Pairwise Representation Alignment for Semi-Supervised EEG-based Emotion Recognition

Guangyi Zhang and Ali Etemad

Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada
{guangyi.zhang, ali.etemad}@queensu.ca

Abstract

We propose a novel semi-supervised architecture for learning strong EEG representations for emotion recognition. Successive to weak and strong data augmentations, our model performs label guessing for large amounts of original and augmented unlabeled data. Next, we apply a convex combination of unlabeled and labeled data. Following, to reduce the potential distribution mismatch between the large amounts of unlabeled data and the limited amount of labeled data, we perform pairwise embedding alignment. We adapt several state-of-the-art semi-supervised approaches for EEG learning, with which we compare our proposed architecture. We perform rigorous experiments with all 10 methods on two public EEG-based emotion recognition datasets, SEED and SEED-IV. The experiments show that our proposed framework achieves overall best results with varying amounts of limited labeled samples.

1 Introduction

Human emotions are highly informative non-verbal cues that can be widely used to enhance human-machine interaction. Many solutions have been proposed for affective computing (Picard 2000) using Electroencephalography (EEG), as EEG is widely used for directly measuring brain activity with high spatio-temporal resolution. Recently, various deep learning models have achieved state-of-the-art performances in emotion recognition tasks due to the capability of learning highly discriminative information from multichannel EEG recordings (Zheng and Lu 2015; Zheng et al. 2018). However, as the significant majority of existing EEG-based deep emotion recognition solutions are '*supervised*' methods (Zheng et al. 2018; Zhang and Etemad 2021a, 2020; Zheng, Zhu, and Lu 2017; Zhang et al. 2018b), they rely on output labels. Labeling EEG, on the other hand, is difficult, time-consuming, costly, and often requires professional annotators (Zheng and Lu 2015; Zheng et al. 2018). As a result, it is important to investigate how to develop effective solutions that can deal with limited labeled training samples.

Semi-Supervised Learning (SSL) is a powerful paradigm to address the problems of the scarcity of labeled training samples, achieving great success in the field of computer vision (Van Engelen and Hoos 2020). For example, pseudo-labeling was proposed to encourage low-entropy predictions of unlabeled data (Lee et al. 2013). To do so, a model was first trained using limited labeled data. The model was then used to predict pseudo-labels for the unlabeled data. Finally, the model was retrained using the entire data with true labels and pseudo labels combined (Lee et al. 2013). Consistency regularization, in conjunction with data augmentation, has lately become popular in SSL studies (Samuli and Timo 2017; Tarvainen and Valpola 2017; Oliver et al. 2018). Common regularization techniques including stochastic augmentation applied to input data and dropout applied throughout the network have been employed in recent SSL frameworks (Samuli and Timo 2017). For instance, the Π-model (Samuli and Timo 2017) trained the network (with dropout) on both the original and augmented inputs, and minimized the distance between their corresponding outputs, as an unsupervised loss. Meanwhile, a supervised cross-entropy loss was only computed for the labeled set. To improve the Π-model, 'temporal ensembling' and 'mean teacher' methods further relied on consistency regularization techniques for SSL (Samuli and Timo 2017; Tarvainen and Valpola 2017).

In addition to the classical SSL approaches mentioned above, ADA-Net (Wang, Li, and Gool 2019) was proposed to minimize the potential distribution mismatch between large unlabeled and few labeled data through representation alignment. MixMatch (Berthelot et al. 2019) was proposed to first minimize the entropy of pseudo-labels for unlabeled data under various augmentations. It then increased the size of the training data by combining labeled and unlabeled data (Zhang et al. 2018a). FixMatch (Sohn et al. 2020) was proposed to simplify the consistency regularization using cross-entropy along with a confidence threshold. AdaMatch (Berthelot et al. 2021) was proposed to improve upon FixMatch by aligning the class distribution between unlabeled and labeled data. ADA-Net, MixMatch, FixMatch and AdaMatch achieved state-of-the-art results in the field of computer vision. However, their use has been mostly limited to computer vision tasks, and has therefore not been explored in other domains such as brain-computer interfaces (BCI) or biological signal processing. Only in a recent paper (Zhang and Etemad 2021b), few of the classical SSL frameworks were used for BCI with EEG representation learning.

