INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

SOT: Self-supervised Learning-Assisted Optimal Transport for Unsupervised Adaptive Speech Emotion Recognition

Ruiteng Zhang[1], *Jianguo Wei*[1,2], *Xugang Lu*[3], *Yongwei Li*[4], *Junhai Xu*[1], *Di Jin*[1], *Jianhua Tao*[5]

[1]College of Intelligence and Computing, Tianjin University, China
[2]Computer College, Qinghai Nationalities University, China [3]National Institute of Information and Communications Technology, Japan [4]Institute of Automation, Chinese Academy of Sciences, China
[5]Department of Automation, Tsinghua University, China

Abstract

In cross-domain speech emotion recognition (SER), reducing the global probability distribution distance (GPDD) between different domains plays a crucial role in unsupervised domain adaptation (UDA), which can be naturally measured by optimal transport (OT). However, owing to the large intra-variations of emotion categories, samples distributed in overlap may induce negative transports. Moreover, OT only considers the GPDD and therefore cannot efficiently transport hard-discriminative samples without utilizing the local structures from intra-class distributions. We propose a self-supervised learning (SSL)-assisted optimal transport (SOT) algorithm for cross-domain SER. First, we regularized OT's transport coupling to mitigate negative transports; then, we designed an SSL module to emphasize local intra-class structure to assist OT in capturing those nontransferable acknowledge. Cross-domain SER experimental results showed that SOT dramatically outperformed state-of-the-art UDAs.

Index Terms: speech emotion recognition, unsupervised domain adaptation, optimal transport, self-supervised learning

1. Introduction

Speech emotion recognition (SER) has great potential value in human–computer interaction [1, 2, 3, 4, 5, 6]. However, SER systems have difficulty maintaining robustness in the Out-of-Domain (OoD) problem [7] because the testing speech data are diverse and complex in real-world applications. This leads to discrepancies in significant probability distributions between training and testing conditions [8].

To address this issue, researchers have proposed various domain adaptation (DA) algorithms for cross-domain SER, which aim to mitigate domain differences between training and testing data [8]. Such algorithms can be broadly categorized into supervised DA (SDA) or unsupervised DA (UDA) algorithms based on the availability of label information in the target data. Discriminative subspace alignment algorithms [9, 10] have been employed in SDA to mitigate domain differences in SER, and generative target sample methods [11, 12] have also received attention with the development of generative adversarial networks (GANs). However, in practical scenarios, the target emotion data are often completely unlabeled, leading to UDA being more practical. Most UDA strategies for SER are currently developed based on an adversarial training strategy [13]. Such policies aim to guide emotional classifiers that cannot to distinguish the latent domain distribution of source or target utterances, thereby mitigating domain mismatches [7, 14, 15]. Additionally, some studies have attempted to calibrate domain information using intra-class relations [16, 17]. However, as various speakers express emotions differently and in different environments, the boundaries between different emotion categories are unclear [18, 19]. Consequently, the above methods often align domain information and sacrifice intrinsic emotion information simultaneously, especially when using adversarial training methods. In contrast, the global probability distribution distance (GPDD) measures the global relationship between two probability distributions, thereby facilitating the transfer of the source probability distribution to the target one while maintaining the intrinsic emotional representations. Unfortunately, current UDAs in SER do not consider GPDDs and may even undermine them [18, 20]. The aforementioned limitations often reduce the robustness of existing cross-domain SER UDAs in real-world scenarios, particularly in cases involving language, background noise, and utterance duration mismatches.

Concerning the global probability distribution distance alignment, optimal transport (OT) [21] provides natural mathematic formulations and has been intensively applied in the machine learning field [22]. OT converts one probability distribution shape to another shape with the least effort based on the geometry difference. However, the correlations between different emotions are related to the properties of human emotion expression and perception [7], resulting in unclear boundaries of different target emotion categories [19]. The distribution of these samples in overlapped regions between different emotions often leads to OT computing massive negative transports (i.e., negatively transporting the speech samples from different emotional categories' samples), dramatically reducing adaptation performance. Moreover, OT only considers the GPDD without utilizing the local structures from intra-class distributions, wasting the emotional representations in the local intra-class structure, especially in hard-discriminative samples. However, related work modeling these wasted emotional representations is still limited in cross-domain SER.

