Supplementary Materials

APPENDIX A
PROOF OF THE COMPUTATION P

For the total c emotion classes, the supervised prototypical representation P is computed as

$$P = \left(Y^T Y\right)^{-1} Y^T f\left(X\right), \tag{S1}$$

where $X = [x_1; x_2, ...; x_n]$ and $Y = [y_1; y_2, ...; y_n]$ are the labeled source data and the corresponding emotional labels (one-hot representation) in the current mini-batch, respectively. $f\left(\cdot\right)$ is a feature extractor with the parameter θ_f. $f\left(X\right)$ is a $N_B \times M$ feature matrix, where N_B is the batch size for stochastic gradient descent and M is the feature dimensionality. $P = [\mu_1; \mu_2; ...; \mu_c]$, where μ_c is the prototypical representation for the cth emotional category.

Proof.

$$\begin{aligned}
P &= [\mu_1; \mu_2; ...; \mu_c] \\
&= \left[\frac{1}{N_1}\sum_{i_1}^{N_1} f(x_{i_1}); \frac{1}{N_2}\sum_{i_1}^{N_2} f(x_{i_2}); ...; \frac{1}{N_c}\sum_{i_c}^{N_c} f(x_{i_c})\right] \\
&= \begin{bmatrix} \frac{1}{N_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{N_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{N_c} \end{bmatrix} \left[\sum_{i_1}^{N_1} f(x_{i_1}); ...; \sum_{i_c}^{N_c} f(x_{i_c})\right] \\
&= \begin{bmatrix} N_1 & 0 & \cdots & 0 \\ 0 & N_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & N_c \end{bmatrix}^{-1} Y^T f(X) \\
&= \left(Y^T Y\right)^{-1} Y^T f\left(X\right),
\end{aligned}$$

where N_j ($j \in [1, ..., c]$) is the sample size in the jth emotion class. x_{i_j} are the EEG samples belonging to the jth emotion class.

APPENDIX B
PROOF OF THEOREM 1

Theorem 1: *Based on the labeled source domain ($\mathbb{S}$), the unlabeled source domain ($\mathbb{U}$), and the target domain ($\mathbb{T}$), the target error $\epsilon_T(h)$ with a hypothesis h ($h \in \mathcal{H}$) is bounded as*

$$\begin{aligned}
\epsilon_T(h) \leq \epsilon_S(h) &+ \min\left\{\mathbb{E}_{\overline{\mathcal{D}}_T}\left[|F_{\overline{T}} - F_T|\right], \mathbb{E}_{\mathcal{D}_T}\left[|F_T - F_{\overline{T}}|\right]\right\} \\
&+ \pi_U(d_{\mathcal{H}}\left(\mathcal{D}_S, \mathcal{D}_U\right) + d_{\mathcal{H}}\left(\mathcal{D}_U, \mathcal{D}_T\right)) + \pi_S d_{\mathcal{H}}\left(\mathcal{D}_S, \mathcal{D}_T\right) \\
&+ \pi_U \min\left\{\mathbb{E}_{\mathcal{D}_S}\left[|F_S - F_U|\right], \mathbb{E}_{\mathcal{D}_U}\left[|F_U - F_S|\right]\right\},
\end{aligned}$$

where $\epsilon_S(h)$ is the error of the labeled source domain, and $d_{\mathcal{H}}$ represents the $\mathcal{H}$-divergence between the given domains. $F_S(x)$, $F_U(x)$, and $F_T(x)$ are the labeling functions of the labeled source domain, the unlabeled source domain, and the target domain, respectively. $F_{\overline{T}}(x) = \pi_S F_S(x) + \pi_U F_U(x)$ is a labeling function for any $x \in Supp(\overline{D}_T)$, which is a weighted summation of $F_S(x)$ and $F_U(x)$ with the weights of π_S and π_U.

Proof. In the standard domain adaptation, the target error $\epsilon_{T(h)}$ for a given classification hypothesis $h \in \mathcal{H}$ is bounded by [1]

$$\epsilon_T(h) \leq \epsilon_S(h) + d_{\mathcal{H}}\left(\mathcal{D}_S, \mathcal{D}_T\right) + \delta, \tag{S2}$$

where $\mathcal{D}_S$ and $\mathcal{D}_T$ refer to the labeled source domain and the target domain, respectively. $\delta = \min\left\{\mathbb{E}_{\mathcal{D}_S}\left[|F_S - F_T|\right], \mathbb{E}_{\mathcal{D}_T}\left[|F_T - F_S|\right]\right\}$. However, under the setting of this study, the source domain is composed of both labeled and unlabeled data. If simply applying the standard domain adaptation in this study by only using the labeled source data, the estimation of the target error would not be accurate enough.