In this paper, we propose a novel semi-supervised framework for EEG-based emotion recognition to reduce the distribution mismatch between large unlabeled and few labeled data. To achieve this, our model first augments EEG data, and then combines the labeled and unlabeled samples into

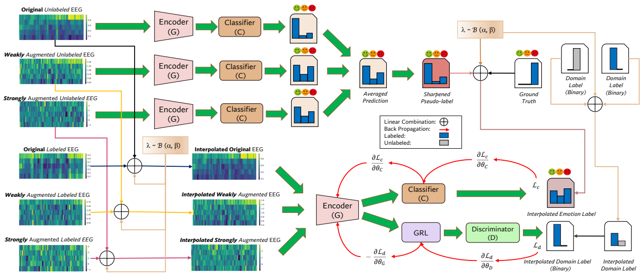
Original Unlabeled EEG
Weakly Augmented Unlabeled EEG
Strongly Augmented Unlabeled EEG
Original Labeled EEG
Weakly Augmented Labeled EEG
Strongly Augmented Labeled EEG
Encoder (G)
Classifier (C)
λ ~ ℬ (α, β)
Interpolated Original EEG
Interpolated Weakly Augmented EEG
Interpolated Strongly Augmented EEG
Averaged Prediction
Sharpened Pseudo-label
Ground Truth
Domain Label (Binary)
Domain Label (Binary)
Linear Combination: ⊕
Back Propagation:
Labeled:
Unlabeled:
$\frac{\partial \mathcal{L}_c}{\partial \theta_C}$
$\frac{\partial \mathcal{L}_c}{\partial \theta_C}$
$\mathcal{L}_c$
Interpolated Emotion Label
GRL
Discriminator (D)
$\mathcal{L}_d$
$-\frac{\partial \mathcal{L}_d}{\partial \theta_G}$
$\frac{\partial \mathcal{L}_d}{\partial \theta_D}$
Interpolated Domain Label (Binary)
Interpolated Domain Label

Figure 1: The architecture of our proposed semi-supervised learning network.

a new set. We then use the averaged predictions of a classifier's predictions to guess the labels of the original and augmented unlabeled data. Next, we apply a convex combination between labeled and unlabeled data. Following, we perform representation alignment by training an emotion classifier and a domain discriminator on the interpolated set. We then compare our solution with various SSL techniques, which we implement and adapt for EEG-based emotion recognition. We test our method, as well as the benchmarks, across varying amounts of labeled data, and show that our solution achieves superior performance in most scenarios.

2 Method

Problem Setup. For a classification problem with k emotion categories, let us denote $\mathcal{D}_l = \{x_i^l, y_i^l\}_{i=1}^M$, $\mathcal{D}_u = \{x_i^u\}_{i=1}^N$, $\mathcal{D}_v = \{x_i^l, y_i^l\}_{i=1}^I$, and $\mathcal{D}$ as the labeled, unlabeled, validation, and entire training sets, where $\mathcal{D}_l \cup \mathcal{D}_u \cup \mathcal{D}_v = \mathcal{D}$ and $\mathcal{D}_l \cap \mathcal{D}_u \cap \mathcal{D}_v = \emptyset$. $\mathcal{D}_l$ is formed by m sample(s) per class selected from $\mathcal{D}$, where $M = m \times k$. Our goal is to propose a robust pipeline to improve the model prediction by leveraging both $\mathcal{D}_l$ and $\mathcal{D}_u$, particularly when $M \ll N$. We are also interested in a barely supervised scenario, where *only one* sample per class is labeled. We propose a novel SSL architecture that is inspired by very recent cutting-edge SSL solutions in the computer vision domain, namely MixMatch (Berthelot et al. 2019), FixMatch (Sohn et al. 2020), and ADA-Net (Wang, Li, and Gool 2019). An overview of our proposed architecture is shown in Figure 1.

