Analysis of Semi-Supervised Methods for Facial Expression Recognition

Shuvendu Roy, Ali Etemad
Dept. ECE and Ingenuity Labs Research Institute
Queen's University, Kingston, Canada
{shuvendu.roy, ali.etemad}@queensu.ca

2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII) | 978-1-6654-5908-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACII55700.2022.9953876

***Abstract*—Training deep neural networks for image recognition often requires large-scale human annotated data. To reduce the reliance of deep neural solutions on labeled data, state-of-the-art semi-supervised methods have been proposed in the literature. Nonetheless, the use of such semi-supervised methods has been quite rare in the field of facial expression recognition (FER). In this paper, we present a comprehensive study on recently proposed state-of-the-art semi-supervised learning methods in the context of FER. We conduct comparative study on *eight* semi-supervised learning methods, namely Pi-Model, Pseudo-label, Mean-Teacher, VAT, MixMatch, ReMixMatch, UDA, and FixMatch, on three FER datasets (FER13, RAF-DB, and AffectNet), when various amounts of labeled samples are used. We also compare the performance of these methods against fully-supervised training. Our study shows that when training existing semi-supervised methods on as little as 250 labeled samples per class can yield comparable performances to that of fully-supervised methods trained on the full labeled datasets. To facilitate further research in this area, we make our code publicly available at: https://github.com/ShuvenduRoy/SSL_FER.**

***Index Terms*—Semi-Supervised Learning, Facial Expressions, Affective Computing**

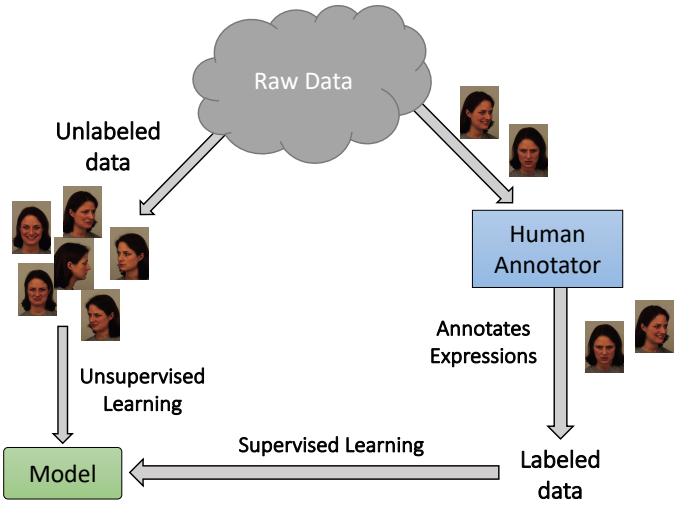


Fig. 1: Overview of semi-supervised learning for FER.

I. Introduction

Facial expressions play an important role in human communications. As a result, growing efforts are being made toward developing facial expression recognition (FER) methods that can facilitate better human-machine interaction systems. Real-world applications of FER systems include driving assistants [1], personal mood management systems [2], [3], health-care assistants [4], emotion-aware multimedia [5], and others. Although FER systems have shown great promise and improvements over the past few years, the FER remains challenging due to several factors such as occlusions, illuminations, scene backgrounds, challenging viewing angles, ethnicity, and demographic factors. Recently, deep learning solutions have shown the potential to effectively perform FER and solve many of these problems [6]–[9]. However, deep learning models require very large human-annotated datasets to achieve their optimal performance, and collecting such large datasets with human annotations is a costly and time-demanding process. Hence, FER solutions capable of learning robust representations from a relatively small amount of labeled data are highly desired.

To deal with the unavailability of large annotated datasets, methods such as self-supervised and semi-supervised learning leverage unlabeled data to learn important features with minimal supervision. While self-supervised methods [7], [10], [11] learn from pseudo-labels generated from unlabeled data, semi-supervised learning methods [12], [13] take a more hybrid approach to utilize small amounts of labeled data to guide the learning of large unlabeled datasets. Despite the success of these methods in image and video representation learning, their use for FER has been less explored. A high-level overview of semi-supervised learning for FER is shown in Fig 1.

Recently, a wide range of semi-supervised learning methods have been proposed which can be categorized into two broad categories: pseudo-labeling [14] and consistency regularization [15]–[17]. In pseudo-labeling [14], a model trained on the labeled data is used to predict the labels for the unlabelled data. The predicted pseudo-labels with higher confidence are subsequently used as labels for the unlabeled data. The model is then trained with the unlabeled data in a supervised setting using the generated pseudo-labels. Consistency regularization-based methods are based on the assumption that a realistic perturbation (e.g. augmentation) on an image does not change its semantics, and therefore the output of the model should remain unchanged. The most common type of these methods reduces the distance between the embeddings of two augmentations of an unlabeled image [15], [16]. More recently, consistency regularization has been combined with pseudo-labeling in a hybrid framework that shows impressive