Motivated by the domain projection method introduced in the domain generalization studies [2], [3], we can project the target domain $\mathcal{D}_T$ onto the convex hull of the source and compute its "projection" as Eq. S3. Based on the given labeled source domain ($\mathcal{D}_S$), the unlabeled source domain ($\mathcal{D}_U$), and the target domain ($\mathcal{D}_T$), the computation is given as

$$\mathcal{D}_{\overline{T}} = \text{argmin}_{\pi_S, \pi_U}\, d_{\mathcal{H}}\left[\mathcal{D}_T, \pi_S\mathcal{D}_S + \pi_U\mathcal{D}_U\right], \tag{S3}$$

where π_S and π_U are belong to Δ_1 and $\pi_S + \pi_U$ is equal to 1. The labeling function of $\overline{\mathcal{D}}_T$ could be represented by a weighted summation of the labeling functions of $\mathcal{D}_S$ and $\mathcal{D}_U$, as

$$f_{\overline{T}}(x) = \pi_S F_S(x) + \pi_U F_U(x), \tag{S4}$$

where $F_S(x)$ and $F_U(x)$ are the corresponding labeling functions of $\mathcal{D}_S$ and $\mathcal{D}_U$. Then, the target error is bounded as

$$\epsilon_T(h) \leq \delta_{\overline{T},T} + \epsilon_{\overline{T}}(h) + d_{\mathcal{H}}\left(\mathcal{D}_{\overline{T}}, \mathcal{D}_T\right), \tag{S5}$$

where $\delta_{\overline{T},T} = \min\left\{\mathbb{E}_{\mathcal{D}_{\overline{T}}}\left[|F_{\overline{T}} - F_T|\right], \mathbb{E}_{\mathcal{D}_T}\left[|F_T - F_{\overline{T}}|\right]\right\}$. The overall source error $\epsilon_{\overline{T}}(h)$ is defined as

$$\epsilon_{\overline{T}}(h) = \pi_S \epsilon_S(h) + \pi_U \epsilon_U(h), \tag{S6}$$

which is a weighted summation of the labeled source error of $\epsilon_S(h)$ and the unlabeled source error of $\epsilon_U(h)$. Considering that the label information of the unlabeled source domain $\mathcal{D}_U$ is unknown in the training process, we can bound the error $\epsilon_U(h)$ as

$$\epsilon_U(h) \leq \delta_{S,U} + \epsilon_S(h) + d_{\mathcal{H}}\left(\mathcal{D}_S, \mathcal{D}_U\right), \tag{S7}$$

where $\delta_{S,U} = \min\left\{\mathbb{E}_{\mathcal{D}_S}\left[|F_S - F_U|\right], \mathbb{E}_{\mathcal{D}_U}\left[|F_U - F_S|\right]\right\}$. The $\mathcal{H}$-divergence between $\mathcal{D}_{\overline{T}}$ and $\mathcal{D}_T$ is given as

$$\begin{aligned} d_{\mathcal{H}}\left(\mathcal{D}_{\overline{T}}, \mathcal{D}_T\right) = 2 \sup_{h \in \mathcal{H}} |\, &\Pr_{x \sim \mathcal{D}_{\overline{T}}}[h(x) = 1] \\ &- \Pr_{x \sim \mathcal{D}_T}[h(x) = 1]|. \end{aligned} \tag{S8}$$

According to the triangle inequality and the sub-additivity of the sup, the upper-bound of the $\mathcal{H}$-divergence can be written as [2], [3]

$$d_{\mathcal{H}}\left(\overline{\mathcal{D}}_T, \mathcal{D}_T\right) \leq \pi_S d_{\mathcal{H}}\left(\mathcal{D}_S, \mathcal{D}_T\right) + \pi_U d_{\mathcal{H}}\left(\mathcal{D}_U, \mathcal{D}_T\right). \tag{S9}$$

Through replacing the $\epsilon_{\overline{T}}(h)$ with $\pi_S \epsilon_S(h) + \pi_U \epsilon_U(h)$ and replacing the upper-bound $d_{\mathcal{H}}\left(\overline{\mathcal{D}}_T, \mathcal{D}_T\right)$ with $\pi_S d_{\mathcal{H}}\left(\mathcal{D}_S, \mathcal{D}_T\right) + \pi_U d_{\mathcal{H}}\left(\mathcal{D}_U, \mathcal{D}_T\right)$ in Eq. S5, the proof completes.

APPENDIX C
DETAILS ABOUT THE EEG-MIXUP AUGMENTATION

In this section, we provide a detailed description of the EEG-Mixup augmentation. The main idea of the EEG-Mixup is to fit the original EEG data in accordance with the independent and identically distributed (IID) assumption and then apply mixup augmentation [4] to generate augmented data. Specifically, we first sample EEG signal x_n and label y_n from the same trial of the same subject, which follows the joint distribution. As shown in Fig. S1 (a), we randomly select two samples from one trial and then generate a new sample ($x_k^{'}$ and $y_k^{'}$) using the Mixup augmentation. Note that for the trials without labels (domains $\mathbb{U}$ and $\mathbb{T}$), we only use the EEG signal for augmentation, which is presented in Fig. S1 (b). To illustrate the EEG-Mixup from an intuitive perspective, we also provide the corresponding pseudo-code in Algorithm S1.

