



Cross-corpus open-set Speech Emotion Recognition Method Based on Spatiotemporal Features with Inverse-Entropy Regularization

ZhaoHui Zhou¹, Hui Luo¹

¹College of Computer and Control Engineering, Northeast Forestry University, Harbin, PRC

837885876@qq.com, luohui0216@163.com

Abstract

We propose a method to address the performance degradation due to the distribution shift and unknown emotion categories in cross-corpus open-set speech emotion recognition. The method combines spatiotemporal feature extraction and inverse-entropy regularization. First, the long-range spatiotemporal dependencies are extracted from emotional audio sequences using a deep fusion network. To further align distributions from the source and target corpora, the MMD regularization is applied to minimize the distance between their joint distributions. Moreover, we propose an inverse-entropy regularization to learn the discriminative information used to reject known classes, which can balance the classification confidence of samples from the known or unknown categories in the open-set setting, allowing the model to predict unknown classes while preventing over-prediction. Experimental results show that our method outperforms baseline models across four cross-corpus speech emotion datasets.

Index Terms: Speech Emotion Recognition, Open-Set, Domain Adaptation, Inverse-Entropy, CNN, LSTM

1. Introduction

Speech Emotion Recognition (SER) plays a pivotal role in human-computer interaction (HCI) systems, with applications ranging from affective computing to adaptive voice assistants [1, 2]. Conventional SER methodologies rely on the closed-set assumption, which restricts test categories to strictly match those encountered during training. However, real-world scenarios inevitably involve unknown emotional categories (e.g., nuanced affective blends or culturally specific expressions) and cross-corpus distributional discrepancies in joint probability spaces. This necessitates open-set SER frameworks capable of dual functionality: precise recognition of known emotions across domains and robust rejection of out-of-distribution categories.

Current open-set SER research confronts two primary challenges: 1) cross-domain distribution shifts between various speech corpora; 2) potential presence of category-level unknowns during cross-corpus evaluation.

To address cross-dataset domain shifts in SER, we propose STIENet (SpatioTemporal-Inverse Entropy Network), a unified optimization framework integrating multiscale spatiotemporal feature learning with inverse entropy regularization. While DAOD [3] establishes theoretical guarantees for domain alignment through its generalized error upper bound, it suffers from residual distributional biases even when label spaces are partially aligned—a critical limitation in emotion recognition where acoustic features exhibit high inter-domain variance. Our architecture addresses this gap by synergizing CNN-extracted

local acoustic patterns with LSTM-modeled temporal dynamics, enhancing shift robustness through hierarchical feature abstraction [4, 5]. STIENet extends DAOD’s framework through two key innovations:

- **Pseudo-unknown category infusion:** We inject dynamically generated pseudo-unknown samples during training, explicitly modeling the open-set decision boundaries that DAOD implicitly assumes.
- **Entropy-aware regularization:** We introduce an inverse entropy term that adaptively smooths domain-specific confidence thresholds, directly countering DAOD’s vulnerability to distribution-biased label correlations.

The joint optimization objective combines DAOD’s theoretical foundation with these enhancements, enabling simultaneous domain alignment and unknown rejection. This dual mechanism specifically mitigates DAOD’s oversight of persistent feature distribution biases in aligned label spaces. Figure 1 illustrates the proposed framework’s architecture and information flow.

2. Related work

This section reviews closed-set and open-set domain adaptation approaches for cross-corpus SER, focusing on scenarios with exclusive source label availability and complete absence of target labels. Specifically, we concentrate on the open-set partial domain adaptation setting where the source label space constitutes a subset of the target label space, with target domains containing exclusive unknown classes.

2.1. Closed-Set Methods

Closed-set approaches learn domain-invariant representations via techniques like autoencoders, NMF, PCA, and LDA [6, 7, 8, 9, 10, 11], some preserving geometric structures. Distributional shifts are mitigated through MMD, graph-based metrics [7, 8, 11], or classifier-regularized optimization [12]. Deep architectures like PCANet, neural networks, and attention-based multimodal systems [9, 13, 14] complement these shallow methods, with adversarial training emerging for cross-corpus SER [15, 16].