Data Augmentation. We first apply strong and weak augmentations on both labeled and unlabeled data. For each x_b^l and x_b^u in the training batch of D_l and D_u, we generate the augmented data as $\mathcal{A}_{s/w}(x_b) = x_b + \mathcal{N}(\mu, \sigma)$, where $x_b \sim [0, 1]$ is the normalized input, and $\mathcal{N}$ is a Gaussian distribution with $\mu = 0.5$. The strength of the augmentation can be tuned by changing σ. We choose 0.8 and 0.2 as σ in additive Gaussian Noise for strong ($\mathcal{A}_s$) and weak ($\mathcal{A}_w$) augmentations respectively, as suggested in (Liu et al. 2019; Zhang and Etemad 2021a).

Label-Guessing. After data augmentation, we enlarge the unlabeled set of data as the concatenation of original (x_b^u), weakly ($\mathcal{A}_w(x_b^u)$), and strongly ($\mathcal{A}_s(x_b^u)$) augmented unlabeled data. We then pass-forward the enlarged data to an encoder (G) and a classifier (C) to generate pseudo-labels (p_b) by averaging the predictions as:

$$p_b = \frac{1}{3}\sum_1^S \mathrm{softmax}(p_m(y \mid x_b^{u,s}; \theta_G, \theta_C)), \tag{1}$$

where $x_b^{u,s}$ denotes $x_b^u, A_w(x_b^u)$ and $A_s(x_b^u)$ when $s = 1, 2, 3$. Here, p_m is the model prediction, and θ_G and θ_C denote the model parameters for the encoder and classifier, respectively. p_b has been further sharpened as $p_b^u = p_b / \sum_1^k p_b$. In addition to emotion pseudo-labels, we also assign binary domain labels (z^u) to the unlabeled set (G_u) as:

$$G_u = \{(\langle x_b^u \cup \mathcal{A}_s(x_b^u) \cup \mathcal{A}_w(x_b^u)\rangle, \langle p_b^u \cup p_b^u \cup p_b^u\rangle, z^u)\}, \tag{2}$$

where $\langle , \rangle$ denotes the concatenation of two or more sets.

Similarly, we enlarge the labeled data as the concatenation of the original (x_b^l), weakly ($\mathcal{A}_w(x_b^l)$), and strongly ($\mathcal{A}_s(x_b^l)$) augmented labeled data. Since the augmentation should not alter the labels of the data, we employ the ground truth (y_b^l) of x_b^l as labels for the augmented data. Moreover, we assign binary domain labels (z^l) to the labeled set (G_l) as:

$$G_l = \{(\langle x_b^l \cup \mathcal{A}_s(x_b^l) \cup \mathcal{A}_w(x_b^l)\rangle, \langle y_b^l \cup y_b^l \cup y_b^l\rangle, z^l)\}. \tag{3}$$

The data and corresponding labels from both G_l and G_u are used for pairwise convex combination, which will be discussed in the next section.

MixUp. Inspired by a powerful data augmentation method MixUp (Zhang et al. 2018a), we generate a new training set by applying convex combinations between the *labeled* and *unlabeled* set (G_l, G_u), according to:

$$\lambda \sim \mathcal{B}(\alpha, \alpha), \tag{4}$$

$$(\tilde{x}_b, \tilde{y}_b, \tilde{z}_b) = \lambda\{G_l\} + (1 - \lambda)\{G_u\}, \tag{5}$$

where $\mathcal{B}$ is Beta distribution with α of 0.25 .

Pairwise Representation Alignment. Following MixUp, we align the distribution of the pairwise labeled and unlabeled representations by training on the new interpolated set ($\tilde{x}_b, \tilde{y}_b, \tilde{z}_b$). To measure the distribution divergence between labeled and unlabeled data, we use:

$$d_H = 2\{1 - \min[\frac{1}{|\tilde{x}_b|}\sum_i^{|\tilde{x}_b|} \arg\max(p_m(z \mid x_i; \theta_G, \theta_D)) \neq z_i]\}, \tag{6}$$

based on (Ganin et al. 2016; Wang, Li, and Gool 2019), in which $x_i \in \tilde{x}_b$ and $z_i \in \lfloor \tilde{z}_b \rceil$. θ_D is the model parameters for the discriminator (D). We minimize this distribution divergence in order to encourage the encoder (G) to align the EEG representations of labeled and unlabeled data as:

$$\min_G d_H = \max_G \min_H [\frac{1}{|\tilde{x}_b|}\sum_i^{|\tilde{x}_b|} \arg\max(p_m(z \mid x_i; \theta_G, \theta_D)) \neq z_i]\}. \tag{7}$$