APPENDIX D
EFFECT OF THE EEG BANDS USED FOR MODEL TRAINING

In this section, we report the performance of the proposed EEGMatch in the individual Delta (δ), Theta (θ), Alpha (α), Beta (β), and Gamma (γ) bands to investigate their impacts on model performance. For simplicity, we conduct our experiments on the SEED database and set the number of labeled trials as three for each source subject ($N = 3$). Table S1 reports the model performance on 15 target subjects using features extracted from various combinations of EEG frequency bands. The results show that the Beta and Gamma bands achieve better recognition results than the other three bands in most subjects. It confirms our findings that Beta and Gamma bands contain the major information for emotion recognition. Furthermore, it is evident that the model trained using all EEG bands demonstrates superior performance when compared to other combination strategies. This phenomenon suggests that a fusion of all EEG bands is essential for achieving optimal performance in our model.

APPENDIX E
EFFECT OF THE NUMBER OF TRAINABLE PARAMETERS

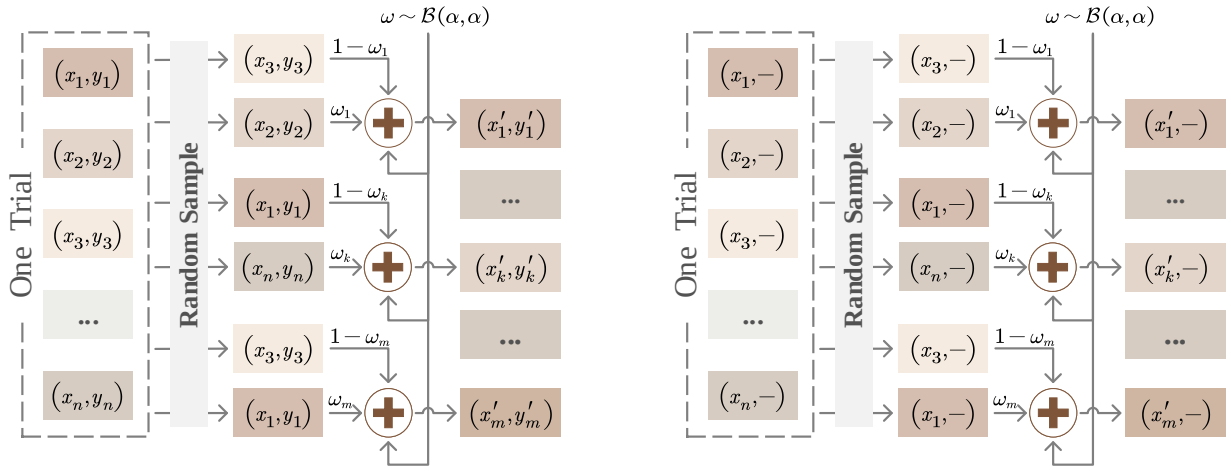
The number of trainable parameters is an important factor that determines the model computation complexity and performance. The computation complexity of the EEGMatch can be estimated by computing the number of trainable parameters in its feature extractor $f(\cdot)$, domain discriminator $d(\cdot)$, and bilinear transformation matrix B. For the feature extractor $f(\cdot)$, it is designed as 310 neurons-H neurons-Relu activation-H neurons-Relu activation-H neurons. For the domain discriminator $d(\cdot)$, it is designed as H neurons-H neurons-Relu activation-dropout layer-H neurons-3 neurons-Softmax activation. The size of matrix B is $H \times H$. H refers to the number of neurons in the hidden layers. Consequently, the number of trainable parameters Num in the EEGMatch can be computed as,

$$Num = (310H + H^2) + (H^2 + 3H) + H^2 = 313H + 3H^2, \tag{S10}$$

The higher the number of trainable parameters, the higher the computation complexity. To investigate its effect on the model performance, we adjust the value of H and report the performance of the EEGMatch in Table S2. The experiments are conducted on the SEED database with the three labeled trials ($N = 3$). The results indicate that the optimal H value is 64, which achieves the best classification performance. Additionally, it demonstrates the second-lowest computation complexity, indicating a favorable balance between performance and complexity.

APPENDIX F
SENSITIVITY ANALYSIS OF THE DYNAMIC THRESHOLDS

To ensure correct optimization during model training, we employ two dynamic thresholds (upper threshold τ_h and lower threshold τ_l) to exclude invalid paired samples. In this section, we use a grid-search method to investigate the impact of initial values of the thresholds on model performance. Specifically, we fix the initial upper threshold τ_h^0 and adjust the initial lower threshold τ_l^0 from 0.1 to 0.5 with a step of 0.1. For simplicity, we conduct our experiments on the SEED database and set the number of labeled trials as three for each source subject ($N = 3$). As



(a) EEG-Mixup on the trials with labels (domain $\mathbb{S}$)

(b) EEG-Mixup on the trials without labels (domains $\mathbb{U}$ and $\mathbb{T}$)

Fig. S1: Diagram of the proposed EEG-Mixup augmentation

TABLE S1: The mean accuracies (%) of the proposed EEGMatch on the SEED database using varying EEG bands.

