The Supplementary Materials for paper: CTL-MTNet: A Novel Mixed Task Net Based on CapsNet and Transfer Learning for Single-Corpus and Cross-Corpus Speech Emotion Recognition

A Dataset

In the experiments, in order to compare with the state-of-the-art methods, the algorithm is tested on four datasets, including the Institute of Automation of Chinese Academy of Sciences (CASIA) [Tao *et al.*, 2008], Berlin Emotional Database (EmoDB)[Burkhardt *et al.*, 2005], Surrey Audio-Visual Expressed Emotion Database (SAVEE)[Jackson and Haq, 2014], and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[Livingstone and Russo, 2018]. The basic information of the four datasets is shown in Table S1, and the distribution information of the sentiment is given in Table S2.

B Evaluation metrics

The weighted average recall (WAR) and unweighted average recall (UAR) are adopted for performance comparisons in this paper, which are defined as follows:

$$WAR = \sum_{\alpha=1}^{E} \frac{\sum_{\beta=1}^{k}(TP_{\alpha}^{\beta} + TN_{\alpha}^{\beta})}{\sum_{\alpha=1}^{E}\sum_{\beta=1}^{k}(TP_{\alpha}^{\beta} + TN_{\alpha}^{\beta} + FP_{\alpha}^{\beta} + FN_{\alpha}^{\beta})} \times \frac{\sum_{\beta=1}^{k} TP_{\alpha}^{\beta}}{\sum_{\beta=1}^{k}(TP_{\alpha}^{\beta} + FN_{\alpha}^{\beta})} \tag{1}$$

$$UAR = \frac{1}{E}\sum_{\alpha=1}^{E} \frac{\sum_{\beta=1}^{k} TP_{\alpha}^{\beta}}{\sum_{\beta=1}^{k}(TP_{\alpha}^{\beta} + FN_{\alpha}^{\beta})} \tag{2}$$

where E denotes the number of emotion classes, k represents the number of speech signals. $TP_{\alpha}^{\beta}, TN_{\alpha}^{\beta}, FP_{\alpha}^{\beta}$ and FN_{α}^{β} represent the true positive, true negative, false positive, and false negative values of class α for speech signal β respectively.

C The experiment settings

In this experiment, 39-dimensional MFCCs are extracted from the Librosa toolbox [McFee *et al.*, 2015] to serve as inputs with the frame shift of 0.0125 s and the frame length of 0.05 s.

The proposed algorithm is implemented with TensorFlow. In the single-corpus task, the batch size is set to 64. The CPAC model is optimized by using Adam algorithm [Kingma and Ba, 2014] with an initial learning rate α of 1.0×10^{-3}, exponential decay rates β_1 = 0.932, β_2 = 0.975, and the weight decay ϵ is set to 1.0×10^{-8}.

In the cross-corpus task, the batch size is set to 512. The CAAM model is optimized by using Adam algorithm [Kingma and Ba, 2014] with an initial learning rate α of 1.0×10^{-3}, exponential decay rates β_1 = 0.935, β_2 = 0.975, and the weight decay ϵ is set to 1.0×10^{-8}. The gradient reversal layer (GRL) [Ganin and Lempitsky, 2015] is employed to train ψ to minimize the MDD loss function. Furthermore, we utilize data augmentation on the source-corpus speech signals for the cross-corpus task to ensure that the training and test data volumes are approximately the same for the case where the source-corpus data volume is significantly smaller than the target-corpus data, such as in the case of using EMODB and CASIA datasets as the source and target corpora respectively.

D Detailed results in the single-corpus task

Due to space limits, the confusion matrix of our proposed algorithm on a single corpus is not given in the main text. The confusion matrices for the highest recognition accuracies obtained by the proposed algorithm on four databases are shown in Figures S1(a)- S1(d). For example, on the CASIA data, the Fear and Sad emotions are the hardest to be separated. These results are also reflected in Fig.2(d) of the main text. After visualization by t-SNE, the Fear and Sad emotions are most closely clustered, leading to the high misidentification. As depicted in Figures S2 and S3, it is obvious that different classes are clustered with clear boundaries in all datasets. It is confirmed that the proposed CPAC learns the emotional features with well discrimination.