As suggested in (Ganin et al. 2016; Wang, Li, and Gool 2019), we optimize this max-min problem by adding a Gradient Reverse Layer (GRL) before the D in order to reverse the gradient in G, as shown in Figure 1. Finally, we train both the emotion classifier and the domain discriminator by minimizing the emotion classification loss (L_c in Eq. 8) and domain discriminator (L_d in Eq. 9) loss with an adversarial training strategy as follows:

$$\mathcal{L}_c = ||\tilde{y}_b - \mathrm{softmax}(p_m(y \mid \tilde{x}_b; \theta_G, \theta_C))||_2^2 \tag{8}$$

$$\mathcal{L}_d = \mathcal{H}(\lfloor \tilde{z}_b \rceil, p_m(z \mid \tilde{x}_b; \theta_G, \theta_D)), \tag{9}$$

where $\mathcal{H}(p, q) = -\sum p(x) \log q(x)$ represents cross-entropy.

Total Loss Function. Our total loss function comprises three parts. The first term is a supervised loss $\mathcal{L}_s = \mathcal{H}(y_b^l, p_m(y \mid x_b^l; \theta_G, \theta_C))$ that has been commonly used in many SSL literature. We adopt the unsupervised loss L_u used in (Sohn et al. 2020) as the second term, as follows:

$$\mathcal{L}_u = \mathcal{H}(y_b^u, p_m(y \mid \mathcal{A}_s(x_b^u); \theta_G, \theta_C)), \tag{10}$$

where $y_b^u = \arg\max(p_m(y \mid \mathcal{A}_w(x_b^u); \theta_G, \theta_C))$. Consequently, we update the total loss as:

$$\mathcal{L}_{total} = \mathcal{L}_s + \eta\mathcal{L}_u + (\lambda\mathcal{L}_c + \mathcal{L}_d). \tag{11}$$

Instead of using a pre-defined threshold τ (Sohn et al. 2020) that may need to be tuned for each dataset individually, we apply a warm-up function η on the unsupervised loss, similar to (Berthelot et al. 2021). The third term is the sum of the classification and discriminator losses trained on the interpolated set as mentioned earlier.

3 Experimental Setup and Results

Datasets. The **SEED** dataset (Zheng and Lu 2015) contains EEG recorded using 62 electrodes at a sampling rate of 1000 *Hz*. In the experiments, 15 film clips with three emotions (neutral, positive, and negative) were selected as stimuli. The studies were completed by a total of 15 individuals, 8 females and 7 males. Each subject takes part in the experiment twice, with each experiment consisting of 15 recording trials. The **SEED-IV** dataset (Zheng et al. 2018) contains 62-channel EEG recordings obtained with a sample frequency of 1000 *Hz*. 72 video snippets with four emotions (neutral, fear, sad, and happy) were used as stimuli. The experiments were carried out by 15 participants (8 females and 7 males). Each subject repeated the experiment three times, with different stimuli used each time. Each experiment has 24 recording trials (6 trials for each emotion).

Pre-processing and Feature Extraction. We follow the exact pre-processing steps described in (Zheng and Lu 2015; Zheng et al. 2018) as follows. We first down-sample the EEG to 200 *Hz*. Then, to reduce artefacts, band-pass filters with frequencies ranging from $0.3-50$ *Hz* and $1-75$ *Hz* were applied to the EEG recordings. Following pre-processing, EEG data were split into continuous segments of similar length (1 second for SEED and 4 seconds for SEED-IV) with no overlap. Differential Entropy (DE) features were extracted from five bands (delta, theta, alpha, beta, and gamma) of each EEG segment. We assume EEG signals obey a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, and thus DE can be calculated as $DE = \frac{1}{2} \log 2\pi e \sigma^2$.