Band	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	Average
δ band	58.99	55.24	50.53	70.95	77.22	66.29	77.34	65.20	64.29	78.43	72.16	61.91	58.37	71.83	67.01	66.38±11.28
θ band	79.29	74.57	44.41	64.26	50.79	47.58	39.54	54.51	59.51	65.35	53.74	49.21	55.72	43.11	67.74	56.62±08.11
α band	55.12	44.76	50.23	48.79	68.01	70.18	32.59	61.46	72.15	52.51	56.71	52.44	50.77	55.36	66.41	55.83±10.19
β band	72.16	64.11	58.45	68.01	67.12	69.98	68.23	66.38	73.16	68.79	83.53	60.28	82.41	68.01	95.79	71.09±09.29
γ band	85.33	60.81	54.39	50.24	67.94	75.78	59.87	85.56	68.89	71.33	91.25	73.48	79.49	85.68	78.11	72.54±11.85
β,γ bands	83.74	77.78	56.12	90.39	72.86	76.93	87.80	85.94	86.38	82.85	86.26	85.59	80.49	75.31	84.29	80.85±08.18
β,γ,δ bands	90.07	87.95	78.69	61.22	85.21	92.16	83.82	83.44	81.56	93.13	79.19	73.36	76.75	100	100	84.43±09.85
β,γ,θ bands	83.32	76.90	64.70	74.19	87.24	66.15	88.16	87.95	81.08	85.65	93.13	100	73.78	94.05	93.93	83.35±10.10
β,γ,α bands	90.75	63.79	60.81	80.49	71.59	78.43	66.00	83.79	70.53	83.12	92.78	93.93	83.29	85.21	93.08	79.84±10.63
All bands	100.00	79.76	80.55	78.26	78.64	88.42	96.94	71.15	90.48	88.80	93.37	83.35	90.78	95.93	83.82	**86.68±07.89**

TABLE S2: The mean accuracies (%) of the proposed EEGMatch on the SEED database.

Complexity	sub1	sub2	sub3	sub4	sub5	sub6	sub7	sub8	sub9	sub10	sub11	sub12	sub13	sub14	sub15	Average
$H = 32$	93.48	93.75	76.43	69.79	93.69	80.64	81.67	67.77	85.98	91.96	97.26	100	74.89	90.71	100	86.54±10.34
$H = 64$	100.00	79.76	80.55	78.26	78.64	88.42	96.94	71.15	90.48	88.80	93.37	83.35	90.78	95.93	83.82	**86.68±07.89**
$H = 128$	96.17	78.02	87.09	64.82	77.34	94.55	84.86	89.09	94.76	83.00	90.75	91.84	78.79	93.28	90.93	86.35±08.35
$H = 256$	90.13	83.26	80.58	82.79	94.19	86.18	83.79	85.27	85.53	88.51	90.60	74.72	71.36	90.60	89.22	85.12±05.92

illustrated in Fig. S2, we can observe that the performance of the proposed EEGMatch is relatively insensitive to the initialization of the dynamic threshold. On the other hand, $\{\tau_h^0, \tau_l^0\} = \{0.9, 0.5\}$ and $\{\tau_h^0, \tau_l^0\} = \{0.6, 0.5\}$ are two optimal settings that can maximize the model performance. Considering that a high upper threshold can reduce the confirmation bias [5], [6], we initialize the upper threshold as $\tau_h^0 = 0.9$ and the lower threshold as $\tau_l^0 = 0.5$ by default.

Appendix G
Effect of Different Learning Rates

In this section, we examine the effect of learning rates on the performance of the proposed EEGMatch. Specifically, we adjust the learning rates in the range of 0.01 to 0.0001 and evaluate the model's performance on the SEED database. Note that the number of labeled trials is set as twelve for each source subject ($N = 12$). The accuracy of the EEGMatch on the target domain is recorded during the training process and illustrated in Fig. S3. The result demonstrates that small learning rates (0.001 and 0.0001) can stabilize the domain adversarial training and improve the model performance on the target domain. However,

Algorithm S1 The training algorithm of EEGMatch

Input:

- emotion class c, max iteration $Maxepoch$, batch size N_B, initial threshold values τ_h^0 and τ_l^0;
- labeled source data $\{X_s, Y_s\}$, unlabeled source data $\{X_u\}$, unknown target data $\{X_t\}$;
- random initialization of prototype representation P_{s^*}, bilinear transformation matrix B, feature extractor $f(\cdot)$, and discriminator $d(\cdot)$;

Output: P_{s^*}, B, $f(\cdot)$, $d(\cdot)$;

##All the equations used below can be found in Section III of the main manuscript.##

EEG-Mixup based data augmentation

generate the augmented data $\{\mathcal{A}(X_s, Y_s)\}$, $\{\mathcal{A}(X_u)\}$, $\{\mathcal{A}(X_t)\}$ using Eq. (1) and (2);
update $\{X_s, Y_s\}$ by concatenating $\{X_s, Y_s\}$ and the augmented data $\{\mathcal{A}(X_s, Y_s)\}$;
update $\{X_u\}$ by concatenating $\{X_u\}$ and the augmented data $\{\mathcal{A}(X_u)\}$;
update $\{X_t\}$ by concatenating $\{X_t\}$ and the augmented data $\{\mathcal{A}(X_t)\}$;
for 1 to $Maxepoch$ **do**
  for b = 1 to N_B **do**
    sample labeled source domain data $\{X_s^b, X_s^b\}$ from the updated $\{X_s, Y_s\}$;
    sample unlabeled source domain data $\{X_u^b\}$ from the updated $\{X_u\}$;
    sample unlabeled target domain data $\{X_t^b\}$ from the updated $\{X_t\}$;

