Supplementary Material for Decoupled Feature Interaction for Sparse EEG-Based Emotion Recognition

I. Experiments

We evaluated the performance of the algorithmic model using the average precision and the F1-score, expressed as follows:

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\ Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \\ F1\ score &= \frac{1}{|C|} * \sum_{i=1}^{C} \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i} \end{aligned} \tag{1}$$

where TP denotes true positive, TN denotes true negative, FP denotes false positive, FN denotes false negative, and C is the number of categories.

A. Baseline Models

- EEGNet [1]: EEGNet is a classical CNN used for EEG-based brain-computer interfaces. It has been used for P300 visual-evoked potentials, Error-Related Negativity (ERN) responses, Movement-Related Cortical Potentials (MRCP), and Sensory Motor Rhythms (SMR).
- Tsception [2]: Tsception uses multi-scale convolution for emotion recognition, consisting of dynamic temporal layers, asymmetric spatial layers, and fusion layers, with a focus on EEG time-frequency and asymmetric EEG patterns.
- FBMSNet [3]: FBMSNet is a neural network for decoding EEG motor imagery. It processes EEG data through filter banks to obtain a multi-view spectral representation as input, then extracts multi-scale time features using mixed depth convolution, performs spatial filtering by spatial convolution blocks, calculates logarithmic variance dimensionalization by time log-variance blocks, and finally completes classification by a full connection layer and softmax activation function.
- STNet [4]: STNet uses spatio-temporal convolution for emotion recognition, including temporal layer, spatial layer, and ensemble layer, which focuses on capturing the temporal dynamics and spatial correlation of EEG.
- MMSF [5]: MMSF is a deep learning network that uses multimodal physiological signals for emotion recognition.
- AttX [6]: AttX is a cross-modal attentional connectivity method for multimodal representation learning. It establishes intermediate connections in a multi-stream network, allowing information to be shared between different modalities for better representation learning.
- MulT [7]: MulT is a cross-modal interaction network that focuses on data alignment and long-term dependencies in time series.
- BSCA [8]: This method consists of a bidirectional crossover and a modal self-attention mechanism, which effectively utilizes cross-modal information and intra-modal key information.
- MS-iMamba [9]: MS-iMamba consists of multi-scale time blocks and space-time fusion blocks. It is capable of capturing local details and global time dependencies between subsequences of different sizes, as well as focusing on the interaction between dynamic time dependencies and spatial features.
- TSFF-Net [10]: TSFF-Net aims to address the challenges in 3-channel motion imagination decoding. This method integrates temporal, spatial and frequency features, thereby overcoming the limitations of single-modal feature extraction networks based on time series or time-frequency modalities.
- CBR [11]: CBR aims to tackle the challenge of depression recognition using only three-electrode EEG data. By leveraging a case-based reasoning framework combined with dynamic time warping for similarity retrieval, this method effectively captures individual EEG patterns and mitigates the limitations of traditional machine learning approaches in small-sample clinical scenarios.

II. Discussion and Analysis

1) Feature Visualization: We visualize the feature separability of the model on the public datasets in Fig. 1

2) Generalization Performance across Unseen Subjects: To further verify the practical applicability and generalization ability of the proposed method, we conducted a systematic evaluation of the model in cross-individual scenarios. Specifically, we employed the Leave-One-Subject-Out (LOSO) method on three public datasets and one private dataset for experiments. For the public datasets, each round would select one subject as the test set, and the remaining subjects would be used for training. The testing of all subjects would be completed in a cycle. For the private dataset with

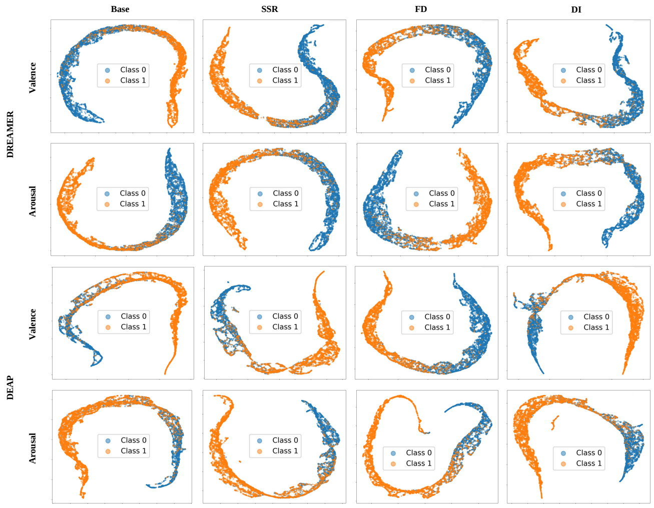

Fig. 1. Visualization of feature separability on the DEAP and DREAMER datasets.

a larger number of subjects, we divided it into 13 subgroups, and in each round, one subgroup would be selected as the test set, while the remaining subgroups would be used as the training sets for cross-group evaluation. The results are shown in Table I, Table II, Table III, and Table IV.