2.2. Open-Set Methods

While closed-set adaptation presumes identical source-target label spaces, open-set approaches address practical scenarios with partial overlaps and domain-specific private classes, unified as unknown categories. The OSNN architecture [17] achieves effective unknown-class identification, whereas the framework in [18] employs discriminative weighting to separate known/unknown classes. LEAD [19] enhances recognition effi-

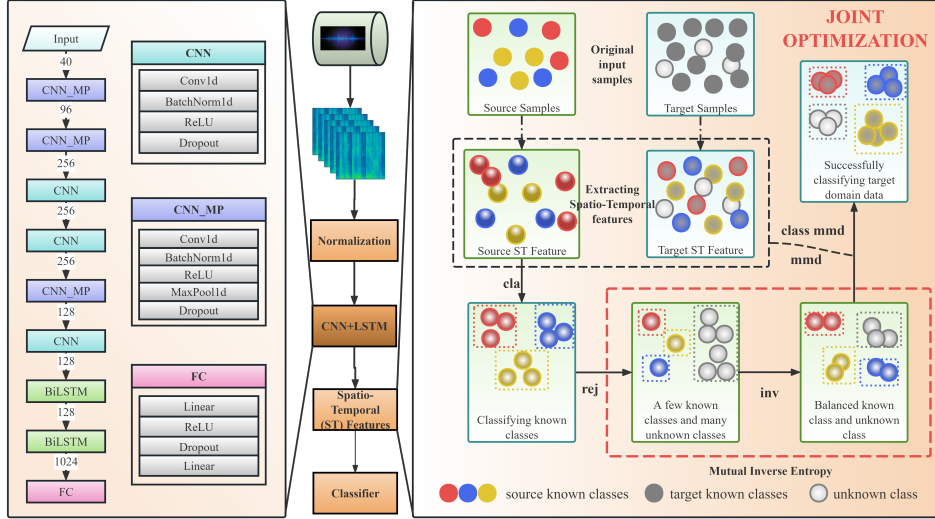


Figure 1: *Parameter sensitivity study for our method.*

ciency through orthogonal feature decomposition and instance-level decision boundaries. DAOD [3] optimizes adaptation via structural risk minimization coupled with open-domain differential regularization.

3. The proposed method

3.1. Formal Definition

Let the source domain data be $D^s = \{(x_i^s, y_i^s)\}_{i=1}^n$, where $y_i^s \in C^s = \{1, \dots, K\}$ represents the known emotional categories; the target domain data is $D^t = \{x_j^t\}_{j=1}^m$, with labels being unknown. The label space includes known classes $C^t = C^s \cup C^{\text{unk}}$, where C^{unk} denotes the unknown emotional categories [3].

The objective of open-set speech emotion recognition is as follows:

- **Known Category Classification:** Correctly classify samples x_j^t from the target domain into the known categories C^s .
- **Unknown Category Rejection:** Reject the prediction for samples x_j^t belonging to the unknown category C^{unk} .

3.2. Deep Spatiotemporal Feature Learning

The inherent time-frequency duality in speech emotion expression necessitates a hierarchical architecture capable of concurrently modeling localized spectral patterns and global temporal dependencies. Our framework addresses this through synergistic integration of convolutional and recurrent processing.

The proposed CNN-LSTM hybrid architecture consists of three core modules:

The convolutional neural network (CNN) module extracts multi-scale acoustic features from raw speech signals through stacked 1D convolutional operations. Initial layers employ large kernels (sizes 15 and 5) to expand channel dimensions and reduce temporal resolution, while deeper layers utilize small kernels (size 3) for fine-grained local pattern extraction. Each convolutional block integrates batch normalization, ReLU activation, max pooling, and dropout regularization, ultimately producing compressed temporal features with 128 channels.

The bidirectional long short-term memory (Bi-LSTM) module then captures long-range contextual dependencies us-

ing a two-stage cascaded structure, where each layer contains 64 hidden units and processes sequences bidirectionally to model forward and backward temporal dynamics.

Finally, the fully connected classifier projects Bi-LSTM outputs into a joint label space through a 1024-dimensional hidden layer with ReLU activation, generating $\text{num_classes} + 1$ outputs—the additional dimension explicitly addresses unknown-class rejection in open-set scenarios. Dropout and batch normalization are applied throughout the network to mitigate overfitting.