semi-supervised two-step pairwise learning

$Cls\left(f\left(X_u^b\right)\right) = softmax\left(f\left(X_u^b\right)^T BP_{s^*}\right)$;
$\widehat{Y}_u^b = Sharpen\left(Cls\left(f\left(X_u^b\right)\right)\right)$; // pseudo labels of $\mathbb{U}$
update $X^{s^*} = \left[X_s^b; X_u^b\right]$, $Y^{s^*} = \left[Y_s^b; \hat{Y}_u^b\right]$;
calculate $P_{s^*} = \left((Y^{s^*})^T(Y^{s^*})\right)^{-1}(Y^{s^*})^T f\left(X^{s^*}\right)$; // prototypical representation
compute R_p^s with X^{s^*} using Eq. (12) and compute R_p^t with X_t^b using Eq. (15); // instance-to-instance similarity matrix
compute R^s with Y^{s^*} using Eq. (13) and compute R^t using Eq. (16); // groundtruth similarity matrix
compute $\mathcal{L}_{pair}^s$ using Eq. (14), and compute $\mathcal{L}_{pair}^t$ using Eq. (17) and (18);
compute $\mathcal{L}_{pair}$ using Eq. (20); // instance-wise pairwise learning loss computation based on $\mathcal{L}_{pair}^s$ and $\mathcal{L}_{pair}^t$
update τ_h and τ_l using Eq. (19); // dynamic thresholds

semi-supervised multi-domain adaptation

compute $\mathcal{L}_{disc}$ using Eq. (25); // semi-supervised multi-domain adaptation loss computation
compute $\mathcal{L}$ using Eq. (26); // the final objective loss computation based on $\mathcal{L}_{pair}$ and $\mathcal{L}_{disc}$
gradient back-propagation;
update network parameters;
  end for
end for

we also observe that an excessively small learning rate of 0.0001 would hinder the model's ability to find the optimal parameters, resulting in decreased performance. Consequently, we set the default learning rate of the EEGMatch as 0.001 to jointly consider both optimal model performance and training stability.

APPENDIX H
EFFECT OF THE NUMBER OF SOURCE SUBJECTS

The number of source subjects used for transfer learning is an important factor for model performance [7], [8]. In this section, we explore the impact of the number of source subjects, denoted as M_s, on the performance of our proposed EEGMatch. Specifically, we adjust the number of source subjects M_s from 1 to 13 with a step of 2 and evaluate the performance of the proposed EEGMatch on the SEED database. In order to ensure that the number of source domain samples is not much smaller than the target domain samples, we set the number of labeled trials as twelve for each source subject. The mean performance across 15 target subjects of the SEED database is presented in Fig. S4. The results show that the model performance remains stable when $M_s \geq 3$, which indicates its robustness across a broader range of source subject quantities. On the other hand, we can determine the minimum number of subjects needed for the model to achieve a specified level of performance. For example, achieving an accuracy exceeding 85% requires a minimum of 3 subjects, while surpassing an accuracy rate of 90% necessitates a minimum of 9 subjects."

Algorithm S2 The pseudo-code of EEG-Mixup

Input:
- labeled source data $\{X_s, Y_s\}$, unlabeled source data $\{X_u\}$, target data $\{X_t\}$;
- The number of source domain subjects Sub, the number of labeled trials in the source domain N, the number of unlabeled trials in the source domain $M - N$;
- The number of target domain subjects Sub_t, the number of unlabeled trials in the target domain M;
- Augmentation ratio $ratio$;

Output: augmented data $\{\mathcal{A}(X_s, Y_s)\}$, $\{\mathcal{A}(X_u)\}$, $\{\mathcal{A}(X_t)\}$;

EEG-Mixup on the labeled source domain
for p_s = 1 to Sub **do**
  for q_s = 1 to N **do**
    collect $D^s_{p_s,q_s} = \{(x^s_n, y^s_n)\}^{N^s_{p_s,q_s}}_{n=1}$ from $\{X_s, Y_s\}$;
    $//N^s_{p_s,q_s}$ is the number of samples from p_s subject at q_s trial.
    for 1 to $ratio \times N^s_{p_s,q_s}$ **do**
      randomly select two samples (x^s_i, y^s_i) and (x^s_j, y^s_j);
      sample ω from distribution $\mathrm{Beta}(\alpha, \alpha)$
      generate new sample by $x^s_z = \omega x^s_i + (1-\omega)x^s_j$, $y^s_z = \omega y^s_i + (1-\omega)y^s_j$;
      add (x^s_z, y^s_z) to $\mathcal{A}(X_s, Y_s)$
    end for
  end for
end for
EEG-Mixup on the unlabeled source domain
for p_u = 1 to Sub **do**
  for q_u = 1 to $M - N$ **do**
    collect $D^u_{p_u,q_u} = \{(x^u_n, y^u_n)\}^{N^u_{p_u,q_u}}_{n=1}$ from $\{X_u\}$;
    $//N^u_{p_u,q_u}$ is the number of samples from p_u subject at q_u trial.
    for 1 to $ratio \times N^u_{p_u,q_u}$ **do**
      randomly select two samples x^u_i and x^u_j;
      sample ω from distribution $\mathrm{Beta}(\alpha, \alpha)$
      generate new sample by $x^u_z = \omega x^u_i + (1-\omega)x^u_j$;
      add x^u_z to $\mathcal{A}(X_u)$
    end for
  end for
end for
EEG-Mixup on the target domain
for p_t = 1 to Sub_t **do**
  for q_t = 1 to M **do**
    collect $D^t_{p_t,q_t} = \{(x^t_n, y^t_n)\}^{N^t_{p_t,q_t}}_{n=1}$ from $\{X_t\}$;
    $//N^t_{p_t,q_t}$ is the number of samples from p_t subject at q_t trial.
    for 1 to $ratio \times N^t_{p_t,q_t}$ **do**
      randomly select two samples x^t_i and x^t_j;
      sample ω from distribution $\mathrm{Beta}(\alpha, \alpha)$
      generate new sample by $x^t_z = \omega x^t_i + (1-\omega)x^t_j$;
      add x^t_z to $\mathcal{A}(X_t)$
    end for
  end for
end for

