Exploring the Boundaries of Semi-Supervised Facial Expression Recognition using In-Distribution, Out-of-Distribution, and Unconstrained Data

Shuvendu Roy, *Student Member, IEEE,* and Ali Etemad, *Senior Member, IEEE*

Abstract—Deep learning-based methods have been the key driving force behind much of the recent success of facial expression recognition (FER) systems. However, the need for large amounts of labelled data remains a challenge. Semi-supervised learning offers a way to overcome this limitation, allowing models to learn from a small amount of labelled data along with a large unlabelled dataset. While semi-supervised learning has shown promise in FER, most current methods from general computer vision literature have not been explored in the context of FER. In this work, we present a comprehensive study on 11 of the most recent semi-supervised methods, in the context of FER, namely Pi-model, Pseudo-label, Mean Teacher, VAT, UDA, MixMatch, ReMixMatch, FlexMatch, CoMatch, and CCSSL. Our investigation covers semi-supervised learning from in-distribution, out-of-distribution, unconstrained, and very small unlabelled data. Our evaluation includes five FER datasets plus one large face dataset for unconstrained learning. Our results demonstrate that FixMatch consistently achieves better performance on in-distribution unlabelled data, while ReMixMatch stands out among all methods for out-of-distribution, unconstrained, and scarce unlabelled data scenarios. Another significant observation is that with an equal number of labelled samples, semi-supervised learning delivers a considerable improvement over supervised learning, regardless of whether the unlabelled data is in-distribution, out-of-distribution, or unconstrained. We also conduct sensitivity analyses on critical hyper-parameters for the two best methods of each setting. To facilitate reproducibility and further development, we make our code publicly available at: github.com/ShuvenduRoy/SSL_FER_OOD.

Index Terms—Facial expression recognition, semi-supervised learning, in-distribution, out-of-distribution, unconstrained.

1 Introduction

Facial Expression Recognition (FER) [1], [2] is a critical application of computer vision that enables computers to identify and understand human expressions, with applications ranging from health care [3], [4], [5] to intelligent vehicles [6]. Deep learning methods have been the driving force behind most of the recent successes in FER. However, one of the major barriers to further improvement of deep learning-based FER is the need for large-scale labelled data. To this end, semi-supervised learning (SSL) has shown immense promise as a solution for improving performance while leveraging minimal supervision.

To tackle the problem of label scarcity, semi-supervised methods learn from a small amount of labelled data in conjunction with large amounts of unlabelled data. Depending on the relationship between the unlabelled and labelled data, there are three well-established forms of SSL: in-distribution (ID) [7], [8], [9], [10], [11], out-of-distribution (OOD) [12], [13], and unconstrained [14], [15] SSL. The ID SSL category assumes that the unlabelled data comes from the same distribution as the labelled data. However, in a practical application, this assumption is hard to satisfy or verify when collecting a sizeable unlabelled dataset. Consequently, more recent works have shifted focus toward more realistic data scenarios, including OOD and unconstrained SSL, that offer greater flexibility and potential for real-world applications. In OOD SSL, the unlabelled set contains samples from the same classes as the labelled set but comes from a different source and, therefore, has a different data distribution. On the other hand, unconstrained SSL assumes (1) that the unlabelled data is OOD, and (2) that the unlabeled set can contain samples that belong to classes that are not necessarily the same as those in the labelled set. This makes the unconstrained setting the most practical scenario for collecting large amounts of unlabelled data and scaling

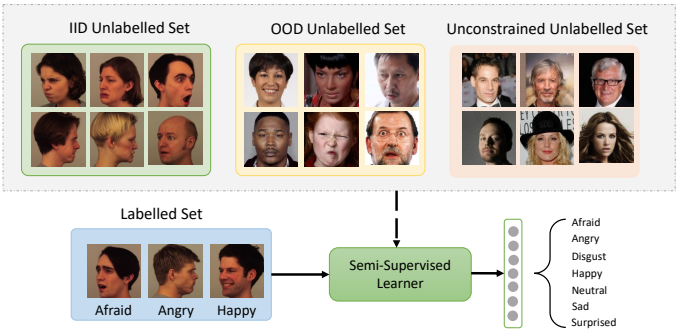


Fig. 1: Semi-supervised FER under ID, OOD, and unconstrained unlabelled data scenarios.