TABLE I
COMPARISON ON THE SEED DATASET UNDER LOSO SETTING. BOLD IS THE BEST, _ REPRESENTS A SUBOPTIMAL RESULT.

Method	Dataset	Emotion State			
		Acc_2 (%)	$F1_2$ (%)	Acc_3 (%)	$F1_3$ (%)
EEGNet [1]	SEED	56.67	48.97	44.54	40.43
Tsception [2]		58.82	57.06	42.91	40.15
FBMSNet [3]		59.04	55.46	44.70	39.97
STNet [4]		52.17	50.80	39.85	37.07
MMSF [5]		59.21	57.18	43.48	41.25
AttX [6]		60.16	58.26	43.82	41.67
MulT [7]		60.52	57.92	44.66	41.69
BSCA [8]		59.46	57.65	44.03	42.21
MS-iMamba [9]		55.50	53.99	42.56	40.18
TSFF-Net [10]		61.49	58.50	42.11	37.31
CBR [11]		49.59	47.24	33.85	31.68
DFI **(Ours)**		**61.52**	**59.54**	**45.03**	**42.56**

As shown in Table I to Table IV, the proposed DFI method demonstrates superior cross-subject generalization performance under the Leave-One-Subject-Out (LOSO) setting across four datasets. On the SEED dataset, DFI achieves the best accuracy and F1 scores in both binary and ternary classification tasks, confirming its robustness in modeling emotional features for previously unseen individuals. On the DEAP and DREAMER datasets, although some methods slightly outperform DFI in specific dimensions, DFI consistently maintains superior or stable performance in both the Valence and Arousal dimensions, indicating its capability for learning consistent representations across modalities and sub-

TABLE II
COMPARISON ON THE DEAP DATASET UNDER LOSO SETTING. BOLD IS THE BEST, _ REPRESENTS A SUBOPTIMAL RESULT.

Method	Dataset	Valence		Arousal	
		Acc_2 (%)	$F1_2$ (%)	Acc_2 (%)	$F1_2$ (%)
EEGNet [1]	DEAP	49.36	40.83	53.76	39.50
Tsception [2]		51.18	44.26	52.74	41.48
FBMSNet [3]		47.28	41.03	52.89	41.38
STNet [4]		**53.71**	37.53	56.37	36.23
MMSF [5]		48.88	47.81	52.77	40.60
AttX [6]		49.28	47.42	55.54	40.92
MulT [7]		47.39	44.28	53.36	42.76
BSCA [8]		49.54	47.95	53.74	41.73
MS-iMamba [9]		50.66	45.57	54.70	41.32
TSFF-Net [10]		50.09	46.71	53.75	42.64
CBR [11]		48.22	46.92	47.25	**46.07**
DFI (Ours)		49.52	**48.25**	**56.88**	43.26

TABLE III
COMPARISON ON THE DREAMER DATASET UNDER LOSO SETTING. BOLD IS THE BEST, _ REPRESENTS A SUBOPTIMAL RESULT.

Method	Dataset	Valence		Arousal	
		Acc_2 (%)	$F1_2$ (%)	Acc_2 (%)	$F1_2$ (%)
EEGNet [1]	DREAMER	55.37	39.83	51.92	48.53
Tsception [2]		54.18	49.87	51.32	48.85
FBMSNet [3]		53.91	49.14	51.90	48.92
STNet [4]		53.57	45.90	49.17	47.06
MMSF [5]		54.02	49.99	51.50	**49.61**
AttX [6]		51.72	50.20	50.69	48.38
MulT [7]		54.10	48.71	49.93	48.38
BSCA [8]		56.54	47.34	49.91	46.93
MS-iMamba [9]		56.00	50.38	51.75	49.42
TSFF-Net [10]		54.21	49.60	49.91	47.00
CBR [11]		55.11	47.64	51.60	47.55
DFI (Ours)		**56.77**	**50.39**	**52.74**	48.82

TABLE IV
COMPARISON ON THE PRIVATE DATASET UNDER LOSO SETTING. ACC_2 AND $F1_2$ REPRESENT THE ACCURACY AND F1-SCORE FOR THE BINARY CLASSIFICATION TASK BETWEEN DEPRESSION AND HEALTHY SUBJECTS. BOLD IS THE BEST, _ REPRESENTS A SUBOPTIMAL RESULT.