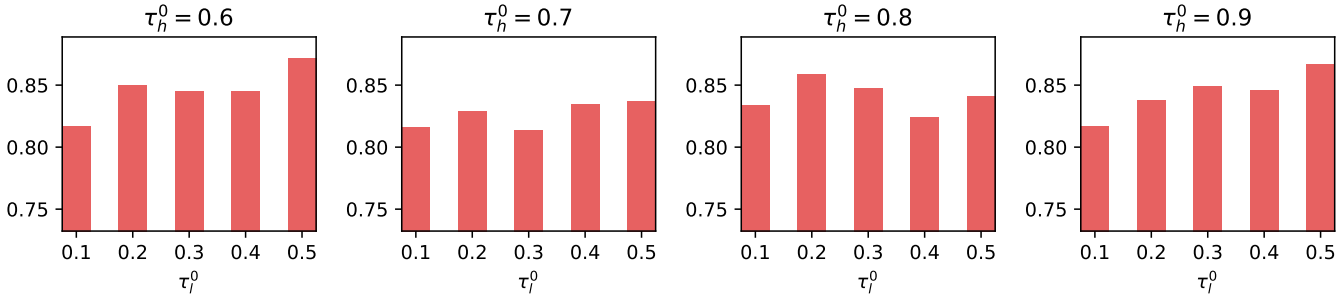
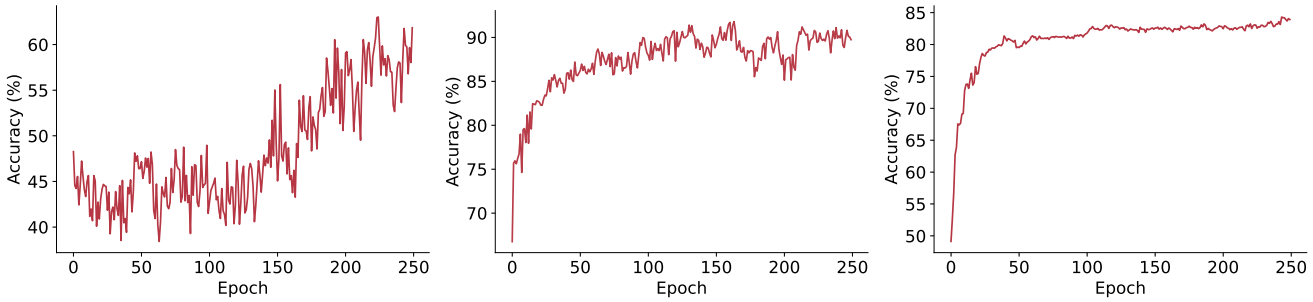


Fig. S2: The mean accuracies (%) of the proposed EEGMatch on target domains under different initial thresholds.



(a) learning rate = 0.01 (b) learning rate = 0.001 (c) learning rate = 0.0001

Fig. S3: The mean accuracies (%) of the proposed EEGMatch under different learning rates.