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

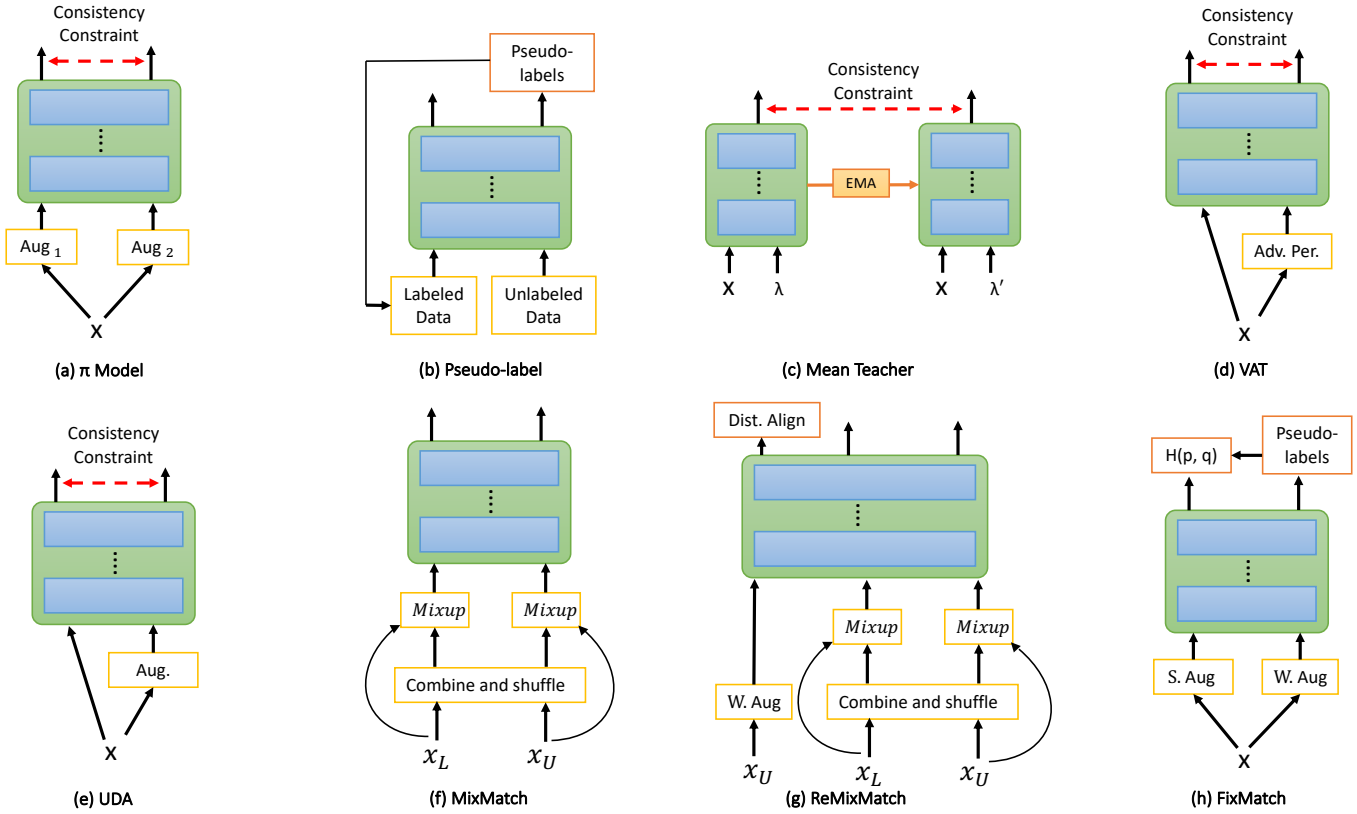


Fig. 2: Overview of the eight state-of-the-art semi-supervised methods studied in this paper for FER. Here, Aug., W. Aug, and S. Aug represent augmentations, weak, and augmentations respectively. Adv. Per. represents adversarial perturbation. EMA is the exponential moving average of the original model.

performance in a wide variety of tasks [18]–[20]. An example of such a hybrid method is FixMatch [18], which first applies a weak augmentation for an unlabeled image and predicts the pseudo-label for it. Then the model is trained with a hard augmentation of the same image using the predicted pseudo-label.

Although there has been some progress in semi-supervised learning in FER, to the best of our knowledge, there is no comprehensive and comparative study on the recent semi-supervised methods applied to FER. In this work, we study eight recently proposed semi-supervised methods on three popular facial expression datasets. The aim of this work is to study the adaptability and performance of such methods on FER without any specialization in expression recognition. We conduct extensive studies with different amounts of labeled data and find impressive results with as little as only 10 labeled samples per class used for training. These methods, when trained with as few as 250 samples per class, show competitive results to the fully-supervised method trained on the same dataset when all of the sample labels are used.

In this paper we make the following contributions:

- We present a comprehensive study on eight recent semi-supervised learning methods on three popular FER datasets.
- We conduct a comparative analysis of the methods and further compare them with fully supervised learning using the same backbone encoder.
- To facilitate quick reproduction and further research on the topic of semi-supervised FER, we release the code for this work which contains the implementations of all the methods in this study.

The remainder of this paper is organized as follows. In the next section, we describe the 8 semi-supervised methods in detail, namely Pi-Model, Pseuro-label, Mean-Teacher, VAT, MixMatch, ReMaxMatch, UDA, and FixMatch. Following, we describe the experiments including the description of the datasets, a comparison of the performances of the semi-supervised methods, and the impact of parameters. Finally, we present the concluding remarks.

