



# Learning Discriminative Features from Spectrograms using Center Loss for Speech Emotion Recognition

Dongyang Dai<sup>1,2</sup>, Zhiyong Wu<sup>1,2,3</sup>, Runnan Li<sup>1,2</sup>, Xixin Wu<sup>3</sup>, Jia Jia<sup>1,2</sup>, Helen Meng<sup>1,3</sup>

<sup>1</sup>Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China  
<sup>2</sup>Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing, China  
<sup>3</sup>Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong



## 1. Introduction

### Motivation

- Extract valid features from raw data for emotion recognition



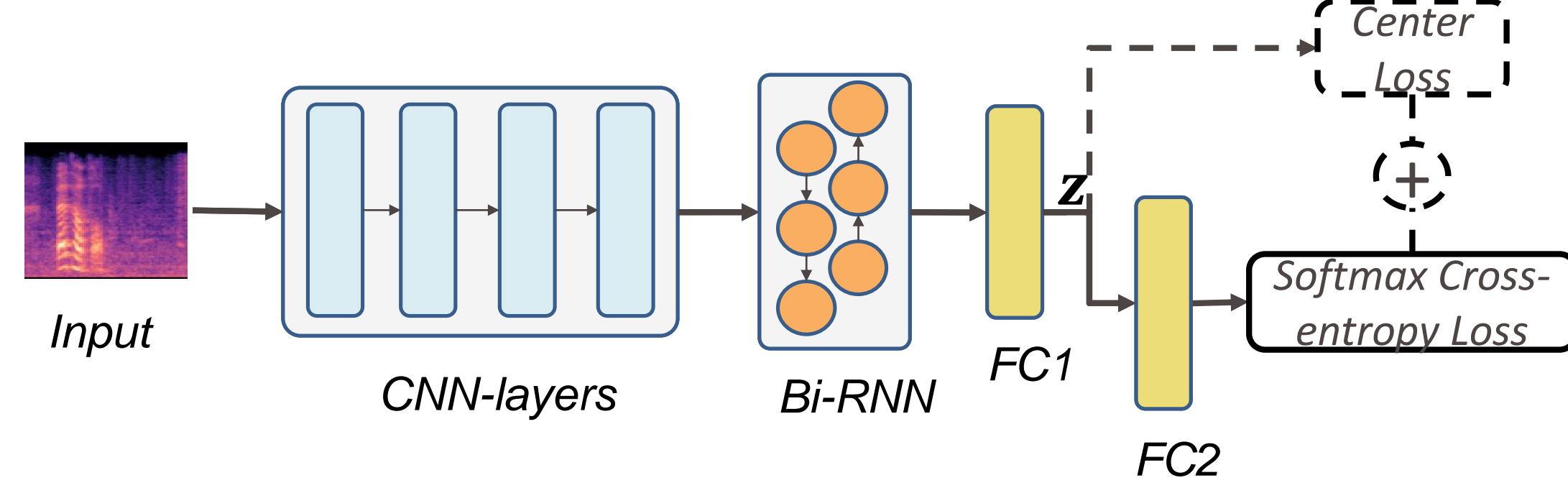
### Challenge

- How to design a suitable model processing directly on raw data
- Emotions are naturally ambiguous