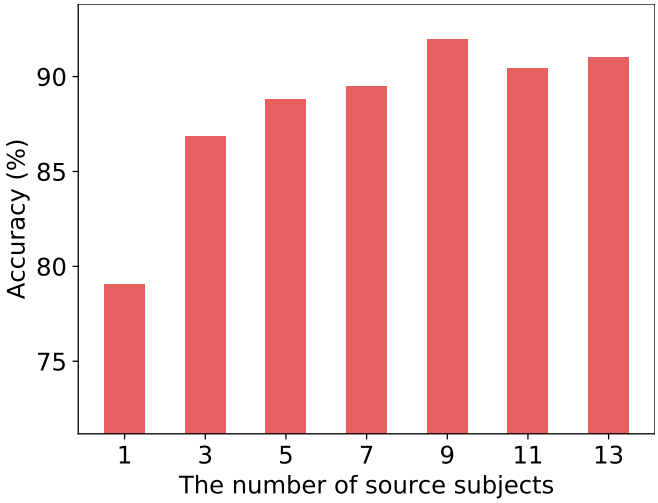


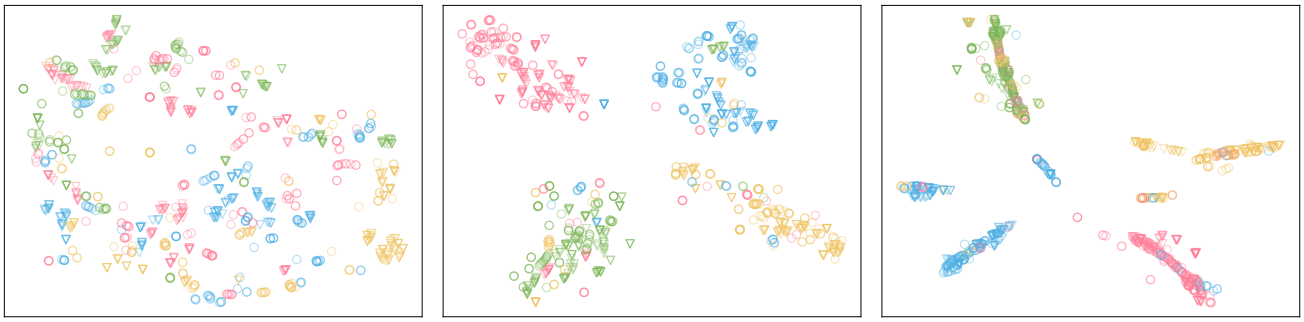
Fig. S4: The mean accuracies (%) of the proposed EEGMatch under different number of source subjects.

Appendix I
Feature visualization on the SEED-IV and SEED-V databases

We employ the t-SNE algorithm [9] to visualize the learned feature representations from the SEED-IV and SEED-V databases. The results are illustrated in Fig. S5 and Fig. S6. To enhance clarity, we assign different colors to the samples with different emotions and assign different shapes to distinguish samples from different domains. The results on the two databases demonstrate that EEGMatch reduces the distribution discrepancy among the three domains throughout the training process. Additionally, it progressively separates the samples with different emotions to minimize the emotion classification error. In summary, the results provide supplementary evidence to support the efficacy of the proposed EEGMatch in EEG-based emotion recognition under a semi-supervised learning framework.

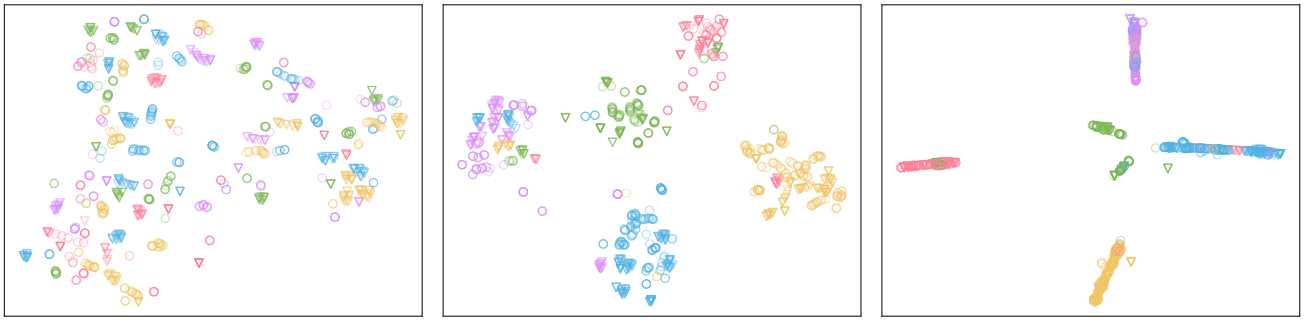
Appendix J
Results on the real-world database.

To further validate the feasibility of the proposed EEGMatch, we extend the application of EEGMatch on a real-world clinical EEG database. This database was obtained from the Hospital Universiti Sains Malaysia (HUSM), featuring patients diagnosed with depression [10]. It includes 27 healthy subjects (38.28±15.64 years old) and 29 subjects diagnosed with Major Depressive Disorder (MDD) (40.33±12.86 years old). Their categorization adheres to the international diagnostic criteria for depression outlined in the Diagnostic and Statistical Manual-IV (DSM-IV). The EEG data was collected using the Brain Master Discovery amplifier, with 19 EEG electrodes and a sampling rate of 256Hz. In the preprocessing, bandpass filtering ranging from 0.5 to 70 Hz and a notch filter at 50 Hz were conducted. The experimental protocol involved recording EEG signals from participants during both



(a) (b) (c)

Fig. S5: A visualization of the learned feature representations (a) before training, (b) at the training epoch of 50, and (c) in the final model. Here, the circle, asterisk and triangle represent the labeled source domain ($\mathbb{S}$), the unlabeled source domain ($\mathbb{U}$), and the target domain ($\mathbb{T}$). The pink, blue, orange, and green colors indicate happy, sad, fear, and neural emotions. The number of labeled trials is set as $N = 20$.