II. Semi-supervised Methods

In this section, we describe the problem setup, and an overview of 8 popular and effective semi-supervised methods for image representation learning.

A. Problem Setup

Assume we have a small labeled dataset $D_l = \{(x_i^l, y_i^l)\}_{i=1}^N$ containing N images and its corresponding class labels.

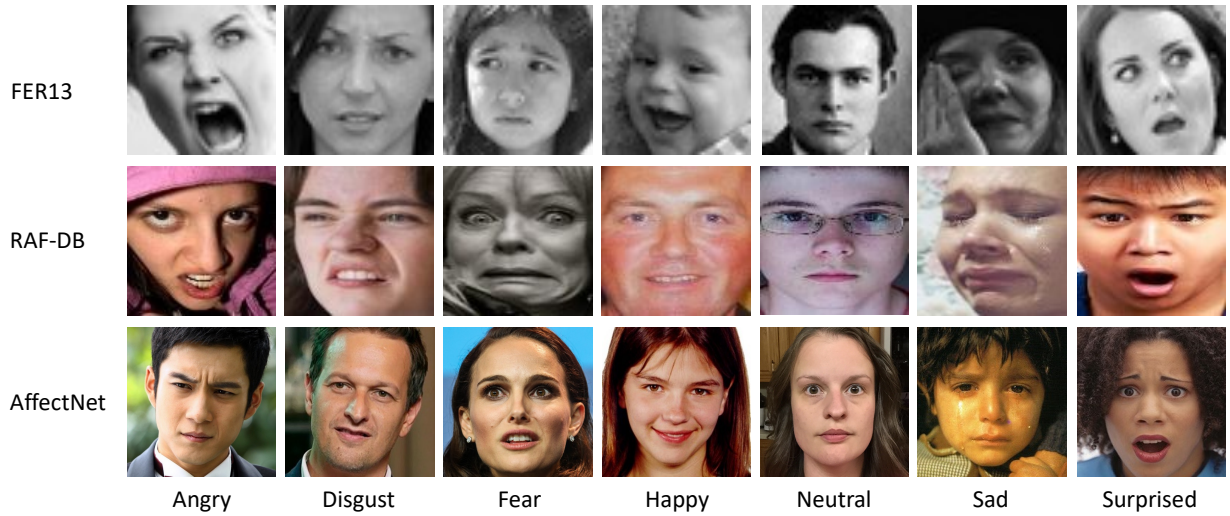


Fig. 3: Examples of images from FER13, RAF-DB, and AffectNet datasets.

Let's assume we also have a large unlabeled dataset $D_u = \{(x_i^u)\}_{i=1}^{M}$, where $M \gg N$. Although we do not have any annotated labels for the images in the unlabeled dataset D_u, our aim is to utilize D_u and D_l together to help the model learn better representations. In this problem setup, we assume a separate validation dataset $D_v = \{(x_i^u, y_i^u)\}_{i=1}^{V}$ that is used to test the final performance of the model. We ensure no overlap between the images in labeled, unlabeled, and validation sets, i.e., $D_l \cap D_u \cap D_v = \emptyset$.

B. Semi-Supervised Methods

We study eight recent semi-supervised methods namely: Pi-Model, Pseuro-label, Mean-Teacher, VAT, MixMatch, ReMaxMatch, UDA, and FixMatch. Following, we present an overview of each of these methods.

1) Pi-Model: Pi-Model [15] is one of the most popular *Consistency Regularization* based semi-supervised learning methods. In this approach, the model generates two augmentations of an image from the unlabelled data, and a regularization loss function reduces the difference in their learned embeddings. At the same time, the pi-model is trained on the labeled data with regular cross-entropy loss in a supervised way. The model also introduces stochastic behavior in the prediction of the model using dropout, random max-pooling, and a randomized augmentation module. The structure of the Pi-Model is shown in Figure 2 (a). The loss on the unsupervised data is represented as

$$\mathbb{E}_{x \in D_l} \mathcal{R}(f(\theta, \tau_1(x)), f(\theta, \tau_2(x))), \tag{1}$$

where τ_1 and τ_2 are two random augmentations applied on the input image x, f is the encoder network, θ represents the parameters of the model f, and $\mathcal{R}$ is the consistency regularization function.

2) Mean Teacher: Mean teacher [21] is built on the concept of consistency regularization, similar to the approach proposed for the Pi-model. However, instead of using the same encoder to generate embeddings of two augmented images, the mean teacher uses an exponential moving average (EMA) of the encoder to predict the embedding for the second image. The regular encoder is called a *student model*, whereas the EMA of the student model is called a *teacher model*. The mean teacher applies the regularization loss on the prediction of the teacher and student models on two augmentations of the same image. A visual illustration of the mean teacher method is depicted in Figure 2 (c). The loss function proposed for the Mean teacher can be represented as

$$\mathbb{E}_{x \in D_u} \mathcal{R}(f(\theta, \tau_1(x)), f(EMA(\theta), \tau_2(x))), \tag{2}$$

where $EMA(\theta)$ is the teacher model. All other notations are similar to that of the Pi-model. The formula for updating the EMA teacher model from the student model is represented as

$$\theta' = m\theta' + (1 - m)\theta, \tag{3}$$

where m is smoothing coefficient for the EMA update.