E Detailed results in the cross-corpus task

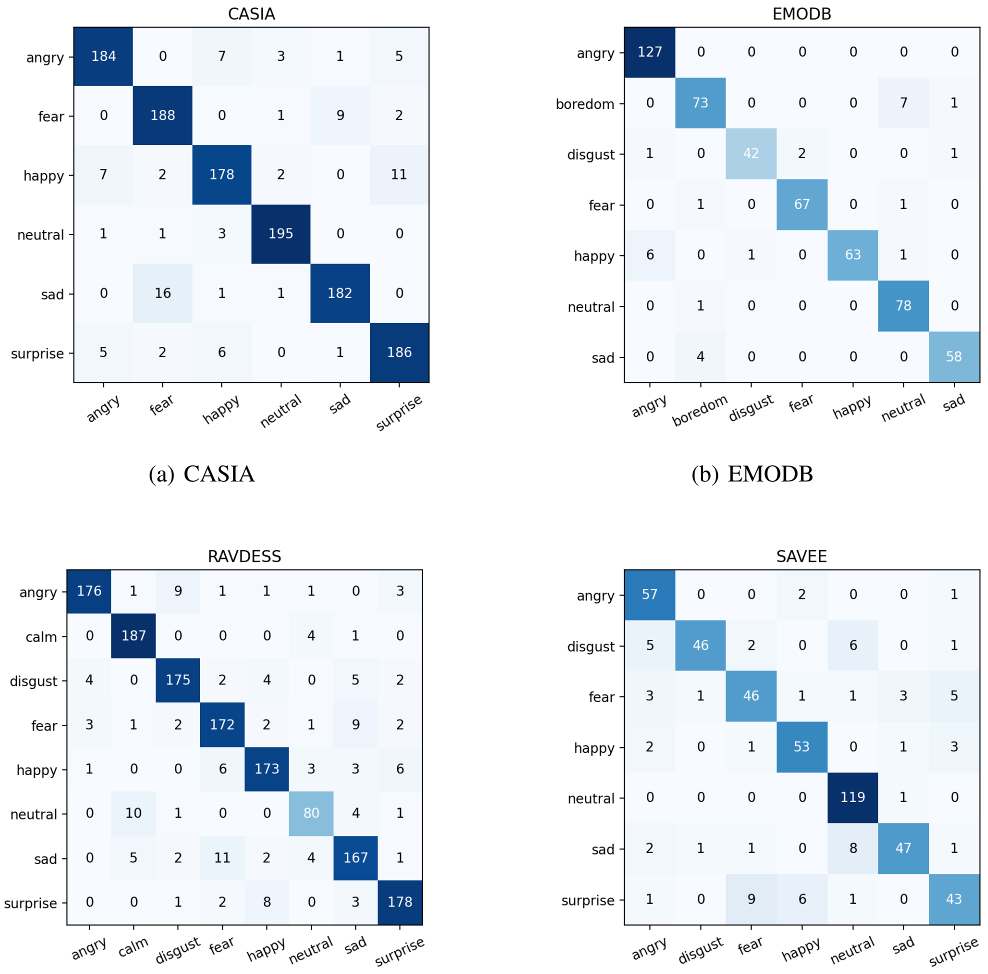
The results on the RAVDESS dataset obtained from our method and those with two representative unsupervised learning methods, CDAN [Long *et al.*, 2017] and DANN [Abdelwahab and Busso, 2018] are compared. The results in Table S3 show that our method outperforms these methods on the RAVDESS dataset. The average unweighted average recall (UAR) and weighted average recall (WAR) obtained by our method are 28.16% and 30.10%, respectively. The proposed

Dataset	Language	Actors	Numbers	Emotion	Sampling rate
RAVDESS	English	24(12 males, 12 females)	1440	8 emotions. Happy, Sad, Angry, Calm, Fear, Neutral, Disgust, Surprise	48 KHz
EMODB	German	10(5 males, 5 females)	535	7 emotions. Happy, Sad, Angry, Boredom, Fear, Neutral, Disgust	16 KHz
SAVEE	English	4 males	480	7 emotions. Neutral, Anger, Disgust, Fear, Happiness, Sadness, Surprise	44.1 KHz
CASIA	Chinese	4(2 males, 2 females)	1200	6 emotions. Angry, Fear, Happy, Neutral, Sad, Surprise	22.05KHz