The combination of CNN and LSTM forms a spatiotemporal feature encoder $\phi(x)$, which generates discriminative emotional representations.

3.3. Domain Adaptation and Distribution Alignment

To mitigate the covariate shift between the source domain and target domain, the projected Maximum Mean Discrepancy (MMD [3]) is adopted as the distribution distance metric:

$$\mathcal{L}_{\text{MMD}} = \left\| \frac{1}{n} \sum_{i=1}^n \phi(x_i^s) - \frac{1}{m} \sum_{j=1}^m \phi(x_j^t) \right\|_2^2 \quad (1)$$

Furthermore, a class-conditional MMD is introduced to align cross-domain distributions of the same emotional features:

$$\mathcal{L}_{\text{MMD-class}} = \sum_{k=1}^K \left\| \frac{1}{n_k} \sum_{\phi(x_i^s) \in D_k^s} \phi(x_i^s) - \frac{1}{m_k} \sum_{\phi(x_j^t) \in D_k^t} \phi(x_j^t) \right\|_2^2, \quad (2)$$

where D_k^s is the set of samples belonging to class k in the source data, and $n_k = |D_k^s|$. Correspondingly, D_k^t is the set of samples belonging to class k (pseudo) in the target data, and $m_k = |D_k^t|$.

3.4. Unknown Class Detection and Inverse Entropy Regularization

To address the dual challenge of unknown-class rejection and known-class preservation, we propose a composite loss mechanism comprising two synergistic components. We formulate

an unknown-class rejection loss that leverages pseudo-labeled samples to train the model in distinguishing known/unknown boundaries:

$$\mathcal{L}_{\text{rej}} = \mathbb{E}_{x \sim D^s} [\text{CE}(f(\varphi(x)), y_{\text{unk}})], \quad (3)$$

where $y_{\text{unk}} = K + 1$ is the pseudo-label for the unknown class, and CE denotes the crossentropy loss. This formulation encourages the feature extractor $\phi(x)$ to map unknown-class candidates (simulated via dropout-based feature perturbation) toward low-confidence regions in the known class probability simplex.

To prevent excessive rejection of ambiguous known-class samples, we introduce an inverse-entropy regularizer:

$$\mathcal{L}_{\text{inv}} = \frac{1}{\mathcal{L}_{\text{rej}} + \varphi}, \quad (4)$$

where φ ensures numerical stability. This term imposes a reciprocal relationship between rejection loss minimization and decision boundary contraction. The inverse-entropy term acts as a Lagrangian multiplier that constrains the rejection loss growth rate during optimization. Let $\nabla_{\theta} \mathcal{L}_{\text{rej}}$ be the rejection loss gradient with respect to parameters θ . The regularization introduces an adaptive gradient scaling factor $-1/(\mathcal{L}_{\text{rej}} + \varphi)^2$, which automatically attenuates parameter updates when \mathcal{L}_{rej} becomes excessively small. This mechanism prevents decision boundary over-expansion while maintaining gradient stability.

3.5. Joint Optimization

The total loss function combines the objectives of each component, optimized through the addition of regularization and balancing:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{cls}} + \beta \mathcal{L}_{\text{rej}} + \gamma \mathcal{L}_{\text{mnd}} + \delta \mathcal{L}_{\text{mnd-class}} + \epsilon \mathcal{L}_{\text{inv}}, \quad (5)$$

where $\mathcal{L}_{\text{cls}} = \mathbb{E}_{(x^s, y^s)} [\text{CE}(f(\phi(x^s)), y^s)]$ is the source domain classification loss. The hyperparameters $\alpha, \beta, \gamma, \delta, \epsilon$ control the contribution of each loss term.

4. Experiments

4.1. Experimental Setup

To rigorously assess the proposed methodology, we established a comprehensive evaluation framework. All experiments were executed in a uniform computational environment to ensure fair comparison. The implementation details are elaborated below.

Our experiments leveraged four benchmark SER corpora: Berlin [20], CASIA [21], eINTERFACE [22], and SAVEE [23], with detailed statistical characteristics presented in Table 1. This configuration generates 12 distinct source→target domain adaptation pairs. We selected three fundamental emotional categories (anger, fear, happiness) as shared classes across corpora, while treating remaining classes in target corpora as unknown samples. To address class imbalance, all datasets were resampled to ensure balanced class distributions through strategic oversampling of underrepresented categories.