To solve these problems, we propose an unsupervised adaptation framework for cross-domain SER, known as self-supervised learning-assisted optimal transport (SOT). Our basic assumption is shown in Fig. 1, which is used to measure the global probability distribution distance between different domains effectively; in addition, the local intra-class structure information in target emotional samples is explored. Specifically, a margin regularization was designed into OT to inhibit the negative transports (the red dash-line in Fig. 1), achieving more effective measures for GPDD of different domains. Moreover, we propose the self-supervised learning assistance exploration (SSLAE) module to emphasize the local intra-class structure to assist OT in capturing those emotional representations, especially in nontransferable samples. The contributions of this paper are listed as follows:

1. We propose margin regularized OT (MOT) for SER, which mitigates negative transports, achieving effective mea-

10.21437/Interspeech.2023-1360

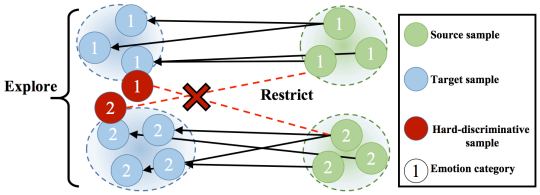

Figure 1: *The basic assumption of our proposal. Black line: correct transports; red dashed line: negative transports. In this figure, the negative transports are restricted to achieve effective measures to the GPDD between different domains. The nontransferable local structures from intra-class distributions are explored by SSLAE.*

sure of the GPDD between different domains (Section 2.2.1);

2. We propose SSLAE to assist OT in capturing emotional features within local intra-class structures, rather than measuring the GPDD solely (Section 2.2.2);

3. We propose a UDA with the siamese framework for SER, called SOT. We conducted three SER experiments, with a mismatch in language, background noise, and audio durations, which were designed using three well-known SER corpora IEMOCAP [23], EMODB [24], and CREMA-D [25]. Experimental results show that the proposed SOT achieved state-of-the-art performance in the OoD SERs (Section 3).

2. Proposed Methods

The proposed SOT is a siamese framework comprising an emotion classifier, a margin regularized OT (MOT) alignment module, and a self-supervised learning-assistance exploration (SSLAE) module. Fig. 2 illustrates the architecture of the proposed SOT. In the training stage, our UDA's emotion encoders share weights, and during inference, it only retains one encoder to produce robust emotion predictions. In Fig. 2, a supervised emotion classifier aims to learn emotional knowledge in the source data. Moreover, to mitigate domain discrepancies, we jointly train the emotion classifier with the MOT and SSLAE branches, emphasizing the GPDD and local intra-class structure, respectively. The MOT alignment module calculates the transport cost of the source and target domains by the joint probability distributions (i.e., marginal and conditional distributions), based on which a margin regularization is proposed to filter the negative transports between different emotional samples. More importantly, the SSLAE module is proposed to assist OT in exploring the local intra-class structure, especially in hard-discriminative samples (red samples in Fig. 1). The SSLAE can learn meaningful domain variables and emotional representations using only positive sample pairs with a siamese architecture. Conversely, existing OTs overlook these structural representations. Next, we introduce the basic theory of domain adaptation in SER and the proposed SOT.

2.1. Domain adaptation for SER

Given a source domain data set $D^s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1,..,N}$, and a target domain data set $D^t = \{(\mathbf{x}_i^t, \mathbf{y}_i^t)\}_{i=1,..,M}$ (in the UDA task, D^t is un-labeled). Due to domain changes (e.g., language mismatch), emotions' probability distributions under different domains are changed significantly, i.e., $p^s(\mathbf{x}, \mathbf{y}) \neq p^t(\mathbf{x}, \mathbf{y})$. The purpose of domain adaptation is to mitigate such discrepancies and train an emotion category classifier that can accurately classify speech samples in the target domain.