### Contribution

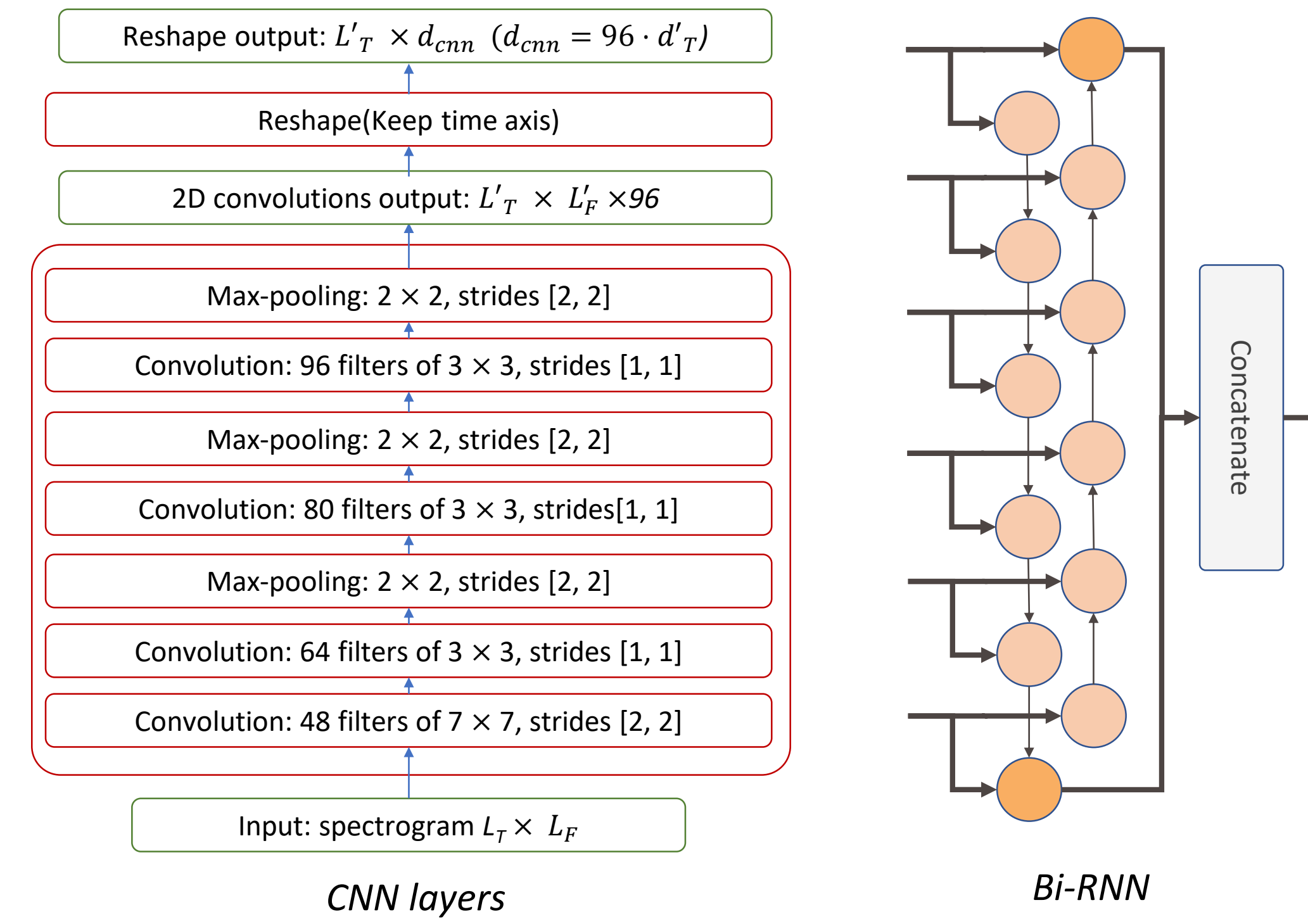
- Extract features and identify emotions directly from spectrograms
- Introduce center loss together with softmax cross-entropy loss in SER task to learn discriminative features
  - Separable inter-class features
  - more compact intra-class features

## 2. Proposed Method



### Model Architecture

- Input: variable length spectrograms (STFT or Mel-spectrograms)
- CNN layers: extract spatial information
- Bi-RNN: compresses the variable length sequence down to a fixed-length vector
- FC1: output  $z \in R^d$  as the learned feature and calculate center loss according to  $z$
- FC2: outputs posterior class probabilities, used to calculate softmax cross-entropy loss
- Softmax Cross-entropy Loss: enables the network to learn separable features
- Center Loss: pulls the features belonging to the same emotion category to their center



### Softmax Cross-entropy Loss

$$L_s = -\frac{1}{\sum_{i=1}^m \omega_{y_i}} \sum_{i=1}^m \omega_{y_i} \log\left(\frac{e^{W_{y_i}^T z_i + b_{y_i}}}{\sum_{j=1}^n e^{W_j^T z_i + b_j}}\right)$$

- $\omega_j$ : in inverse proportion to the sample number of the  $j$ -th class in training set

### Center Loss

$$L_c = \frac{1}{\sum_{i=1}^m \omega_{y_i}} \sum_{i=1}^m \omega_{y_i} \|z_i - c_{y_i}\|^2$$

$$c_j^{t+1} = \begin{cases} (1-\alpha)c_j^t + \alpha \bar{c}_j^t & \sum_{i=1}^m \delta(y_i = j) > 0 \\ c_j^t & \sum_{i=1}^m \delta(y_i = j) = 0 \end{cases}$$

$$\bar{c}_j = \frac{\sum_{i=1}^m \delta(y_i = j) z_i}{\sum_{i=1}^m \delta(y_i = j)}$$

- $L_c$ : center loss
- $c_j$ : the global class center of features corresponding to the  $j$ -th emotion, updated per mini-batch iteration
- $\bar{c}_j$ : the  $j$ -th class center of features from a mini-batch
- $\alpha$ : controls the update rate of  $c_j$

### Joint Loss

$$L = L_s + \lambda L_c$$

- $\lambda$ : trades off center loss against softmax cross-entropy loss.