3) VAT: Virtual Adversarial Training (VAT) [17] is conceptually similar to the Pi-model. While the Pi-model regularizes the embedding of two augmentations on the same image, VAT introduces the concept of adversarial attack as an alternate to augmentation. More specifically, it generates an adversarial transformation of the input. Following, consistency regularization is applied to the input image and the transformed image. A visual illustration of VAT is shown in Figure 2 (d). The loss function for VAT is represented as follows

$$\mathbb{E}_{x \in D_u} \mathcal{R}(f(\theta, x), f(\theta, \gamma^{adv}(x))), \tag{4}$$

where γ^{adv} adversarial perturbation operator.

4) UDA: Unsupervised domain adaptation (UDA) [16] is another consistency regularization method that showed a large improvement in the performance by replacing the regular augmentation module with recently proposed hard augmentation techniques such as AutoAugment [22] and RandAugment [23] that generate very dynamic and diverse augmentations

of an input image. Figure 2 (e) presents the diagram of the UDA method that is very similar to Pi-model and VAT, except for the augmentation module which is replaced with hard augmentations. The loss function of the UDA can be represented as

$$\mathbb{E}_{x \in D_u} \mathcal{R}(f(\theta, x), f(\theta, \tau(x))), \tag{5}$$

where τ is the hard augmentation module.

5) Pseudo-label: Pseudo-label [14] proposed a very simple yet efficient solution for semi-supervised learning. The concept of this method acts as the fundamental block for many of the current state-of-the-art methods. In Pseudo-label, the model predicts the output class probability for each of the unlabeled data which is considered as a pseudo-label for the unlabeled image. The model is then trained in a supervised setting using the labels for the labeled data and the pseudo-labels for the unlabeled data. A visual illustration of the Pseudo-label method is shown in Figure 2 (b). The loss function for the Pseudo-label method can be represented as

$$\mathcal{L} = \mathcal{L}(y_i^l, f(\theta, x_i^l)) + \lambda \mathcal{L}(y_i^u, f(\theta, x_i^u)), \tag{6}$$

where y_i^u is the prediction pseudo-label for the unlabeled image x_i^u, and λ is a coefficient that balances the impact of the two loss functions.

6) MixMatch: MixMatch [24] is a hybrid semi-supervised learning method that unifies consistency regularization with pseudo-labeling. Like the pseudo-labeling methods, MixMatch also predicts the pseudo-labels for the unlabeled data and utilizes them to train the model in supervised settings along with the labeled data. However, the new distinctive component of this algorithm is a Mixup operation that generates mixed inputs and mixed labels by interpolating between the labeled and unlabeled images and their corresponding labels. The Mixup operation of MixMatch can be represented with the following formula:

$$x' = \alpha x_l + (1 - \alpha) x_u, \tag{7}$$

where x_l and x_u are the labeled and unlabeled input images, and α balances influence of unlabeled images on the generated mixed image. The value of α is randomly sampled from a beta distribution.

MixMatch also introduces the concept of generating multiple instances of an unlabeled image with multiple augmentations. More specifically, MixMatch applies one augmentation on the labeled data (x_i^l, y_i^l), but k weak augmentations on an unlabeled image x_j^u to generate k instances, whose predictions are then averaged to generate one pseudo-label $\hat{y_j}$ for each unlabeled image.

The Mixup operation on a batch of labeled data ($d_l \in D_l$) and unlabeled data ($d_u \in D_u$) generates the dataset d_l' and d_u'. Accordingly, the combined MixMatch loss function of the labeled and unlabeled data can be represented as:

$$\mathcal{L}_l = \frac{1}{|d_l'|} \sum_{x,y \in d_l'} H(y, f(x, \theta)), \tag{8}$$

$$\mathcal{L}_u = \frac{1}{C|d_u'|} \sum_{x',y' \in d_u'} \|y\prime - f(x, \theta)\|_2^2, \tag{9}$$

$$\mathcal{L} = \mathcal{L}_l + \lambda \mathcal{L}_u. \tag{10}$$

Here, $H(p, q)$ is the cross-entropy between the distribution p and q, C is the number of classes, and λ is a hyper-parameter to balance the influence of the labeled and unlabeled loss terms. A visual illustration of the MixMatch method is shown in Figure 2 (f).

7) ReMixMatch: ReMixMatch [19] is an extension of MixMatch with two new ideas: distribution alignment and augmentation anchoring. The distribution alignment encourages the distribution of the predictions on the unlabeled data to be similar to that of the labeled data. Augmentation anchoring is added as a replacement for the consistency regularization of MixMatch to encourage the representation of strongly augmented images to be similar to that of weakly augmented images. Here the augmentation anchoring technique uses one weakly augmented image against multiple strongly augmented images. The method also introduced a new strong augmentation method called CTAugment that is more suitable in semi-supervised learning settings. The ReMixMatch method is illustrated in Figure 2 (g).