2.2. Proposed self-supervised learning-assisted OT

The proposed UDA framework follows the deep speech emotion recognition model, which is defined below:

$$\hat{\mathbf{y}} = f(\mathbf{x}; \theta_{G_f}, \theta_{G_y}) = G_y \circ G_f(\mathbf{x}); \quad \mathbf{e} = G_f(\mathbf{x}), \tag{1}$$

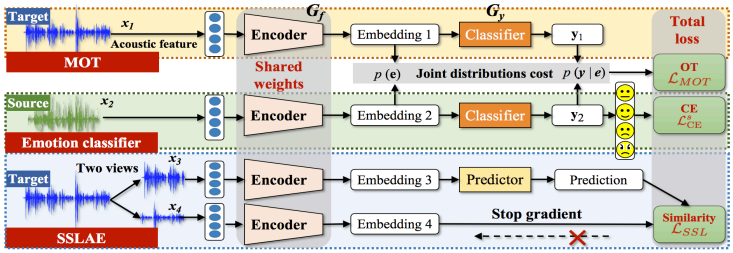

Figure 2: *The details of the proposed SOT adaptation framework.*

where $G_f(\cdot)$ and $G_y(\cdot)$ are the embedding extractor and classifier transforms with parameter sets θ_{G_f} and θ_{G_y}, respectively. $\mathbf{e}$ is the embedding, and $\hat{\mathbf{y}}$ denotes the estimated emotion label. Fig. 2 shows the proposed SOT. This includes two unsupervised domain adaptation modules, MOT (Section 2.2.1) and SSLAE (Section 2.2.2), and a supervised emotion classifier (Section 2.2.2). Here, we describe them in detail.

2.2.1. Margin regularized OT

Optimal transport (OT) [21] is an effective measure to maintain intrinsic emotional features during UDA by calculating the GPDD based on the geometric shape of the distributions. Mathematically, the OT distance between the source p^s and target p^t distributions is defined as follows:

$$d_{\mathrm{OT}}(p^s, p^t) \triangleq \min_{\gamma \in \prod(p^s, p^t)} \sum_{i,j} C(\mathbf{z}_i^s, \mathbf{z}_j^t)\gamma(\mathbf{z}_i^s, \mathbf{z}_j^t), \tag{2}$$

where γ is the transport coupling between the two domains, and $C(\mathbf{z}_i^s, \mathbf{z}_j^t)$ is the transport cost between examples $\mathbf{z}_i^s$ and $\mathbf{z}_j^t$ that are sampled from distributions p^s and p^t, respectively.

Through Eq. (2), we design an unsupervised domain adaptation framework for SER, as shown in the yellow block of Fig. 2. Owing to the unclear boundaries of different emotion categories in an unknown target domain [7], both the joint distributions (i.e., marginal distribution $p(\mathbf{e})$ and conditional distribution $p(\mathbf{y}|\mathbf{e})$) are considered to calculate the transport cost matrix. The distribution difference across domains can be aligned by minimizing the following joint cost function:

$$C_{\mathrm{adpt}}(\mathbf{x}^s, \mathbf{x}^t) = \alpha C_{\mathrm{fea}}(G_f(\mathbf{x}^s), G_f(\mathbf{x}^t)) + \beta C_{\mathrm{cls}}(f(\mathbf{x}^s), f(\mathbf{x}^t)), \tag{3}$$

where $C_{\mathrm{fea}}(\cdot)$ and $C_{\mathrm{cls}}(\cdot)$ measure the distribution cost of marginal and conditional predictions, respectively. Euclidian distance is used to compute $C_{\mathrm{fea}}(\cdot)$ and $C_{\mathrm{cls}}(\cdot)$ [26]. α and β are the weighting coefficients for these two costs, respectively. Eq. (3) measures the distribution discrepancy across domains, and correspondingly, the adaptation loss function based on Eq. (2) is defined as follows:

$$\mathcal{L}_{\mathrm{OT}}(p^s, p^t) = \min_{\gamma \in \prod(p^s, p^t)} \sum_{i,j} C_{\mathrm{adpt}}(\mathbf{x}_i^s, \mathbf{x}_j^t)\gamma(\mathbf{x}_i^s, \mathbf{x}_j^t), \tag{4}$$

where $\mathbf{x}_i^s$ and $\mathbf{x}_j^t$ are the source and target samples (i and j are sample indexes), respectively.