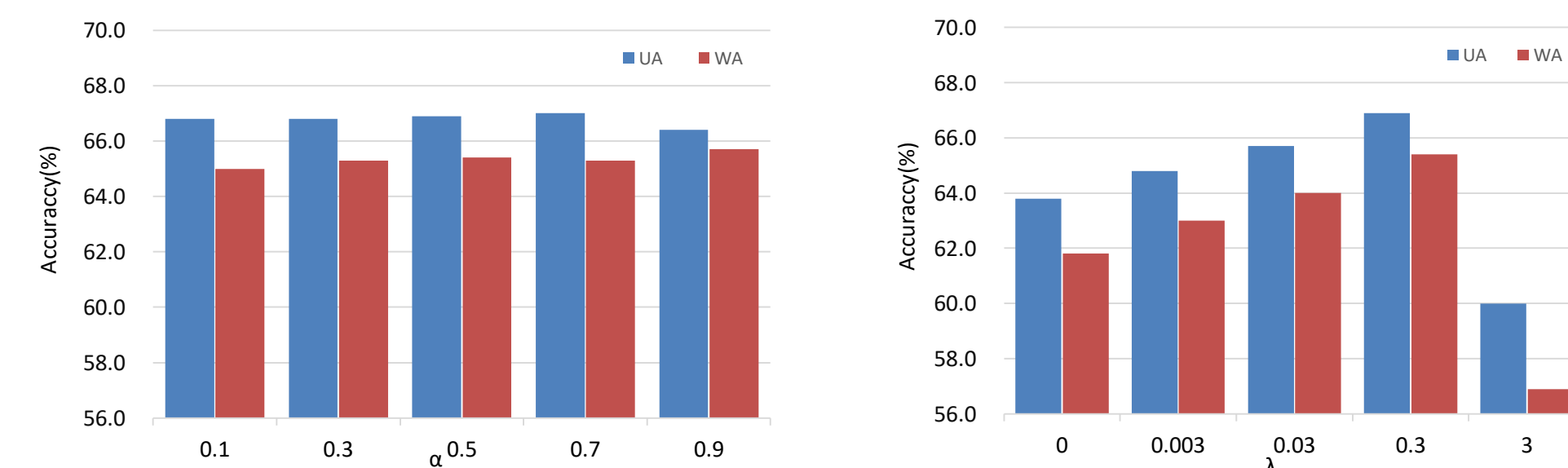
## 3. Experiments and Results

### Experimental Setup

- Data
  - Dataset: IEMOCAP
  - Neutral, angry, happy, sad and excited (merges happy and excited as happy, 5531 utterances)
  - 5 subsets (keep the emotion distribution), 4 subsets for training, half of the last subset as development set and half as test set
- Settings of spectrograms
  - Model input: log scale STFT spectrogram or Mel-spectrogram
  - Hamming window
  - Window size: 40msec
  - Window Shift: 10msec
  - Sample rate: 16KHz
  - DTF length: 1024
  - The number of Mel bands: 128
- Metrics
  - The unweighted accuracy: UA, the mean value of the recall for each class
  - The weighed accuracy: WA, the number of correctly classified samples divided by the total amount of samples

### Experiments

- The effect of hyperparameter  $\alpha$  and  $\lambda$  on Mel-spectrogram
  - (left) fixing  $\lambda = 0.3$ , (right) fixing  $\alpha = 0.5$
  - not sensitive to  $\alpha$
  - can be significantly improved with proper value of  $\lambda$