8) FixMatch: FixMatch [18] is a unified framework that combines consistency regularization with pseudo-labeling in a very simplified way to generate a very simple semi-supervised learning framework. For an unlabeled image, FixMatch first applies weak augmentations and predicts the pseudo-labels for them. Next, it applies a hard augmentation on the same images and trains the model in supervised settings with pseudo-labels. FixMatch only uses the pseudo-labels if the confidence of the predictions is higher than a pre-defined threshold (p_{cutoff}). FixMatch uses only standard shift and flip augmentations for its weak augmentation module. For the hard augmentation module, it uses RandAugment [23], CTAugment [19], and CutOut [25]. A visual illustration of FixMatch is depicted in Figure 2 (h).

III. Experiments and Performance

In this section, we first describe the three datasets used to evaluate the eight state-of-the-art semi-supervised methods. We then present the implementation setup in detail, including the encoder architecture and training protocol. We then present the results. Finally, we present a sensitivity analysis of the semi-supervised methods with respect to their hyper-parameters.

A. Datasets

1) FER13 [26]: This is a facial expression dataset that has been collected and labeled automatically by the Google image search API. All the images in this dataset are re-scaled to a

TABLE I: The performance of different semi-supervised methods on FER13, RAF-DB, and AffectNet datasets, when only 10, 25, 100, and 250 labeled samples are used for training. Here, bold and underline represent the best and second best accuracy for each setting.

	FER13				RAF-DB				AffectNet			
Method / m	10 labels	25 labels	100 labels	250 labels	10 labels	25 labels	100 labels	250 labels	10 labels	25 labels	100 labels	250 labels
Π-model	37.09	40.87	50.66	56.42	39.86	50.97	63.98	71.15	24.17	25.37	31.24	32.40
Mean Teacher	45.21	<u>55.14</u>	52.17	58.06	62.05	45.17	45.57	**76.85**	19.54	20.21	20.80	44.05
VAT	24.95	**55.22**	51.55	55.64	<u>63.10</u>	45.82	62.05	59.45	17.68	<u>35.02</u>	37.68	37.92
UDA	<u>46.72</u>	49.89	50.62	<u>60.68</u>	46.87	**53.15**	58.86	60.82	27.42	32.16	37.25	37.64
Pseudo-label	32.79	36.04	49.21	54.88	58.31	39.11	54.07	67.40	18.00	21.05	33.05	37.37
MixMatch	45.69	46.41	<u>55.73</u>	58.27	36.34	43.12	<u>64.14</u>	73.66	**30.80**	32.40	39.77	<u>48.31</u>
ReMixMatch	41.07	43.25	44.62	57.49	37.35	42.56	42.86	61.70	29.28	33.54	<u>41.60</u>	46.51
FixMatch	**47.88**	49.90	**59.46**	**62.20**	**63.25**	<u>52.44</u>	**64.34**	<u>75.51</u>	<u>30.08</u>	**38.31**	**46.37**	**51.25**

TABLE II: Fully supervised training with full and partial training data, as well as the top semi-supervised performance.

	Fully-sup.		Semi-sup. (best)
Dataset	**All labeled data**	**250 labels/class**	**250 labels/class**
FER13	64.57	53.58	62.20
RAF-DB	80.47	65.87	76.85
AffectNet	54.91	40.28	51.25

resolution of 48×48. The dataset contains over 35K images of 7 basic expressions where 28K images belong to the training split. Figure 3 (top) shows a few examples of the dataset.

2) RAF-DB [27]: This is an in-the-wild dataset that contains around 15K images with 12K images on the training split. A total of 315 human annotators worked on the annotation of this dataset and each label is assured to be annotated by around 40 individual annotators. This dataset also contains 7 expression classes. Some examples of RAF-DB dataset is shown in Figure 3 (middle).

3) AffectNet [28]: This is a very large-scale in-the-wild dataset collected from the Internet with three search engines and expression keywords. We took the images for 7 basic expressions (a total of around 284K images), similar to the FER13 and RAF-DB datasets. Some examples of the AffectNet dataset are shown in Figure 3 (bottom).

B. Implementation details

Here we present the implementation details including the encoder and the training protocol used in this study. Unless mentioned otherwise, we use the same setup for all the results reported in this paper.

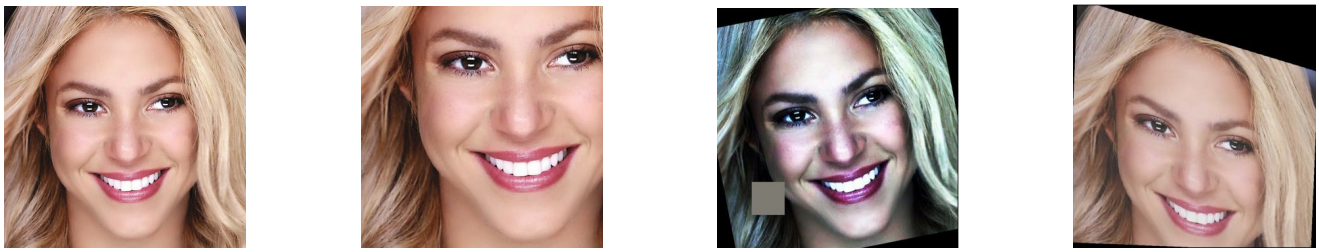
1) Encoder: We use ResNet [29] as the general backbone encoder for all the semi-supervised methods, as our preliminary experiments showed that this style architecture outperforms others such as VGG-style networks [30]. More specifically, we use the ResNet-18 as our default encoder which is trained with an input resolution of 224×224. A variant of ResNet called WideResNet, which has been specially designed to work well with images of very low resolution, is used in our study for the FER13 dataset.