Negative transfer in OT occurs due to the probabilistic aliasing of hard-discriminative samples, which are often located in regions where different emotions overlap (the red dashed line in Fig. 1). To avoid negative transports, it is necessary to limit the transfer of target samples that are too distant in probability from the source samples. Based on the above consideration, we propose a margin regularization to filter these couplings with excessive costs, by which we set a threshold b of the transport cost to obtain the admissible couplings as:

$$w_{i,j} = \begin{cases} 1, & C(\mathbf{x}_i^s, \mathbf{x}_j^t) \leq b \\ 0, & C(\mathbf{x}_i^s, \mathbf{x}_j^t) > b. \end{cases} \tag{5}$$

In this equation, if the transport cost is smaller than b, the coupling between the two samples is allowed; otherwise, the cou-

Table 1: *Statistics of the per-emotion recognition corpus in experiments.*

Corpus	Speakers	Neutral	Happy	Angry	Sad
IEMOCAP [23]	10	1,708	1,636	1,103	1,084
EMODB [24]	10	79	71	127	62
CREMA-D [25]	91	1,087	1,271	1,271	1,270

pling is discarded. However, this hard controlling of the couplings may pose a risk in discarding admissible couplings or accepting non-admissible couplings. Therefore, we designed a soft weighting on the coupling to smooth the regularization, which is defined below:

$$\tilde{w}_{i,j} = \sigma\left(-scale * \left(C(\mathbf{x}_i^s, \mathbf{x}_j^t) - b\right)\right), \quad (6)$$

where $\sigma(\cdot)$ is a sigmoid function, and $scale$ is a scaling parameter. In this work, we set b=1, and scale parameter $scale$=5.0. Finally, the loss function defined in Eq. (4) is changed to:

$$\mathcal{L}_{\mathrm{MOT}}(p^s, p^t) = \min_{\gamma \in \prod(p^s, p^t)} \sum_{i,j} C(\mathbf{x}_i^s, \mathbf{x}_j^t)\gamma(\mathbf{x}_i^s, \mathbf{x}_j^t)\tilde{w}_{i,j}. \quad (7)$$

In this formulation, the soft weighting $\tilde{w}_{i,j}$, defined in Eq. (6), can be regarded as a margin regularization of coupling and can be explicitly added to fulfill the function of MOT.

2.2.2. SSL-assisted exploration

In Section 2.2.1, we regularized the transport coupling of OT to mitigate the negative transports. Although OT can effectively measure the GPDD between different domains, it cannot capture the local structures from intra-class distributions, particularly in hard-discriminative samples. To explore these representations in the target domain, inspired by SimSiam [27], we propose the SSL assisted exploration (SSLAE) module to explore the latent emotion representations involved in the target samples that cannot be utilized by OT. The architecture of the SSLAE is depicted in the blue block of Fig. 2. In this figure, a target utterance is randomly cropped to two views: $\mathbf{x}_3$ and $\mathbf{x}_4$ (3s duration). Then, a siamese network [27] where one branch with stoped-gradient utilizes $\mathbf{x}_3$ and $\mathbf{x}_4$ to learn the discriminative class structure and latent domain information on the target data. The predictor include two fully-connection layers with the ReLU activation function. In the training stage, SSLAE attempts to bring the probability distance of prediction $\mathbf{p}$ close to $\mathbf{e}_4$ by minimizing the negative cosine similarity, which is defined as follows:

$$\mathcal{D}(\mathbf{v}, \mathbf{u}) = -\frac{\mathbf{v}}{||\mathbf{v}||_2} \cdot \frac{\mathbf{u}}{||\mathbf{u}||_2}, \quad (8)$$

where $\mathbf{v}$ and $\mathbf{u}$ are the compared vectors. The loss function for the part with the self-supervised assistance module is as follows:

$$\mathcal{L}_{\mathrm{SSL}} = \frac{\mathcal{D}(\mathbf{p}_3, stopgrad(\mathbf{e}_4))}{2} + \frac{\mathcal{D}(\mathbf{p}_4, stopgrad(\mathbf{e}_3))}{2}, \quad (9)$$

where $stopgrad$ denotes the stop-gradient propagation for the network. In the training stage, the encoder on $\mathbf{x}_3$ receives the stop gradient from $\mathbf{e}_3$ in the first term but receives gradients from $\mathbf{p}_3$ in the second term (and vice versa for $\mathbf{x}_4$).