Method	Dataset	Emotion State	
		Acc_2 (%)	$F1_2$ (%)
EEGNet [1]	Private Dataset	82.50	79.89
Tsception [2]		85.02	83.36
FBMSNet [3]		89.65	88.70
STNet [4]		83.24	80.54
MMSF [5]		87.86	87.02
AttX [6]		93.73	93.21
MulT [7]		91.21	90.33
BSCA [8]		92.56	91.95
MS-iMamba [9]		89.97	89.10
TSFF-Net [10]		92.88	92.28
CBR [11]		68.57	64.56
DFI (Ours)		**94.08**	**93.62**

jects. We attribute this performance advantage to the structural-aware feature decoupling and dynamic interaction mechanisms integrated into the model design. By distinguishing between stable emotional structures (invariant features) and subject-dependent expressions (adaptive features), and aligning them via multi-scale cross-attention mechanisms, DFI can maintain semantic consistency even when the input distribution shifts. Nevertheless, this study has certain limitations. The current experiments employ only a subset of frontal channels from each dataset, without fully utilizing the whole-brain signals. While this lightweight configuration is more suitable for practical wearable applications and edge deployment scenarios, it may lead to information loss in more complex or fine-grained emotion modeling tasks. Future work could explore channel reconstruction, cross-channel transfer, or adaptive channel selection mechanisms to further enhance the model's representational capacity and generalization performance while maintaining simplicity of the recording setup.

3) Comparison under the complete channel: To further verify the effectiveness and universality of the proposed method, we conducted the verification using all the channels, as shown in Table V and Table VI. From the experimental results, it can be seen that the proposed DFI method has achieved relatively stable and overall superior performance on all three datasets. This to some extent indicates that the proposed method can achieve stable performance in the multi-channel EEG emotion recognition task and has good applicability under different dataset settings.

TABLE V
THE RESULTS OBTAINED BY USING ALL CHANNELS ON SEED. A UNIFIED TRAINING-TEST DATA SPLIT WAS ADOPTED TO EVALUATE ALL THE COMPARISON METHODS. BOLD IS THE BEST, AND _ DENOTES THE SUBOPTIMAL ONE.

Method	**Dataset**	**Emotion State**			
		Acc$_2$ (%)	**F1$_2$ (%)**	**Acc$_3$ (%)**	**F1$_3$ (%)**
EEGNet [1]	SEED	80.30	80.05	58.92	56.11
Tsception [2]		94.34	94.52	85.64	86.12
FBMSNet [3]		81.78	81.94	62.41	62.38
STNet [4]		80.99	81.15	62.65	62.79
MMSF [5]		87.91	88.08	68.64	68.96
AttX [6]		94.36	94.54	84.25	84.78
MulT [7]		88.90	89.07	73.02	73.24
BSCA [8]		94.34	94.52	86.47	87.01
MS-iMamba [9]		83.63	83.79	63.50	63.45
TSFF-Net [10]		89.09	89.26	71.35	71.95
CBR [11]		49.79	49.73	33.14	32.61
DFI **(Ours)**		**96.24**	**96.43**	**91.98**	**92.57**

TABLE VI
THE RESULTS OBTAINED BY USING ALL CHANNELS ON DEAP AND DREAMER. A UNIFIED TRAINING-TEST DATA SPLIT WAS ADOPTED TO EVALUATE ALL THE COMPARISON METHODS. BOLD IS THE BEST, AND _ DENOTES THE SUBOPTIMAL ONE.

Method	**Dataset**	**Emotion State**	
		Acc$_4$ (%)	**F1$_4$ (%)**
EEGNet [1]	DEAP	43.75	39.82
Tsception [2]		62.29	60.90
FBMSNet [3]		42.75	38.21
STNet [4]		43.28	41.16
MMSF [5]		49.41	47.92
AttX [6]		68.31	67.24
MulT [7]		48.08	46.40
BSCA [8]		77.95	77.42
MS-iMamba [9]		45.00	42.40
TSFF-Net [10]		43.61	38.69
CBR [11]		25.71	24.96
DFI (Ours)		**83.13**	**82.78**
EEGNet [1]	DREAMER	45.74	41.97
Tsception [2]		86.12	86.64
FBMSNet [3]		51.71	49.60
STNet [4]		42.63	41.23
MMSF [5]		54.95	54.15
AttX [6]		80.09	80.59
MulT [7]		61.62	61.45
BSCA [8]		85.81	86.44
MS-iMamba [9]		52.02	50.03
TSFF-Net [10]		50.91	46.32
CBR [11]		76.36	76.07
DFI (Ours)		**92.99**	**93.81**