2) Training protocol: For semi-supervised training, we first randomly select $N = n \times C$ images for the labeled dataset D_l, where C is the total number of classes and n is the number of labeled images to be taken par class. In this study, we present the result for all the methods with $n \in \{10, 25, 100, 250\}$ samples. Following FixMatch, we have trained all the models for 2^{20} iterations with a batch-size of 64 for the labeled data and a 7-times larger batch size for the unlabeled data. The cutoff confidence value for selecting a pseudo-label is set to 0.95 for all the pseudo-label based methods. The models are trained with SGD optimizer with a momentum of 0.9 and an initial learning rate of 0.03. A cosine learning rate decay is used as a scheduler to reduce the learning rate over the training iterations. Weight decay is used as a regularizer with a value of 0.0005. For the EMA model, a moving average weight of 0.999 is used. For the models with sharpening distribution, a temperature value of 0.5 is utilized.

To compare the performance of the semi-supervised methods, we also train the same backbone encoder in a fully-supervised setting. For the experiment with fully supervised training, we present the results in two settings: (1) using the full training data for training, and (2) using 250 images per class for training (maximum amount of labeled data used by the semi-supervised methods). For a fair comparison, we keep all the training settings the same where applicable. For fully supervised training, the augmentation module consists of random resizing, random horizontal flip, and random crop. The fully supervised models are also trained for 2^{20} iterations with SGD optimizer and a cosine learning rate decay scheduler. All the methods are trained with the proposed weak and strong augmentations in the original method. Some example of weak and strong augmentations applied to a facial image is illustrated in Figure 4.

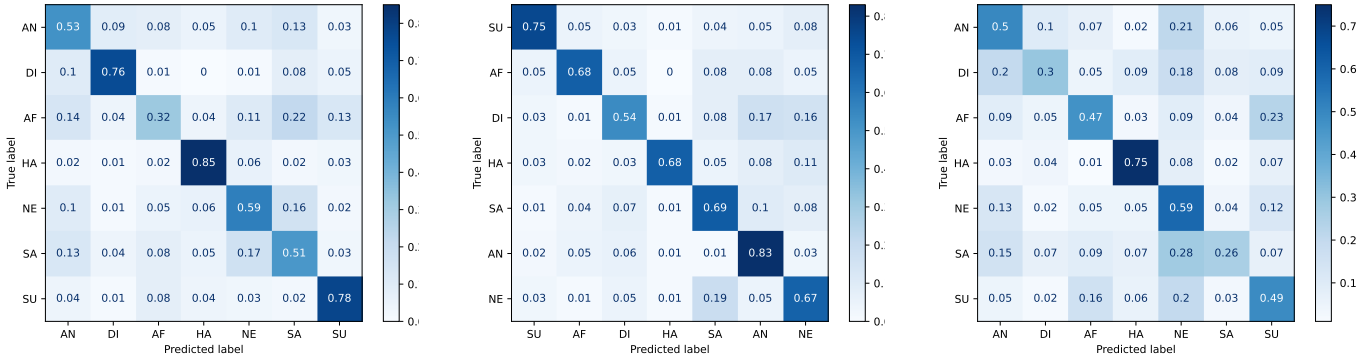
C. Results

Table I presents the results for the eight semi-supervised methods on FER13, RAF-DB, and AffectNet datasets. Note that, the experiments for Pseudo-label, Mean Teacher, MixMatch, and FixMatch are done with the best hyper-parameters for each dataset that are found from the sensitivity study which will be presented in Section III-D. As standard practice in semi-supervised literature [18], [24], we present the results



(a) Input image (b) Weakly augmented (c) Strongly aug. 1 (d) Strongly aug. 2

Fig. 4: Examples of weak and strong augmentations applied on a sample facial image.



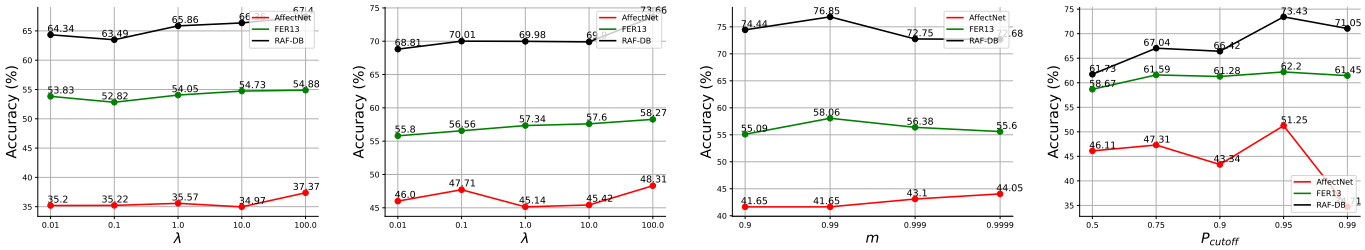
(a) *FER13 (FixMatch)* (b) *RAF-DB (Mean Teacher)* (c) *AffectNet (FixMatch)*