In addition to adaptation loss functions, a classification loss is adopted in our UDA. Here, the cross entropy (CE) loss with Softmax $\mathcal{L}_{\mathrm{CE}}^s(\cdot)$ is used. Therefore, the total loss includes the source emotion classification loss and two adaptation losses:

$$\mathcal{L}_{\mathrm{Total}} = \min_{\gamma, \theta_{G_f}, \theta_{G_y}} \left(\mathcal{L}_{\mathrm{CE}}^s(\cdot) + \eta\mathcal{L}_{\mathrm{MOT}}(\cdot) + \mu\mathcal{L}_{\mathrm{SSL}}(\cdot)\right), \quad (10)$$

where η and μ are the weighting coefficients.

In sum, three modules should be optimized in SOT: the emotion classifier (the green block in Fig. 2); MOT-based adaptation loss (Eq. (7), the yellow block in Fig. 2); and SSL exploration loss (Eq. (9), the blue block in Fig. 2).

Table 2: *Baseline performance in the source data (UA and WA).*

System	Years	UA (%)	WA (%)
HuBERT [5]	2021	-	68.90
Residual-BLSTM [31]	2022	70.11	69.31
MHSA-FACA [32]	2022	72.01	72.83
Ours	2023	69.73	70.01

3. Experiments

3.1. Experimental configurations

Details of the source and target domains: To investigate UDAs and verify our proposed hypotheses, we analyzed three common OoD problems in SER and utilized speech data from four emotional categories (*neutral*, *happy*, *angry*, and *sad*) in the audio modality, consistent with previous studies. Table 1 presents the statistics of the speech corpora used.

• Task 1: **Language mismatch SER.** Under this condition, the source audio samples were sampled from IEMOCAP [23], a well-known SER data set in English. The German SER corpus EMODB [24] was selected as the target domain.

• Task 2: **Recording conditions mismatch SER.** The presence of background noise can pose a significant challenge for SER systems, particularly because it can adversely impact the ability of models to discern and accurately classify prosodic features of speech, which are critical for emotion detection [7]. We selected IEMOCAP to conduct these sub-experiments. 5-fold cross-validation was performed. Then, 80% samples of IEMOCAP (clean speech) were treated as the source data, and the remaining samples were added with background noise (noise, music, and babble) from MUSAN [28] at three signal-to-noise ratio (SNR) levels (0, 5, and 10 dBs) to generate three target domain data sets.

• Task 3: **Audio duration mismatch SER.** Overly short speech utterances can result in the loss of important emotional information, which can negatively impact the ability of SER models to associate contextual emotional cues and ultimately lead to a decrease in performance. To further investigate the generalization of the proposed SOT, we conducted SER experiments under this condition. IEMOCAP (4.46-s average utterance duration) was the source domain, and CREMA-D [25] (2.63 s) was the target domain.

SER system: The input acoustic feature was extracted from Hubert-Base [5, 6], which had the same configurations as that used in [29]. ECAPA-TDNN [30] with 512 dimensional was used as the backbone model (6.2M parameters). To investigate the proposed SOT more comprehensively, we not only compared its performance with that reported in existing studies but also implemented several UDA methods in SER to run the comparison experiments based on our SER benchmark, including the domain adversarial training (DAT) [15] and soft label-based domain adversarial training (DASL) [7] model. In the training stage, adam with momentum 0.9, weight decay 4e-6, and initial learning rate 0.001 was utilized. The size of mini-batch was 16. α and β in Eq. (3) were empirically set as 0.1, 0.001. In the evaluation stage, unweighted and weighted accuracies (UA and WA) were used as the evaluation metrics.

3.2. Experimental results

We evaluated the accuracy of our SER system in the source domain, achieving 69.7% UA and 70.0% WA (5-fold leave-one-session-out cross-validation was performed), as shown in Table 2, which matches the state-of-the-art UDA's performance. Then, we established systemic domain adaptation benchmarks for three OoD tasks based on our SER system.