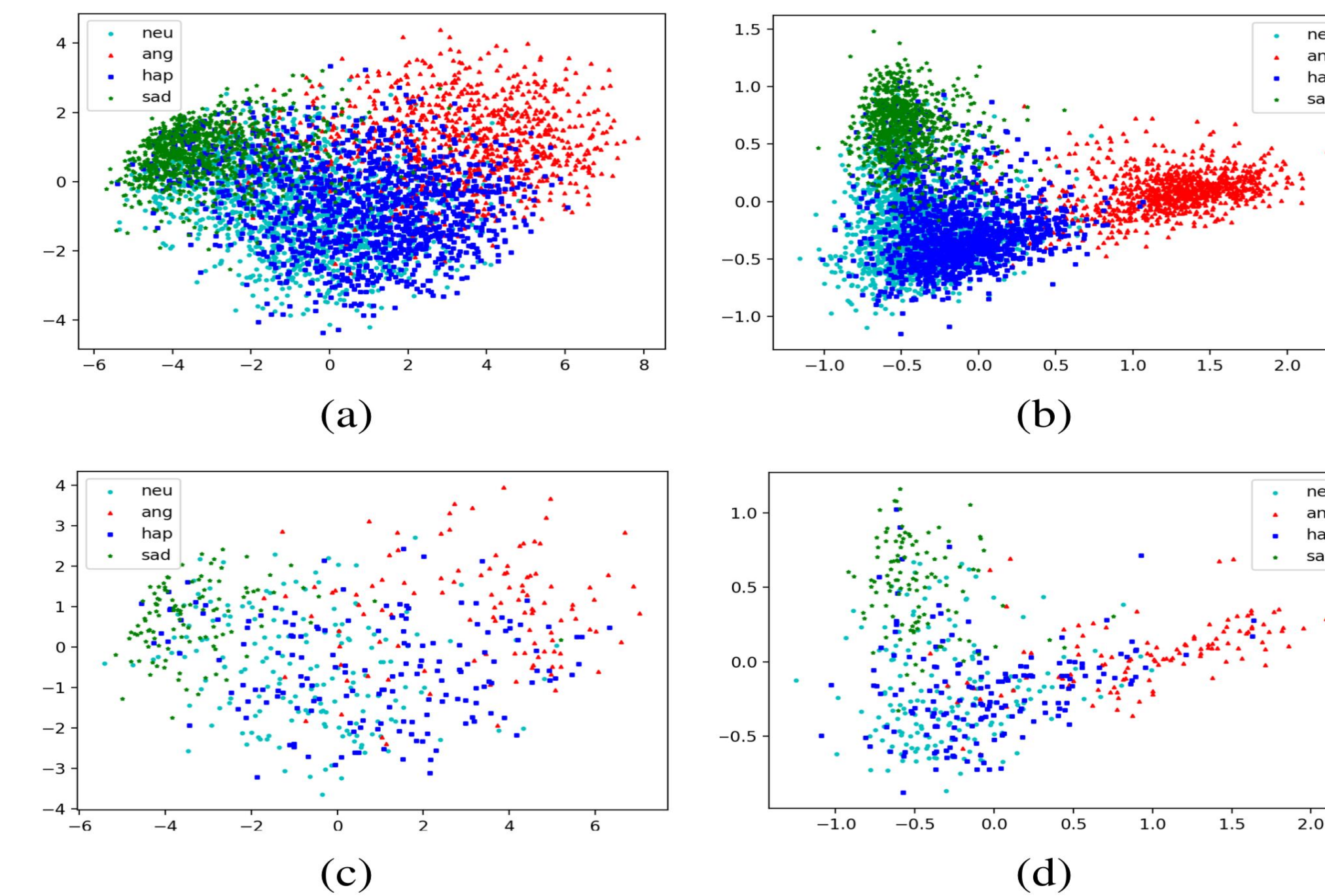


- Experiments with different  $\lambda$  on Mel and STFT

	Setting1	Setting2	Setting3	Setting4
$\lambda, \alpha$	$\lambda=0$	$\lambda=0.3, \alpha=0.5$	$\lambda=0$	$\lambda=0.3, \alpha=0.5$
Input	Mel	Mel	STFT	STFT
The UA and WA on setting 1 ~ setting 4				
	Setting1	Setting2	Setting3	Setting4
UA(%)	63.80	66.86	60.97	65.13
WA(%)	61.83	65.40	58.93	62.96

- Confusion matrix on setting1/setting2/setting3/setting4(%)

- PCA embedding of feature  $z$ 
  - (a) training set on setting 1, (b) training set on setting 2, (c) test set on setting 1, (d) test set on setting 2



	neutral	angry	happy	sad
neutral	57.5/63.7/54.4/57.3	9.5/6.7/9.3/7.3	16.4/16.7/18.5/19.6	16.6/12.7/17.7/15.7
angry	11.9/10.8/12.7/10.3	69.1/70.5/68.1/72.0	15.5/16.7/16.7/15.3	3.5/2.0/2.5/2.2
happy	21.1/21.9/21.6/20.5	16.2/13.1/18.6/16.1	51.1/55.6/47.6/51.8	11.5/9.4/12.2/11.4
sad	13.8/12.8/16.1/12.5	2.6/2.5/3.9/2.8	6.0/7.0/6.2/5.3	77.6/77.7/73.7/79.3

## 4. Conclusion

### Conclusion

- Introducing center loss with proper  $\lambda$  could effectively improve the SER performance on both STFT spectrogram and Mel-spectrogram input
- Mel-spectrogram input, reducing the dimension based on human hearing characteristics, over performs STFT spectrogram input
- The 2-D PCA embedding illustrates the discriminative power when using center loss, which enables the neural network to learn more effective features for SER

## 5. Acknowledgment

- This work is supported by National Natural Science Foundation of China (NSFC) (61433018, 61375027), joint research fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N CUHK404/15) and National Social Science Foundation of China (13&ZD189)