Evaluation Protocols. We apply the same evaluation protocols that were used in the original dataset publications (Zheng and Lu 2015; Zheng et al. 2018). For SEED, we use the first 9 trials for training and the remaining 6 trials for testing. For SEED-IV, we use the first 16 trials for training, and the rest 8 trials for testing.

Implementation details. In all the experiments, we train for 30 epochs with a batch size of 8. We employ Adam optimizer (Kingma and Ba 2014) with the default learning rate of 0.001. Tuning of the hyper-parameters has been done on the validation set D_v (with no leakage with the test set). A pre-defined threshold (τ) has been used in FixMatch and AdaMatch. We perform a grid search in $[0.0 - 1.0]$ with a step size of 0.1. In FixMatch, we set $\tau = 0.9$ for both datasets, while in AdaMatch, we set $\tau = 0.6$ and $\tau = 0.5$ for SEED and SEED-IV, respectively. Our experiments were carried out on two NVIDIA GeForce RTX 2080 Ti and six Tesla P100 GPUs using PyTorch (Paszke et al. 2019).

The convolutional encoder consists of two 1-D convolutional blocks, where each block contains a 1-D convolutional layer followed by a 1-D batch normalization layer and a LeakyReLU activation. The classifier and discriminator share the same structure containing two fully connected layers with a dropout rate of 0.5.

SSL Benchmarks. For comparison and evaluation of our work, we adapt, implement, and where necessary, modify four state-of-the-art SSL methods, ADA-Net (Wang, Li, and Gool 2019), MixMatch (Berthelot et al. 2019), FixMatch (Sohn et al. 2020), and AdaMatch (Berthelot et al. 2021), in addition to five classical SSL methods, Π-model (Samuli and Timo 2017), temporal assembling (Samuli and Timo 2017), mean teacher (Tarvainen and Valpola 2017), convolutional autoencoders (Tong, Wu, and Wang 2019), and pseudo labeling (Lee et al. 2013), for EEG representation learning. We employ the same convolutional module in these benchmarks as in the one used in our proposed method. For the decoder component of the convolutional autoencoder, we use two transposed convolutional 1-D blocks. In each block, a 1-D transposed convolutional layer is followed by a 1-D batch normalization layer and ReLU activation. We implement the Π-model, temporal ensembling, mean teacher, pseudo-labeling, and convolutional autoencoder with the same algorithm settings (e.g., loss function, unsupervised weight, etc.) as used in (Zhang and Etemad 2021b).

Performance. We experiment with varying amounts of labeled samples per class used, namely $1, 3, 5, 7, 10, 25$. As shown in Table 1, classical SSL methods generally show inferior performance in all 6 designated label scenarios. Convolutional autoencoder consistently outperforms other classical approaches and slightly outperforms FixMatch. In all of the scenarios, ADA-Net and AdaMatch consistently outperform the classical SSL approaches. When only one sample per class is provided, AdaMatch achieves the second best result (shown with underline). When more samples are available, ADA-Net obtains the second best results. Our method consistently achieves the best (shown in bold) performance. In particular, our proposed architecture outperforms the sec-

Table 1: The performance of our method in comparison to other semi-supervised methods on SEED dataset.

Method	1 label	3 labels	5 labels	7 labels	10 labels	25 labels
Π-model	0.6025±0.0957	0.6787±0.1014	0.7243±0.1132	0.7494±0.1084	0.7635±0.1093	0.7787±0.1088
Temporal Ens.	0.5922±0.0902	0.6995±0.0907	0.7380±0.0978	0.7715±0.0957	0.7980±0.0953	0.8383±0.0873
Mean Teacher	0.5397±0.0824	0.6275±0.0998	0.6642±0.0946	0.6990±0.1132	0.7148±0.0898	0.7709±0.0966
Conv. AutoEnc.	0.7139±0.1220	0.8003±0.1169	0.8286±0.1089	0.8474±0.0970	0.8546±0.0977	0.8734±0.0896
Pseudo Label	0.6802±0.1320	0.7811±0.1202	0.7957±0.1078	0.8221±0.1103	0.8411±0.0979	0.8532±0.0938
ADA-Net	0.7200±0.1297	0.8366±0.1074	0.8704±0.0869	0.8923±0.0802	0.8929±0.0784	0.9082±0.0707
MixMatch	0.6897±0.1393	0.8089±0.1280	0.8394±0.1030	0.8546±0.0964	0.8584±0.0924	0.8688±0.0878
FixMatch	0.6636±0.1384	0.7626±0.1156	0.7904±0.1068	0.8179±0.1056	0.8314±0.0998	0.8444±0.0909
AdaMatch	0.7403±0.1178	0.8259±0.1026	0.8362±0.1084	0.8584±0.0969	0.8671±0.0909	0.8802±0.0880
Ours	**0.7777±0.1205**	**0.8652±0.1002**	**0.8838±0.0909**	**0.8955±0.0862**	**0.9050±0.0774**	**0.9114±0.0752**