• Task 1: Language mismatch SER. The experiment involved training the model on the source set (IEMOCAP) and running the UDA algorithm on the target domain (EMODB) without any label information. The results presented in Table 3 show that the SER models had limited performance under the OoD problem, despite achieving good accuracy in the source

Table 3: *Adaptation performance on the language mismatch condition. Source domain: IEMOCAP; target domain: EMODB.*

Description	Adaptation method	UA (%)	WA (%)
Zong *et al.* [9]	DALSR	63.23	66.08
Liu *et al.* [34]	DoSL	62.91	67.85
Ahn *et al.* [35]	FLUDA	56.80	-
Liu *et al.* [36]	TRaSL	61.02	65.49
Our implement	No adaptation	58.69	57.74
Our implement	DAT	60.29	59.81
Our implement	DASL	62.88	62.72
Proposed	SOT	**70.65**	**70.39**

Table 4: *Adaptation performance on the background noise OoD tasks (based on IEMOCAP) with different SNR levels.*

Description	Method	0 dB		5 dB		10 dB	
		UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)
Tiwari *et al.* [12]	GNM	46.34	-	50.48	-	55.24	-
Our implement	No adap.	44.33	41.98	52.29	51.77	59.40	59.60
Our implement	DAT	46.70	46.42	51.49	51.40	59.85	59.94
Our implement	DASL	48.01	47.58	52.65	52.57	61.21	61.28
Proposed	SOT	**50.42**	**50.43**	**56.10**	**56.17**	**63.79**	**63.99**

domain. For instance, our baseline model achieved 69.7% UA on IEMOCAP but only 58.7% UA on EMODB. While adversarial training was found to mitigate language OoD reduction, as evident in DASL's 62.9% UA on EMODB, the performance gain of the method was limited because it destroyed the intrinsic characteristics of emotion representation. In contrast, the proposed SOT achieved a remarkable 70.7% UA on the target domain, representing a relative increase of 11% compared to the best baseline system. This outcome can be attributed to the proposed SOT's ability to reduce domain discrepancies through the GPDD measurement and latent variable exploration, which maximally preserves the intrinsic features of the discriminant.

• Task 2: Recording conditions mismatch SER. Maintaining the performance of SER systems is challenging when background noise impairs prosodic features that are critical for emotion detection [7]. The performance of adversarial training-based UDA methods, such as DAT and DASL, in adapting to this challenge was limited, as demonstrated in Table 4. By contrast, the proposed SOT achieved satisfactory performance and outperformed the state-of-the-art systems.

• Task 3: Audio duration mismatch SER. The experimental results are presented in Table 5, where the best UDA system based on adversarial training achieved a limited 52.4% UA. In contrast, the proposed SOT achieved remarkable performance with a 60.4%, indicating an 8% relative increase compared to the state-of-the-art system (CWW + Unsup. [33]).

In summary, we conducted experiments on three challenging yet recurring OoD problems in cross-domain SER to evaluate our proposed hypothesis. The experimental results show that our UDA framework achieved state-of-the-art performance, attributed to its ability to effectively mitigate domain differences while preserving the intrinsic information in discriminating emotion.

4. Discussions

To identify the separate contributions of the proposed MOT and SSLAE, we conducted an ablation experiment on the language mismatch condition. The results are shown in Table 6, demonstrating the effectiveness of each proposed module.

Although both the GPDD and intra-class structure information are important, finding their appropriate balance in UDA (η and μ in Eq. (10)) is crucial for achieving optimal adaptation performance. For this, we conducted SER experiments on the above three conditions. The results are summarized in Fig. 3, in which the proposed SOT achieved the best performance for the language mismatch task when the weights of MOT and SSLAE were 1.0 and 0.01. The best performance for the adaptation in background noise was achieved under $\eta = 0.5$ and $\mu = 0.01$. In the utterance duration mismatch task, SOT obtained the best performance under $\eta = 0.5$ and $\mu = 0.005$, respectively.