Table 1: Overview of the Speech Emotion Recognition Corpus

Datasets	Language	# Sample	# Emotion
Berlin(B)	German	535	7
eINTERFACE(e)	English	1257	6
SAVEE(S)	English	480	7
CASIA(C)	Chinese	900	6

Our implementation adopted a hybrid CNN-LSTM architecture, whose effectiveness for speech emotion recognition (SER) has been demonstrated in prior studies. For comparative analysis, we implemented five state-of-the-art open-set recognition methods as baselines: OSNN [17], TNMF [24], ATI [25], LEAD [19], DAOD [3] and OSBP [26], strictly following their original configurations and parameter settings.

Following established practices in open-set recognition [3, 21], we employed two complementary evaluation metrics:

- **OS Accuracy:** Measures the overall classification accuracy across all test samples, including instances of unknown class.
- **OS* Accuracy:** Quantifies classification accuracy only measured on the shared 3 classes.

4.2. Cross-Corpus Comparisons

A comprehensive set of 12 cross-corpus experiments was performed using all possible pairwise combinations of the speech corpora. Each experimental configuration was executed for 30 epochs with hyperparameters dynamically adjusted through consideration of training loss convergence patterns. All experiments were repeated 10 times with different random seeds, and the reported values in Tables 2, 3 represent the optimal performance achieved across these independent trials (all values reported as percentages).

As evidenced by the comprehensive results, STIENet demonstrates superior performance across all 12 crosscorpus configurations, achieving state-of-the-art results in both OS (47.61% on average) and OS* (68.33% on average) metrics. The differential improvement patterns between OS Accuracy and OS* Accuracy provide critical insights into STIENet’s unknown-class handling capability. The conventional approaches (OSNN, TNMF, ATI) show severe performance degradation under substantial domain gaps. For instance, in the C→B configuration with maximum acoustic-prosodic discrepancy, OSNN achieves merely 21.25% OS Accuracy (vs. STIENet’s 49.17%), while TNMF collapses to 11.02% and ATI is 23.42%. This suggests they inadequately capture cross-corpus invariant patterns and unknown classes. LEAD’s average OS Accuracy plummets to 35.79%, primarily due to its strict source-free assumption conflicting with our experimental protocol where source data remains accessible. This highlights the criticality of source-target interaction in open-set SER. STIENet demonstrates superior open-set balancing with an average OS-OS* gap of only 20.72%, compared to DAOD’s significant 28.17% discrepancy. This indicates our method’s unique ability to jointly optimize knownclass precision and unknown-class rejection.

4.3. Ablation Studies

To systematically evaluate the contribution of individual components, we conducted rigorous ablation experiments across 10 independent trials for each hyperparameter configuration, with performance quantified by mean accuracy. Comprehensive ablation tests revealed that removal of any proposed loss component induces significant performance degradation, as illustrated in Figure 2. The catastrophic performance degradation when α was nullified ($\alpha = 0$) suggests that the source-domain cross-entropy loss (\mathcal{L}_{cls} in Equation 5) provides essential supervisory signals for preserving class-discriminative feature representations, acting as a stabilizing anchor against domain shift. The removal of ϵ regularization induces substantial performance degradation across all evaluation metrics, revealing its critical role in open-set decision boundary calibration. These findings

Table 2: Accuracy Table I for Different Methods on Various Dataset Combinations

Models	B→C		B→e		B→S		C→B		C→e		C→S	
	OS	OS*	OS	OS*	OS	OS*	OS	OS*	OS	OS*	OS	OS*
OSNN	26.33	49.00	29.15	58.26	23.24	58.65	21.25	48.13	22.69	44.45	23.46	67.24
TNMF	25.24	46.97	22.18	44.29	19.12	49.27	11.02	24.85	25.25	50.09	17.33	44.48
ATI	30.40	49.44	32.22	41.28	24.38	60.18	23.42	49.20	27.81	51.01	31.61	57.17
LEAD	0.50	0.00	31.99	20.95	41.25	13.89	36.45	17.31	33.33	6.49	50.83	11.70
OSBP	22.17	9.00	32.83	38.00	32.50	45.67	28.71	38.17	37.00	49.33	29.38	31.33
DAOD	27.21	35.17	27.88	34.39	22.67	22.78	54.18	62.79	29.57	37.48	36.83	40.56
STIENet	46.33	60.56	45.66	62.09	49.17	65.56	60.36	80.48	38.39	51.18	48.83	63.78