Table 2: The performance of our method in comparison to other semi-supervised methods on SEED-IV dataset.

Method	1 label	3 labels	5 labels	7 labels	10 labels	25 labels
Π-model	0.4993±0.1230	0.5465±0.1466	0.5765±0.1436	0.5879±0.1477	0.6014±0.1514	0.6192±0.1530
Temporal Ens.	0.5276±0.1315	0.5976±0.1448	0.6300±0.1435	0.6526±0.1407	0.6592±0.1371	0.6725±0.1363
Mean Teacher	0.4703±0.1184	0.5156±0.1235	0.5505±0.1327	0.5660±0.1291	0.5666±0.1178	0.5797±0.1300
Conv. AutoEnc.	0.5319±0.1858	0.5952±0.1813	0.6301±0.1674	0.6483±0.1575	0.6640±0.1726	0.6596±0.1662
Pseudo Label	0.5231±0.1793	0.5808±0.1676	0.6036±0.1792	0.6084±0.1759	0.6213±0.1903	0.6271±0.1836
ADA-Net	0.5465±0.1696	0.6647±0.1790	0.6883±0.1960	**0.7344±0.1412**	0.7223±0.1413	**0.7500±0.1504**
MixMatch	0.5608±0.1592	0.6503±0.1579	0.6942±0.1631	0.7092±0.1602	**0.7231±0.1627**	0.7320±0.1519
FixMatch	0.5337±0.1733	0.6357±0.1557	0.6343±0.1626	0.6462±0.1557	0.6650±0.1591	0.6854±0.1558
AdaMatch	**0.5830±0.1595**	0.6652±0.1658	0.6912±0.1645	0.6811±0.1580	0.6931±0.1687	0.7143±0.1604
Ours	0.5787±0.1809	**0.6853±0.1634**	**0.6966±0.1596**	0.7077±0.1647	0.7215±0.1608	0.7232±0.1615

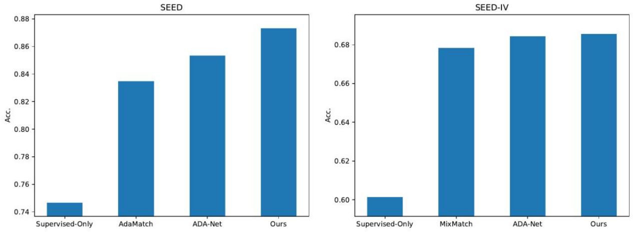

Figure 2: Average performance of the top three SSL methods in comparison to a supervised method for both datasets.

ond best methods by over 3.7% and 2.8% in very few labeled scenarios $(1, 3)$, demonstrating its superiority.

As shown in Table 2, when compared to other more recent solutions, traditional SSL methods achieve poorer performance. Temporal ensembling and convolutional autoencoder have comparable performances and produce the best two results among classical approaches. In every labeled scenario, ADA-Net, MixMatch, FixMatch, and AdaMatch outperform all classical methods. In particular, when only 1 sample per class is labeled, our proposed method achieves the second best result (0.5787 ± 0.1809), approaching the best method by AdaMatch (only 0.43% difference). Our method obtains the best results when more labeled samples are given $(3, 5)$. Especially when 3 samples per class are labeled, our method outperforms the second best result (ADA-Net) by 2%. When more labeled samples are provided, our method approaches the highest performance.