Table 5: *Adaptation performance on the duration mismatch condition.*

Description	Adaptation method	UA (%)	WA (%)
Ahn *et al.* [33]	CWW	53.80	-
Ahn *et al.* [33]	CWW + Unsup.	55.80	-
Parry *et al.* [37]	CR	53.71	-
Our implement	No adaptation	48.88	46.08
Our implement	DAT	50.29	48.66
Our implement	DASL	52.40	50.99
Proposed	SOT	**60.39**	**59.70**

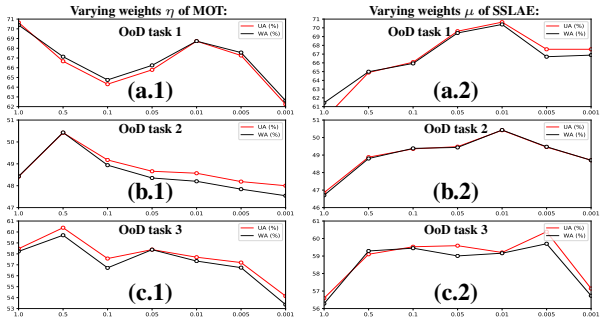

Figure 3: *Performance (UA and WA) based on SOT training while varying weights of MOT (.1) and SSLAE (.2) for three OoD tasks. Task 1: (a); task 2: (b); task 3: (c).*

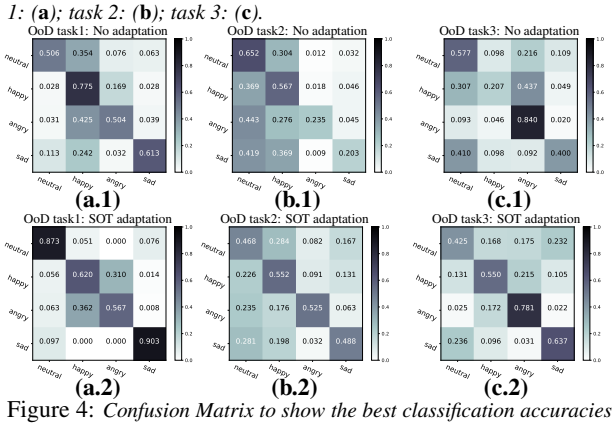

Figure 4: *Confusion Matrix to show the best classification accuracies without adaptation (.1) and with SOT adaptation (.2) for three OoD tasks. Task 1: (a); task 2: (b); task 3: (c).*

Furthermore, to properly study adaptive systems, it is crucial to evaluate the performance improvement of each emotion category between the unadapted baseline and the proposed adaptive system. In Fig. 4, we present the confusion matrix to illustrate the observed changes. By comparing figures 4 (a/b/c.1) and (a/b/c.2), we can see that the proposed SOT significantly improved the accuracy of *neutral*, *angry*, and *sad*, while only slightly affected the accuracy of *happy*. Additionally, the SOT system could correct most of the sentiment errors in the target domain for other unmatched scenes.

Table 6: *Ablation studies for SOT under the language mismatch.*

	SOT	No SSLAE	No MOT	No SSLAE and MOT
UA (%)	70.65	67.96	62.19	58.69
WA (%)	70.39	67.69	62.52	57.74

5. Conclusion

In this study, we proposed SOT, an unsupervised domain adaptation algorithm for cross-domain SER. We first regularized OT's transport coupling to mitigate negative transports, ensuring that OT can effectively measure the GPDD of different emotional categories. Then, we designed an SSLAE module to emphasize local intra-class structure, assisting OT in capturing the emotional representations, especially in hard-discriminative samples. Experimental results indicate that the proposed SOT dramatically outperformed the state-of-the-art UDA algorithms in cross-domain SER. It would be interesting for future works to decide whether to transport the local feature or the GPDD according to the individual speaker characteristics.

6. Acknowledgements

This work was supported by National Key R&D Program of China (No. 2020YFC2004103), Qinghai science and technology program (No. 2022-ZJ-T05), the project of Tianjin science and technology program (No. 21JCZXJC00190), the National Natural Science Foundation of China (Nos. 62272340 and 62201571).

7. References

[1] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech communication*, vol. 53, no. 9-10, pp. 1062–1087, 2011.

[2] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.

[3] Z. Liang, B. Liu, and J. Tao, "Smin: Semi-supervised multimodal interaction network for conversational emotion recognition," *IEEE Transactions on Affective Computing*, 2022.