Fig. 5: Confusion matrices for the best semi-supervised model on FER13, RAF-DB, and AffectNet datasets.

for 10, 25, 100, and 250 labels per class, i.e., a total of 70, 175, 700, and 1750 labeled images for training. For the FER13 dataset, these training settings represent only 0.2%, 0.6%, 2.0%, and 6.0% of the total training data respectively, while for RAF-DB these training numbers represent 0.5%, 1.4%, 5.7% and 14.2% of the total training data. Finally for AffectNet, we end up using 0.03%, 0.07%, 0.28% and 0.7% of the total training set. Our experiments in this table show that for FER13, FixMatch outperforms the other methods (shown in bold) when 10, 100, and 250 labeled samples are used, whereas VAT performs the best when 25 labels are included in the training. In terms of the second best (shown with underline), there is no clear pattern among the different methods. As expected, the more labeled samples are used during training, the better the performance becomes (250 labeled samples $>$ 100 labeled samples $>$ 25 labeled samples $>$ 10 labeled samples). For RAF-DB, FixMatch performs strongly again by showing the best results for 10 and 100 labeled samples, while being the second-best approach for 25 and 250 labeled samples. UDA and Mean Teacher exhibit the best results when 25 and 250 labeled samples are used, respectively. Finally, AffectNet shows a similar trend where FixMatch outperforms the other solutions when 25, 100, and 250 labeled samples are used. When only 10 labels are used, FixMatch proves to be the second best method following MixMatch as the best method.

Overall, the result from Table I shows that out of 12 experiments (3 datasets, 4 label settings), FixMatch outperforms the others in 8 instances, while achieving the second best in 3 other settings. MixMatch obtains the next best results outperforming the others in 1 instance and being the second best in 3 more instances. 3 other top results are achieved by VAT, UDA, and Mean Teacher each with one best result among the 12 settings.

We summarize the results above in Table II by presenting the best results (best semi-supervised method trained with 250 labeled samples) from Table I. We also present the results of fully-supervised training with all the labeled data available during training, as well as fully-supervised training when only 250 labeled samples are present at the training. Here we observe that for FER13, the semi-supervised solution is only 2.4% less than the model trained in a fully-supervised fashion with all the training data but still 8.6% better than training with the same amount of labels but with a fully-supervised setup (not the semi-supervised approach). For RAF-DB the semi-supervised method is only 3.6% less than the model trained on the full dataset, but a considerable 11.0% better than the supervised model trained on the same amount of labeled data. Finally, for AffectNet, the semi-supervised approach is 3.7% lower than the fully-supervised model while using 140 times less labeled data. This performance is 11% better than the fully-supervised model trained with the same amount of labeled samples.



(a) Accuracy vs. λ for the ***Pseudo-label*** method

(b) Accuracy vs. λ for the ***MixMatch*** method.

(c) Accuracy vs. m of EMA for the ***Mean Teacher*** method.

(d) Accuracy vs. p_{cutoff} for the ***FixMatch*** method.

Fig. 6: Sensitivity study of various parameters for different semi-supervised methods on all three datasets.

To further analyze the performance of the models, we evaluate the confusion matrices of the predictions of the best semi-supervised models on test data for each dataset respectively, and show the results in Figure 5. Our analysis shows that even with training the models with small amounts of labeled data (250 samples), the semi-supervised approaches can learn to make strong predictions with reasonable mistakes. Specifically, we observe that for FER13 (Figure 5a), the majority of the mistakes occur by predicting 'sad' instead of 'afraid'. In RAF-DB (Figure 5b), on other hand, 'neutral' was often misclassified as 'sad', while for AffectNet (Figure 5c), 'sad' conversely classified as 'neutral' the most.

D. Sensitivity study

In this work, we adopt state-of-the-art semi-supervised methods for the task of FER where we use the training setups suggested in the original papers. However, since the problem domain of FER is quite different from that of general object classification (which is what the majority of the original semi-supervised literature has focused on), we also explore some of the important hyper-parameters of these methods. Specifically, we present a sensitivity study on the λ value of Pseudo-label and MixMatch, m for the EMA model of Mean Teacher, and the p_{cutoff} value of FixMatch. The other semi-supervised solutions do not have any major hyper-parameters to tune. The experiments are done on all three datasets (FER13, RAF-DB, and AffectNet), and the results are presented in Figure 6. In all cases, 250 labeled samples are used.

As shown in Figure 6a, the experiment on the λ parameter of the Pseudo-label method shows improvement for higher values of λ, which means giving higher importance to the unsupervised loss term results in better accuracy. In the original paper [14], a λ of 1.0 was used as the default value for general object classification on the CIFAR10 dataset. However, our experiment shows that for FER, even higher values of λ can be used which will further improve the results. We observed a maximum improvement of around 3% accuracy for AffectNet with this optimal value of λ when compared to the default value. As shown in Figure 6b, a similar trend is also observed for λ in MixMatch, where a maximum improvement of approximately 4% is observed on RAF-DB with higher emphasis on the unlabeled loss term.