Table S1: The detailed information of speech emotion datasets

Name	Neutral	Anger	Disgust	Fear	Happy	Sad	Surprise	Boredom	Calm
RAVDESS	96	192	192	192	192	192	192	–	192
EMODB	79	127	46	69	71	62	–	81	–
SAVEE	120	60	60	60	60	60	60	–	–
CASIA	200	200	–	200	200	200	200	–	–

Table S2: The details of data distributions in four datasets



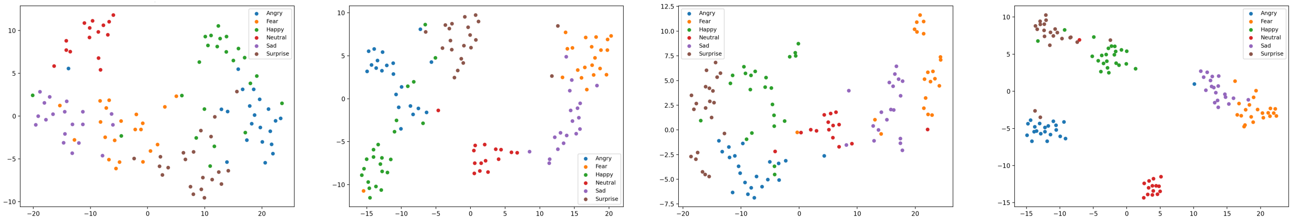
(a) CASIA

(b) EMODB

(c) RAVDESS

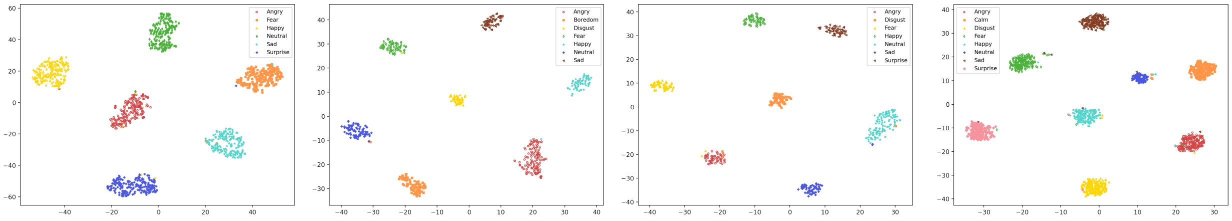
(d) SAVEE

Figure S1: The confusion matrices of the single-corpus task obtained by CTL-MTNet on the CASIA, EMODB, SAVEE and RAVDESS datasets.



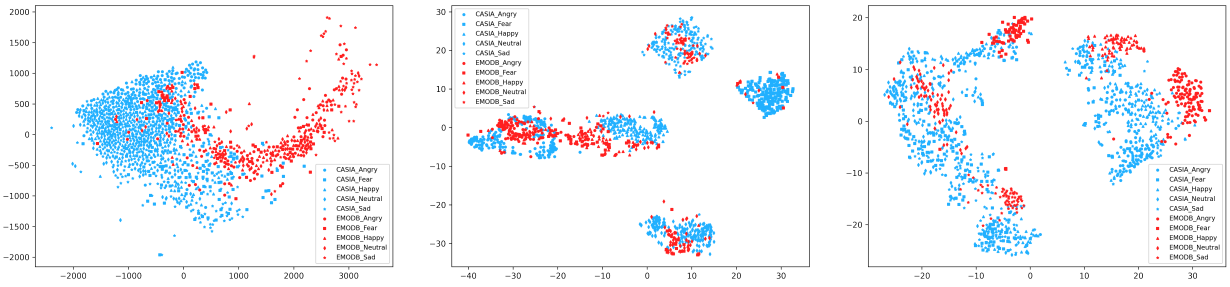
(a) Algorithm 1 (b) Algorithm 2 (c) Algorithm 3 (d) CPAC

Figure S2: t-SNE visualization of the high-level features tested with the ablation algorithms on CASIA



(a) CASIA (b) EMODB (c) SAVEE (d) RAVDESS

Figure S3: t-SNE visualization of the high-level features trained with the proposed algorithm with the CASIA, EMODB, SAVEE and RAVDESS datasets.