[4] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*. IEEE, 2016, pp. 5200–5204.

[5] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[6] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, pp. 1194–1198.

[7] H. Li, Y. Kim, C.-H. Kuo, and S. S. Narayanan, "Acted vs. Improvised: Domain Adaptation for Elicitation Approaches in Audio-Visual Emotion Recognition," in *Proc. Interspeech 2021*, 2021, pp. 3395–3399.

[8] Y. Ahn, S. J. Lee, and J. W. Shin, "Multi-corpus speech emotion recognition for unseen corpus using corpus-wise weights in classification loss," *Proc. Interspeech 2022*, pp. 131–135, 2022.

[9] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE signal processing letters*, vol. 23, no. 5, pp. 585–589, 2016.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[11] B.-H. Su and C.-C. Lee, "A conditional cycle emotion gan for cross corpus speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop*. IEEE, 2021, pp. 351–357.

[12] U. Tiwari, M. Soni, R. Chakraborty, A. Panda, and S. K. Kopparapu, "Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions," in *Proc. ICASSP 2020*. IEEE, pp. 7194–7198.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[14] J. Gideon, M. G. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1055–1068, 2019.

[15] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *Proc. ICASSP 2018*. IEEE, pp. 4889–4893.

[16] Y. Gao, S. Okada, L. Wang, J. Liu, and J. Dang, "Domain-invariant feature learning for cross corpus speech emotion recognition," in *Proc. ICASSP 2022*. IEEE, pp. 6427–6431.

[17] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample-problem," *Proc. NIPS*, vol. 19, 2006.

[18] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Transactions on Affective Computing*, 2022.

[19] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[20] Z. Chen, S. Wang, and Y. Qian, "Self-supervised learning based domain adaptation for robust speaker verification," in *Proc. ICASSP 2021*. IEEE, pp. 5834–5838.

[21] R. Flamary, N. Courty, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 1, 2016.

[22] R. Zhang, J. Wei, X. Lu, W. Lu, D. Jin, L. Zhang, and J. Xu, "Optimal transport with a diversified memory bank for cross-domain speaker verification," in *Proc. ICASSP*. IEEE, 2023.

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[25] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[26] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy, "Joint distribution optimal transportation for domain adaptation," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[27] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. CVPR 2021*, pp. 15 750–15 758.

[28] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[29] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory, "Speaker normalization for self-supervised speech emotion recognition," in *Proc. ICASSP 2022*. IEEE, pp. 7342–7346.

[30] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[31] D. Hu, X. Hu, and X. Xu, "Multiple Enhancements to LSTM for Learning Emotion-Salient Features in Speech Emotion Recognition," in *Proc. Interspeech 2022*, 2022, pp. 4720–4724.

[32] J. Kim, Y. An, and J. Kim, "Improving Speech Emotion Recognition Through Focus and Calibration Attention Mechanisms," in *Proc. Interspeech 2022*, 2022, pp. 136–140.

[33] Y. Ahn, S. J. Lee, and J. W. Shin, "Multi-Corpus Speech Emotion Recognition for Unseen Corpus Using Corpus-Wise Weights in Classification Loss," in *Proc. Interspeech 2022*, pp. 131–135.

[34] N. Liu, Y. Zong, B. Zhang, L. Liu, J. Chen, G. Zhao, and J. Zhu, "Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning," in *Proc. ICASSP 2018*. IEEE, pp. 5144–5148.

[35] Y. Ahn, S. J. Lee, and J. W. Shin, "Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation," *IEEE Signal Processing Letters*, vol. 28, pp. 1190–1194, 2021.

[36] N. Liu, B. Zhang, B. Liu, J. Shi, L. Yang, Z. Li, and J. Zhu, "Transfer subspace learning for unsupervised cross-corpus speech emotion recognition," *IEEE Access*, vol. 9, pp. 95 925–95 937, 2021.

[37] J. Parry, E. DeMattos, A. Klementiev, A. Ind, D. Morse-Kopp, G. Clarke, and D. Palaz, "Speech Emotion Recognition in the Wild using Multi-task and Adversarial Learning," in *Proc. Interspeech 2022*, 2022, pp. 1158–1162.