4) Quantitative Analysis of Dense–Sparse Performance Gap: To further investigate whether the proposed framework truly addresses the core limitation introduced by sparse EEG channels, we conduct a dedicated dense–sparse gap analysis on the SEED binary classification task. Instead of only reporting absolute performance under sparse settings, we explicitly examine how much the performance gap between dense-channel EEG and sparse-channel EEG can be narrowed. Specifically, we evaluate both the baseline model and the proposed DFI framework under dense-channel and frontal sparse-channel conditions. As shown in Fig. 2, under the baseline setting, dense EEG achieves an accuracy of 95.18%, whereas frontal sparse EEG only reaches 70.17%, resulting in a substantial performance gap of 25.01%. A similar phenomenon is observed for the F1-score, where the gap reaches 26.13%. This clearly confirms that sparse-channel acquisition leads to a severe degradation of emotion recognition performance due to the loss of spatial coverage. After introducing the proposed DFI framework, dense-channel performance reaches 96.24%, while sparse-channel performance is significantly improved to 89.71%. As a result, the dense–sparse accuracy gap is reduced to 6.53%, and the F1-score gap is reduced to 6.55%. Compared with the baseline model, this indicates that approximately three quarters of the dense–sparse performance gap is effectively eliminated. These results suggest that although the spatial information loss caused by sparse channels is physically irreversible, the proposed representation enhancement, feature decoupling, and dynamic interaction modeling significantly improve the efficiency and stability of extracting emotion-discriminative information from limited observations. Rather than attempting to recover missing information, the proposed

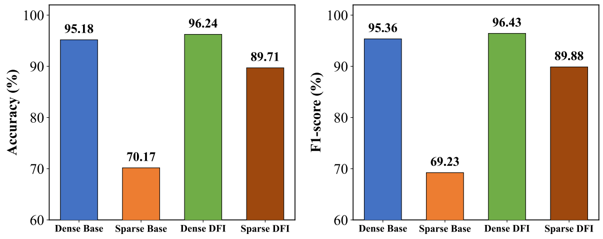

Fig. 2. Quantitative analysis of dense–sparse gap narrowing on SEED dataset

framework increases the utilization of discriminative structures embedded in sparse EEG, which is directly reflected by the substantial convergence of the dense–sparse performance gap.

References

[1] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.

[2] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan, "Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2238–2250, 2022.

[3] K. Liu, M. Yang, Z. Yu, G. Wang, and W. Wu, "Fbmsnet: A filter-bank multi-scale convolutional neural network for eeg-based motor imagery decoding," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 2, pp. 436–445, 2022.

[4] Z. Zhang, Y. Liu, and S.-h. Zhong, "GANSER: A self-supervised data augmentation framework for EEG-based emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2048–2063, 2022.

[5] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, "An explainable deep fusion network for affect recognition using physiological signals," in *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 2069–2072, 2019.

[6] A. Bhatti, B. Behinaein, P. Hungler, and A. Etemad, "Attx: Attentive cross-connections for fusion of wearable signals in emotion recognition," *ACM Transactions on Computing for Healthcare*, vol. 5, no. 3, pp. 1–24, 2024.

[7] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019, p. 6558, 2019.

[8] X. Zhang, X. Wei, Z. Zhou, Q. Zhao, S. Zhang, Y. Yang, R. Li, and B. Hu, "Dynamic alignment and fusion of multimodal physiological patterns for stress recognition," *IEEE Transactions on Affective Computing*, vol. 15, no. 2, pp. 685–696, 2023.

[9] X. Zhou and X. Peng, "Multi-scale spatiotemporal representation learning for EEG-based emotion recognition," *arXiv preprint arXiv:2409.07589*, 2024.

[10] Z. Miao and M. Zhao, "Time–space–frequency feature fusion for 3-channel motor imagery classification," *Biomedical Signal Processing and Control*, vol. 90, p. 105867, 2024.

[11] H. Cai, X. Zhang, Y. Zhang, Z. Wang, and B. Hu, "A case-based reasoning model for depression based on three-electrode eeg data," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 383–392, 2018.