840 $\pm 0.035$	<b>0.462</b> $\pm 0.017$	<i>0.700</i> $\pm 0.018$	0.654 $\pm 0.015$	<b>0.754</b> $\pm 0.008$	
LDLLRR	<i>0.671</i> $\pm 0.039$	0.839 $\pm 0.035$	<i>0.463</i> $\pm 0.017$	0.700 $\pm 0.018$	0.655 $\pm 0.015$	<i>0.754</i> $\pm 0.008$	
LDLIAR w/ $L_2$	<b>0.630</b> $\pm 0.131$	<b>0.761</b> $\pm 0.102$	0.481 $\pm 0.043$	<b>0.718</b> $\pm 0.046$	<b>0.695</b> $\pm 0.027$	0.747 $\pm 0.015$	Art Painting
LDLLDM	0.684 $\pm 0.101$	0.908 $\pm 0.134$	<b>0.466</b> $\pm 0.043$	0.693 $\pm 0.029$	0.613 $\pm 0.034$	<b>0.752</b> $\pm 0.016$	
LDLDPA	0.902 $\pm 0.199$	0.974 $\pm 0.191$	0.697 $\pm 0.160$	0.643 $\pm 0.057$	<i>0.654</i> $\pm 0.043$	0.690 $\pm 0.025$	
LDLFCC	0.676 $\pm 0.115$	0.894 $\pm 0.154$	0.467 $\pm 0.044$	0.697 $\pm 0.038$	0.621 $\pm 0.046$	0.751 $\pm 0.017$	
LDLLRR	<i>0.676</i> $\pm 0.115$	<i>0.893</i> $\pm 0.155$	<i>0.466</i> $\pm 0.044$	<i>0.697</i> $\pm 0.038$	0.621 $\pm 0.046$	<i>0.751</i> $\pm 0.017$	
LDLIAR w/ $L_2$	<b>0.694</b> $\pm 0.056$	<b>0.871</b> $\pm 0.028$	<b>0.349</b> $\pm 0.029$	<b>0.721</b> $\pm 0.015$	<b>0.617</b> $\pm 0.010$	<b>0.838</b> $\pm 0.005$	M2B
LDLLDM	0.866 $\pm 0.181$	1.069 $\pm 0.186$	0.403 $\pm 0.086$	0.690 $\pm 0.035$	0.587 $\pm 0.027$	0.806 $\pm 0.031$	
LDLDPA	0.996 $\pm 0.136$	1.224 $\pm 0.091$	0.487 $\pm 0.049$	0.668 $\pm 0.026$	0.572 $\pm 0.025$	0.777 $\pm 0.017$	
LDLFCC	<i>0.836</i> $\pm 0.144$	<i>1.034</i> $\pm 0.146$	<i>0.377</i> $\pm 0.057$	<i>0.694</i> $\pm 0.028$	<i>0.591</i> $\pm 0.027$	<i>0.810</i> $\pm 0.020$	
LDLLRR	0.875 $\pm 0.160$	1.072 $\pm 0.142$	0.400 $\pm 0.055$	0.687 $\pm 0.031$	0.586 $\pm 0.026$	0.802 $\pm 0.021$	

## B.2 Discussion on Experimental Results

To prevent the  $L_2$  regularization term from confusing our experimental conclusions, we initially exclude it from all comparative algorithms and treat  $L_2$  regularization as a separate baseline in the main text. However, given that  $L_2$  regularization is a widely adopted and empirically effective technique for enhancing model generalization, we in this section incorporate  $L_2$  into all comparison algorithms and comprehensively evaluated their performance across multiple datasets. The performance is shown in Table 2. To distinguish from the main text, we prefix all comparative algorithms with “LDL-” in Table 2, while denoting our proposed algorithm as LDLIAR<sub>w/  $L_2$</sub>  (where “w/  $L_2$ ” represents possessing  $L_2$  regularization). It can be seen that the experimental results and conclusions on the high-entropy datasets are similar to those in the main text; our algorithm achieves superior performance on the low-entropy datasets.

## References

- [1] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.
- [2] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, 1998.
- [3] Tingting Ren, Xiuyi Jia, Weiwei Li, Lei Chen, and Zechao Li. Label distribution learning with label-specific features. In *International Joint Conference on Artificial Intelligence*, pages 3318–3324, 2019.
- [4] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3D facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006.