In Figure 2, we compare the average performance (across $1, 3, 5, 7, 10, 25$ labels) of the top 3 SSL methods to a supervised-only method. The supervised method trains a backbone CNN (the same model used in the SSL methods) only on labeled data. We observe that for SEED, our method achieves the best results of 0.8731 ± 0.0917, followed by ADA-Net (0.8534 ± 0.0922) and AdaMatch (0.8347 ± 0.1008). In SEED-IV, our method achieves the best results with 0.6855 ± 0.1651, followed by ADA-Net (0.6844 ± 0.1629) and MixMatch (0.6783 ± 0.1592).

4 Conclusion

In this research, we propose a novel semi-supervised method for EEG-based emotion recognition. Our model relies on data augmentation, label guessing, MixUp, and pairwise alignment between the distributions of unlabeled and labeled data. We conduct extensive experiments against a number of other methods, where only $1, 3, 5, 7, 10$, and 25 samples per class are labeled, and evaluate the performance on two large publicly available datasets, SEED and SEED-IV. In SEED, our architecture consistently outperforms all other methods across all 6 designated labeled scenarios. In SEED-IV, our proposed method shows competitive performance across all the labeled scenarios. In both datasets, our method achieves excellent results in the very few labeled scenarios $(1, 3)$. The results also show that our framework considerably outperforms the supervised-only method, addressing the challenge of scarcity of labeled EEG data.

References

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. *Advances in Neural Information Processing Systems*, 32.

Berthelot, D.; Roelofs, R.; Sohn, K.; Carlini, N.; and Kurakin, A. 2021. AdaMatch: A Unified Approach to Semi-Supervised Learning and Domain Adaptation. *arXiv preprint arXiv:2106.04732*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896.

Liu, W.; Qiu, J.-L.; Zheng, W.-L.; and Lu, B.-L. 2019. Multimodal emotion recognition using deep canonical correlation analysis. *arXiv preprint arXiv:1908.05349*.

Oliver, A.; Odena, A.; Raffel, C. A.; Cubuk, E. D.; and Goodfellow, I. 2018. Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. *Advances in Neural Information Processing Systems*, 31: 3235–3246.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037.

Picard, R. W. 2000. *Affective Computing*. MIT press.

Samuli, L.; and Timo, A. 2017. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, volume 4, 6.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems*, 33.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.

Tong, L.; Wu, H.; and Wang, M. D. 2019. CAESNet: Convolutional AutoEncoder based Semi-supervised Network for improving multiclass classification of endomicroscopic images. *Journal of the American Medical Informatics Association*, 26(11): 1286–1296.

Van Engelen, J. E.; and Hoos, H. H. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2): 373–440.

Wang, Q.; Li, W.; and Gool, L. V. 2019. Semi-supervised learning by augmented distribution alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1466–1475.

Zhang, G.; and Etemad, A. 2020. RFNet: Riemannian fusion network for EEG-based brain-computer interfaces. *arXiv preprint arXiv:2008.08633*.

Zhang, G.; and Etemad, A. 2021a. Capsule Attention for Multimodal EEG-EOG Representation Learning with Application to Driver Vigilance Estimation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Zhang, G.; and Etemad, A. 2021b. Deep Recurrent Semi-Supervised EEG Representation Learning for Emotion Recognition. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8. IEEE.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018a. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhang, T.; Zheng, W.; Cui, Z.; Zong, Y.; and Li, Y. 2018b. Spatial–temporal recurrent neural network for emotion recognition. *IEEE transactions on cybernetics*, 49(3): 839–847.

Zheng, W.-L.; Liu, W.; Lu, Y.; Lu, B.-L.; and Cichocki, A. 2018. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics*, 49(3): 1110–1122.

Zheng, W.-L.; and Lu, B.-L. 2015. Investigating Critical Frequency Bands and Channels for EEG-based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development*, 7(3): 162–175.

Zheng, W.-L.; Zhu, J.-Y.; and Lu, B.-L. 2017. Identifying stable patterns over time for emotion recognition from EEG. *IEEE Transactions on Affective Computing*.