(a) (b) (c)

Fig. S6: A visualization of the learned feature representations (a) before training, (b) at the training epoch of 50, and (c) in the final model. Here, the circle, asterisk and triangle represent the labeled source domain ($\mathbb{S}$), the unlabeled source domain ($\mathbb{U}$), and the target domain ($\mathbb{T}$). The pink, blue, orange, purple, and green colors indicate happy, sad, neural, fear, and disgust emotions. The number of labeled trials is set as $N = 10$.

closed-eye (EC) and open-eye (EO) conditions, with each session lasting 5 minutes. During the open-eye condition, participants were instructed to relax while minimizing eye movements to ensure the quality of the collected data.

In this study, we employ a cross-subject validation protocol to evaluate the performance of the proposed EEGMatch in handling real-world clinical data. To ensure a robust assessment, we implement a ten-fold cross-validation strategy to compute mean performance and standard deviations. In each validation round, 9 subsets are used as training data, while the remaining 1 subset serves as testing data. This process is iterated 10 times, ensuring that each subset is used as testing data once. To evaluate the model performance under label scarcity conditions, we divided the training subsets into $\mathbb{S}$ and $\mathbb{U}$ domains based on a specified ratio. For instance, if the $\mathbb{S} : \mathbb{U}$ ratio is $1 : 1$, 50% of subjects in the training sets are labeled ($\mathbb{S}$) while the remaining 50% are unlabeled ($\mathbb{U}$). We adjust the ratio from $1 : 2$ to $2 : 1$ and present the results in Table S3. Here, the sample balance of the two groups in three domains is considered. The results show that EEGMatch attains an average accuracy of 65.35±10.07 in the EC experiment, with a 1.33% improvement compared to the existing literature. In the EO experiment, EEGMatch achieves an average accuracy of 67.08±09.24, indicating a 2.39% enhancement compared to other models. In summary, the results demonstrate the efficacy of EEGMatch in alleviating the label-missing problem in a real-world database.

References

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1, pp. 151–175, 2010.

[2] I. Albuquerque, J. Monteiro, M. Darvishi, T. H. Falk, and I. Mitliagkas, "Generalizing to unseen domains via distribution matching," *arXiv preprint arXiv:1911.00804*, 2019.

[3] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," *Advances in neural information processing systems*, vol. 31, 2018.

TABLE S3: Experimental results on the MDD database using cross-subject validation protocol with incomplete labels. EC: closed-eye; EO: open-eye.

Session	$\mathbb{S}:\mathbb{U}$	MixMatch* [11]	AdaMatch* [12]	FlexMatch* [13]	SoftMatch* [14]	**EEGMatch**
EC	2 : 1	66.56±06.81	66.60±08.72	66.43±06.50	66.63±03.98	**71.32±08.16**
	1 : 1	**63.48±07.06**	62.57±09.34	62.12±06.73	62.24±06.18	61.85±10.98
	1 : 2	62.01±07.52	61.11±09.61	61.74±07.06	60.19±09.35	**62.88±07.94**
Average Performance		64.02±07.38	63.43±09.52	63.43±07.09	63.02±07.37	**65.35±10.07**
EO	2 : 1	65.47±07.81	64.54±08.27	61.48±07.42	61.18±06.99	**67.71±08.48**
	1 : 1	65.65±04.83	64.28±06.05	63.52±09.36	63.64±09.38	**66.25±09.94**
	1 : 2	62.95±06.93	61.70±08.06	63.83±09.53	63.28±09.18	**67.27±09.18**
Average Performance		64.69±06.76	63.51±07.63	62.94±08.88	62.70±08.65	**67.08±09.24**

[4] H. Zhang, C. Moustapha, D. Yann N., and L.-P. David, "Mixup: Beyond empirical risk minimization," *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[5] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "FixMatch: simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[6] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[7] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, "Domain adaptation for EEG emotion recognition based on latent representation similarity," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 344–353, 2020.

[8] Y. Peng, W. Wang, W. Kong, F. Nie, B.-L. Lu, and A. Cichocki, "Joint feature adaptation and graph adaptive label propagation for cross-subject emotion recognition from EEG signals," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1941–1958, 2022.

[9] V. D. M. Laurens and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2605, pp. 2579–2605, 2008.

[10] W. Mumtaz, "MDD patients and healthy controls EEG data (new). figshare," *Dataset. MDD Patients and Healthy Controls EEG Data generated by https://doi. org/10.6084/m9. figshare*, vol. 4244171, p. v2, 2016.

[11] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "MixMatch: a holistic approach to semi-supervised learning," *arXiv preprint arXiv:1905.02249*, 2019.

[12] D. Berthelot, R. Roelofs, K. Sohn, N. Carlini, and A. Kurakin, "AdaMatch: a unified approach to semi-supervised learning and domain adaptation," *arXiv preprint arXiv:2106.04732*, 2021.

[13] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 408–18 419, 2021.

[14] H. Chen, R. Tao, Y. Fan, Y. Wang, J. Wang, B. Schiele, X. Xie, B. Raj, and M. Savvides, "SoftMatch: Addressing the quantity-quality trade-off in semi-supervised learning," *arXiv preprint arXiv:2301.10921*, 2023.