Table 3: Accuracy Table II for Different Methods on Various Dataset Combinations

Models	e→B		e→C		e→S		S→B		S→C		S→e	
	OS	OS*	OS	OS*	OS	OS*	OS	OS*	OS	OS*	OS	OS*
OSNN	42.71	51.24	32.21	28.20	27.06	45.19	36.08	70.25	45.00	14.00	53.18	3.11
TNMF	61.21	14.46	49.06	15.30	57.09	22.24	10.22	28.24	55.00	9.01	43.98	5.34
ATI	51.46	22.72	47.47	16.22	29.44	48.00	17.20	34.48	53.33	11.00	47.26	16.21
LEAD	48.60	9.70	47.25	8.30	47.71	11.70	41.68	21.54	38.50	15.30	39.90	14.70
OSBP	32.21	39.33	34.38	41.17	32.50	34.50	31.80	36.33	33.75	40.50	33.71	42.33
DAOD	41.35	45.95	30.37	34.67	29.33	30.56	49.25	61.78	36.42	45.17	31.63	35.54
STIENet	56.82	75.76	40.33	53.75	42.80	56.11	58.00	74.18	42.21	56.28	40.52	54.03

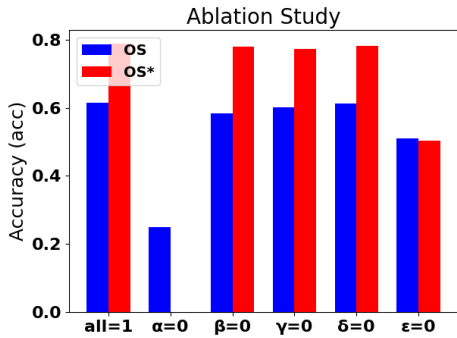


Figure 2: Bar Chart Comparison for Ablation Study. All = 1 means all five parameters ($\alpha, \beta, \gamma, \delta, \epsilon$) are set to 1. Others = 0 means each parameter is individually disabled (set to 0) to evaluate its impact.

validate our theoretical claim in Section 3.4) that ϵ -mediated regularization is essential for learning rejection-aware representations without compromising known-class discriminability.

4.4. Parameter Sensitivity Analysis

To systematically investigate the operational characteristics of key hyperparameters, we performed comprehensive sensitivity evaluations using the C→B cross-corpus configuration (selected as the representative case study due to space constraints). Each parameter was independently swept through its operational range while maintaining others at optimal values identified in Section 4.3). Performance metrics were recorded across 10 randomized trials to ensure statistical reliability.

As evidenced by the complete parameter response surface in Figure 3(α, γ, δ), our method exhibits remarkable insensitivity to auxiliary parameters $\alpha(>0)$, γ and δ . This parameter robustness contrasts sharply with the critical sensitivity observed for β , and ϵ in Figure 3(β, ϵ), and we suggest that β range in

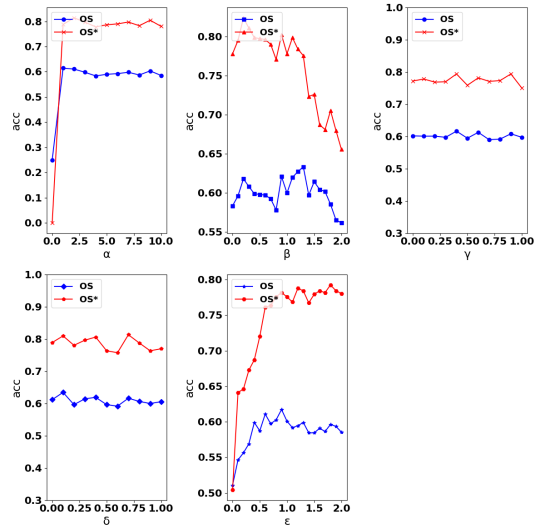


Figure 3: Parameter Sensitivity Analysis Plot

[1.0, 1.5] and ϵ range in [1.5, 2.0].