(a) CASIA and EMODB without CAAM (b) CASIA transfer to EMODB (c) EMODB transfer to CASIA

Figure S4: t-SNE visualization of the original MFCC features and domain-invariant emotion features trained with the proposed algorithm for the CASIA and EMODB datasets.

Source Corpus		CASIA	RAVDESS	EMODB	RAVDESS	SAVEE	RAVDESS	Average
Target Corpus		RAVDESS	CASIA	RAVDESS	EMODB	RAVDESS	SAVEE	
CDAN	UAR	24.06	22.92	24.58	21.05	21.04	28.17	23.64
	WAR	25.00	24.54	27.20	31.86	23.38	25.00	26.16
DANN	UAR	23.75	20.50	25.31	20.94	20.10	26.83	22.91
	WAR	25.12	20.50	27.89	31.86	22.34	25.56	25.55
CAAM	UAR	**25.31**	**25.90**	**32.08**	**28.62**	**24.38**	**32.67**	**28.16**
	WAR	**28.01**	**25.90**	**35.65**	**37.10**	**26.74**	**27.22**	**30.10**

Table S3: The performance comparisons on the RAVDESS dataset.

Method	MDD	Source Corpus	CASIA	RAVDESS	EMODB	RAVDESS	SAVEE	RAVDESS	Average
		Target Corpus	RAVDESS	CASIA	RAVDESS	EMODB	RAVDESS	SAVEE	
Algorithm 4	No	UAR	22.92	19.90	23.44	20.32	19.37	26.83	22.13
		WAR	24.31	19.90	26.04	31.37	21.41	23.89	24.49
CAAM	Yes	UAR	**25.31**	**25.90**	**32.08**	**28.62**	**24.38**	**32.67**	**28.16**
		WAR	**28.01**	**25.90**	**35.65**	**37.10**	**26.74**	**27.22**	**30.10**

Table S4: The ablation study on the RAVDESS dataset.

Source Corpus		CASIA	EMODB	CASIA	SAVEE	EMODB	SAVEE	Average
Target Corpus		EMODB	CASIA	SAVEE	CASIA	SAVEE	CASIA	
CDAN+SVM	UAR	49.22	41.15	41.70	36.95	28.11	45.06	40.36
	WAR	53.66	39.50	52.78	35.50	37.50	51.22	45.03
DANN+SVM	UAR	32.88	35.64	20.00	30.56	29.26	28.89	29.54
	WAR	42.68	31.50	37.50	26.50	44.44	39.02	36.94
NMFTSL+SVM	UAR	61.84	44.50	51.04	47.00	43.47	50.96	49.80
	WAR	61.76	44.50	59.00	47.00	48.10	53.96	52.39
CAAM+SVM	UAR	**66.83**	**56.50**	**63.12**	**61.58**	**50.51**	**72.30**	**61.81**
	WAR	**69.51**	**56.50**	**70.83**	**61.50**	**55.56**	**71.95**	**64.31**

Table S5: The performance comparisons on the CASIA, EMODB, and SAVEE datasets.

method achieves +4.52% and +3.94% relative improvements for the average UAR and WAR compared to those obtained in [Long *et al.*, 2017], +5.25% and +4.55% compared to those obtained in [Abdelwahab and Busso, 2018]. Such great performance on cross-corpus datasets demonstrates that our method can be generalized to various speakers and different kinds of cross-language environments.

Moreover, the ablation experiments with Algorithm 4 that removes the domain adaptation method for the target corpus and trains only on the source verify the importance of the component added by the proposed method. As is shown in Table 4, applying MDD can further gain +6.03% and +5.61% relative improvements for the average UAR and WAR compared to Algorithm 4.

After visualization by t-SNE, the visualized embedding of common sentiment representations in Fig.S4(b) and Fig.S4(c) shows that the proposed CAAM yields superior performance in feature alignment between source corpus and target corpus. It is confirmed that the proposed CAAM can balance well emotional discrimination and sentiment feature alignment.