The smoothing coefficient m of the EMA in the Mean Teacher method is known to impact the performance of this semi-supervised solution. In the original paper [21], higher values of m (close to 1) were suggested for optimal performance. In general, $m = 0.99$ works well on object classification, which as depicted in Figure 6c, also gives the best accuracy for FER13 and RAF-DB datasets. However, AffectNet shows better performance with even higher values of m.

Finally, we present the sensitivity study on p_{cutoff} value for FixMatch in Figure 6d. The experiment shows that the best performance is achieved with a cutoff probability of 0.95 across all the datasets, which is in line with the observations in the original FixMatch method [18]. A large drop is observed for higher and lower cutoff values on all three datasets.

IV. Summary

This paper presented a comprehensive study on different semi-supervised methods for FER using recent and state-of-the-art semi-supervised methods that have been originally proposed for general computer vision tasks. In particular, we adopted Pi-Model, Pseudo-label, Mean-Teacher, VAT, MixMatch, ReMixMatch, UDA, and FixMatch techniques in the context of FER and compared their performance against each other as well as fully-supervised settings on three popular datasets (FER13, RAF-DB, and AffectNet). Our study showed that even when the semi-supervised methods use a small portion of labeled data, they can achieve performances that are very competitive to the fully supervised methods trained using the full labeled dataset. We also compared the results with supervised models that use the same amount of labeled training data to that of semi-supervised methods, and observed significant improvements for the semi-supervised methods. Finally, we studied the impact of different hyper-parameters used in the semi-supervised methods to obtain a better understanding of the optimum settings for semi-supervision in FER. This work will serve as a baseline for future research on FER using semi-supervised approaches.

Acknowledgements

We would like to thank BMO Bank of Montreal and Mitacs for funding this research. We are also thankful to SciNet HPC Consortium for helping with the computation resources.

REFERENCES

[1] H. Leng, Y. Lin, and L. Zanzi, "An experimental study on physiological parameters toward driver emotion recognition," in *International Conference on Ergonomics and Health Aspects of Work with Computers*, 2007, pp. 237–246. 1

[2] M. Thrasher, M. D. Van der Zwaag, N. Bianchi-Berthouze, and J. H. Westerink, "Mood recognition based on upper body posture and movement features," in *International Conference on Affective Computing and Intelligent Interaction*, 2011, pp. 377–386. 1

[3] D. Sanchez-Cortes, J.-I. Biel, S. Kumano, J. Yamato, K. Otsuka, and D. Gatica-Perez, "Inferring mood in ubiquitous conversational video," in *12th International Conference on Mobile and Ubiquitous Multimedia*, 2013, pp. 1–9. 1

[4] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi, and M. Shimura, "Usage of emotion recognition in military health care," in *Defense Science Research Conference and Expo*, 2011, pp. 1–5. 1

[5] Y. Cho, S. J. Julier, and N. Bianchi-Berthouze, "Instant stress: detection of perceived mental stress through smartphone photoplethysmography and thermal imaging," *JMIR Mental Health*, vol. 6, no. 4, 2019. 1

[6] A. Sepas-Moghaddam, A. Etemad, F. Pereira, and P. L. Correia, "Capsfield: Light field-based face and expression recognition in the wild using capsule routing," *IEEE Transactions on Image Processing*, vol. 30, pp. 2627–2642, 2021. 1

[7] S. Roy and A. Etemad, "Self-supervised contrastive learning of multi-view facial expressions," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 253–257. 1

[8] M. Kolahdouzi, A. Sepas-Moghaddam, and A. Etemad, "Face trees for expression recognition," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–5. 1

[9] A. Sepas-Moghaddam, A. Etemad, F. Pereira, and P. L. Correia, "Facial emotion recognition using light field images with deep attention-based bidirectional lstm," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3367–3371. 1

[10] S. Roy and A. Etemad, "Spatiotemporal contrastive learning of facial expressions in videos," pp. 1–8, 2021. 1

[11] M. Pourmirzaei, F. Esmaili, and G. A. Montazer, "Using self-supervised co-training to improve facial representation," *arXiv preprint arXiv:2105.06421*, 2021. 1

[12] J. Jiang and W. Deng, "Boosting facial expression recognition by a semi-supervised progressive teacher," *IEEE Transactions on Affective Computing*, 2021. 1

[13] L. Wang, S. Wang, J. Qi, and K. Suzuki, "A multi-task mean teacher for semi-supervised facial affective behavior analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3603–3608. 1

[14] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896. 1, 4, 7

[15] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *Advances in neural information processing systems*, vol. 29, 2016. 1, 3

[16] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020. 1, 3

[17] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018. 1, 3

[18] 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. 2, 4, 5, 7

[19] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," *arXiv preprint arXiv:1911.09785*, 2019. 2, 4

[20] 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. 2

[21] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017. 3, 7

[22] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 113–123. 3

[23] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703. 3, 4

[24] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019. 4, 5

[25] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017. 4

[26] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International conference on neural information processing*. Springer, 2013, pp. 117–124. 4

[27] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861. 5

[28] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017. 5

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 5

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 5