5. Conclusions

We proposed STIENet, a cross-corpus open-set SER framework that integrates spatiotemporal feature learning and inverse-entropy regularization to address domain shifts and unknown-class detection. The CNN-LSTM hybrid architecture captures robust acoustic-temporal patterns, while the entropy-aware regularization sharpens decision boundaries between known and unknown emotions. Experiments on four SER corpora demonstrate state-of-the-art performance, with ablation studies confirming the necessity of core components. The framework’s parameter robustness and cross-domain generalization highlight its practicality for real-world deployment.

6. Acknowledgment

This work is supported by Natural Science Foundation of Heilongjiang Province of China under grant No. LH2023F002 and National Science Foundation of China under grant No. 62101114.

7. References

- [1] R. C. et al, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.
- [2] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot," in *2010 International Conference on Artificial Intelligence and Computational Intelligence*, vol. 1. IEEE, 2010, pp. 537–541.
- [3] Z. Fang, J. Lu, F. Liu, J. Xuan, and G. Zhang, "Open set domain adaptation: Theoretical bound and algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4309–4322, Oct. 2020.
- [4] J. Parry, D. Felps, and S. Sundaram, "Analysis of deep learning architectures for cross-corpus speech emotion recognition," in *Interspeech*, 2019, pp. 1656–1660.
- [5] T. Swain, U. Anand, Y. Aryan, S. Khahra, A. Raj, and S. Patnaik, "Performance comparison of lstm models for ser," in *Proceedings of International Conference on Communication, Circuits, and Systems: IC3S 2020*. Springer, 2021, pp. 427–433.
- [6] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, Sep. 2014.
- [7] H. Luo and J. Han, "Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2047–2060, 2020.
- [8] H. Luo and J. Han, "Cross-corpus speech emotion recognition using semi-supervised transfer non-negative matrix factorization with adaptation regularization," in *Interspeech*, 2019, pp. 3247–3251.
- [9] Z. Huang, W. Xue, Q. Mao, and Y. Zhan, "Unsupervised domain adaptation for speech emotion recognition using pcanet," *Multimedia Tools and Applications*, vol. 76, pp. 6785–6799, 2017.
- [10] P. Song and W. Zheng, "Feature selection based transfer subspace learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 373–382, Jul. 2018.
- [11] W. Zhang and P. Song, "Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 307–318, 2019.
- [12] N. Liu, M. Li, H. Li, S. Shan, and R. Ji, "Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5144–5148.
- [13] H. Kaya, D. Fedotov, A. Yesilkanat, O. Verkholyak, Y. Zhang, and A. Karpov, "Lstm based cross-corpus and cross-task acoustic emotion recognition," in *Interspeech*, 2018, pp. 521–525.
- [14] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," 2018.
- [15] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, Dec. 2018.
- [16] 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, Oct. 2019.
- [17] P. R. M. Junior, R. M. de Souza, R. d. O. Werneck, B. V. Stein, D. V. Pazinato, W. R. de Almeida, O. A. Penatti, R. d. S. Torres, and A. Rocha, "Nearest neighbors distance ratio open-set classifier," *Machine Learning*, vol. 106, no. 3, pp. 359–386, 2017.
- [18] H. Liu, Z. Cao, M. Long, J. Wang, and Q. Yang, "Separate to adapt: Open set domain adaptation via progressive separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2927–2936.
- [19] S. Qu, X. Wang, Q. Lv, B. Zhang, L. Liu, and R. Jin, "Lead: Learning decomposition for source-free universal domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 334–23 343.
- [20] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Inter-speech*, vol. 5, 2005, pp. 1517–1520.
- [21] ChineseLDC, "Casia-chinese emotional speech corpus," <http://www.chineseldc.org/>, 2005.
- [22] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.
- [23] S. Haq, P. J. Jackson, and J. D. Edge, "Audio-visual feature selection and reduction for emotion classification," in *APSP*, 2008, pp. 185–190.
- [24] P. Song, S. Ou, W. Zheng, Y. Jin, and L. Zhao, "Speech emotion recognition using transfer non-negative matrix factorization," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5180–5184.
- [25] P. P. Busto and J. Gall, "Open set domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 754–763.
- [26] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada, "Open set domain adaptation by backpropagation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 153–168.