F The supervised learning-based results for the cross-corpus task

The current mainstream methods on the cross-corpus task can be divided into two categories: supervised learning-based and unsupervised learning-based [Zhang *et al.*, 2021]. In the main text, we have shown our results on unsupervised learning methods. In order to further compare to the results obtained by the state-of-the-art methods on the field of supervised learning such as NMFTSL [Luo and Han, 2020], we first run these methods to extract each domain-invariant feature representation for the source and target corpus for CDAN, DANN, and NMFTSL, and then employ a linear SVM to evaluate their performances.

The results in Table S5 show that our method outperforms those three methods. The average UAR and WAR obtained by our method are 61.81% and 64.31%, respectively, achieving +21.45% and +19.28% relative improvements of the average UAR and WAR compared to [Long *et al.*, 2017], +32.27% and +27.37% compared to [Abdelwahab and Busso, 2018], and +12.01% and +11.92% compared to [Luo and Han, 2020]. Such superior performance on the cross-corpus task demonstrates that our method can be easily transferred to supervised learning with the highest performance.

References

[Abdelwahab and Busso, 2018] Mohammed Abdelwahab and Carlos Busso. Domain adversarial for acoustic emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2423–2435, 2018.

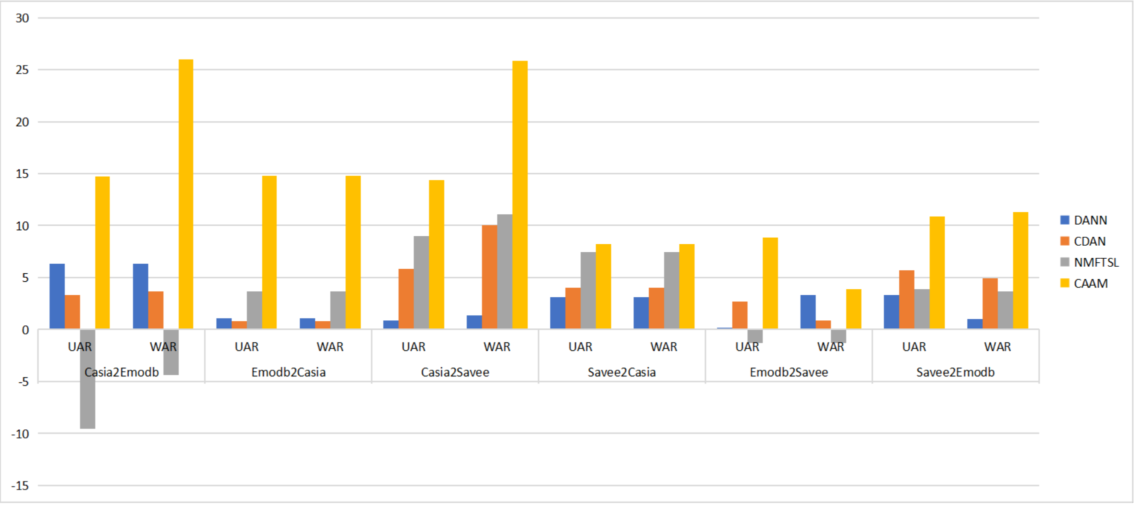


Figure S5: Comparison of the results of different databases and methods in cross-corpus task.

[Burkhardt *et al.*, 2005] Felix Burkhardt, Astrid Paeschke, M. Rolfes, et al. A database of german emotional speech. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, 2005.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[Jackson and Haq, 2014] Philip Jackson and SJUoSG Haq. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*, 2014.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Livingstone and Russo, 2018] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[Long *et al.*, 2017] Mingsheng Long, Zhangjie Cao, Jianmin Wang, et al. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017.

[Luo and Han, 2020] Hui Luo and Jiqing Han. Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2047–2060, 2020.

[McFee *et al.*, 2015] Brian McFee, Colin Raffel, Dawen Liang, et al. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer, 2015.

[Tao *et al.*, 2008] Jianhua Tao, Fangzhou Liu, Meng Zhang, and Huibin Jia. Design of speech corpus for mandarin text to speech. In *The Blizzard Challenge 2008 workshop*, 2008.

[Zhang *et al.*, 2021] Shiqing Zhang, Ruixin Liu, et al. Deep cross-corpus speech emotion recognition: Recent advances and perspectives. *Frontiers in Neurorobotics*, 15, 2021.