• *S. Roy and A. Etemad are with the Department of ECE and Ingenuity Labs Research Institute, Queen's University, Kingston, Canada. E-mail: shuvendu.roy@queensu.ca; ali.etemad@queensu.ca*

up semi-supervised learning [14].

In the context of FER, we identify the following open research problems regarding SSL: (1) While a few prior studies have explored the ID SSL in the context of FER, no prior works have explored the OOD SSL or unconstrained SSL in FER. (2) Many recent prominent methods originally proposed in general computer vision literature have not been explored in this context. In our recent work [16], we investigated and benchmarked the performance of several popular semi-supervised learning methods for FER. However, our previous study only focused on the ID SSL.

In this paper, we extend our previous study [16] by exploring semi-supervised FER under more realistic data settings, including OOD and unconstrained unlabeled data. Furthermore, we also report the performance of ID SSL with a small unlabelled set. We also expand the scope of our study by including a few of the more recent semi-supervised methods. More specifically, we study 11 recent semi-supervised approaches, namely Pi-model [17], Pseudo-label [7], Mean Teacher [18], VAT [8], UDA [9], MixMatch [19], ReMixMatch [11], FlexMatch [20], CoMatch [21], and CCSSL [22]. We conduct extensive experiments for different unlabelled set configurations using five public FER datasets, namely FER13 [23], RAF-DB [24], AffectNet [25], KDEF [26], and DDCF [27], along with a non-FER dataset for the unconstrained data, namely CelebA [28]. Figure 1 depicts an overview of our study on semi-supervised learning from different unlabelled set configurations. Findings from these experiments suggest that FixMatch performs the best among the 11 semi-supervised methods for conventional ID semi-supervised learning, but it performs poorly in other challenging settings. For both OOD and unconstrained unlabelled data, ReMixMatch exhibits the best performance. ReMixMatch also outperforms other methods in the low-data scenario. We also show hyper-parameter sensitivity studies for each of these semi-supervised settings to further boost performance.

Our contributions in this work are four-fold:

- We present a comprehensive and extensive study on 11 recent semi-supervised methods for FER and their comparison to fully supervised learning using six public datasets.
- We compare the performance of all methods under various data scenarios where the unlabelled data is ID, OOD, unconstrained, or very small.
- Our study finds that FixMatch consistently exhibits the best performance for ID unlabelled data, while ReMixMatch is the top-performing approach for OOD, unconstrained, and scarce unlabelled data. We also find that semi-supervised learning improves performance over supervised learning in all the tested scenarios.
- We make our code publicly available for quick reproducibility and further developments in this field: github.com/ShuvenduRoy/SSL_FER_OOD.

2 Related Work

In this section, we review the existing literature on semi-supervised learning from two main perspectives that are relevant to our work: (a) with ID data, (b) with OOD data, and (c) semi-supervised methods used specifically for FER.

2.1 ID SSL

In recent years, there has been an increasing interest in the literature on semi-supervised learning due to its potential for training large models with small amounts of labelled data. Most of these methods have been demonstrated to perform well on unlabelled data that come from the same distribution as the labelled set. The existing literature on semi-supervised learning can broadly be divided into two categories: entropy minimization [7], [29] and consistency regularization [8], [9], [18].

Entropy minimization-based methods learn by predicting the pseudo-labels of unlabelled samples with low entropy. Pseudo-label [7] was one of the first works in this direction, where the pseudo-labels for unlabelled samples are predicted and added to the labelled set if their entropy is low. Noisy-Student [29] generates these pseudo-labels with a pre-trained encoder, while Meta Pseudo-label [30] uses a teacher network to update the pseudo-labels based on the student network's performance. In contrast, consistency regularization-based methods aim to generate consistent predictions for different perturbations of the same input. Pi-model [17] was one of the first works in this direction, where two augmentations of an unlabelled image are forced to generate the same class prediction. Virtual adversarial training (VAT) [8] replaces the augmentation with adversarial perturbations and enforces consistency on the predictions. Unsupervised domain adaptation (UDA) [9] shows that replacing simple augmentations with hard augmentations [31], [32] could bring significant improvement to semi-supervised methods. Since then, most of the semi-supervised methods have used some form of hard augmentation in their pipeline.

Another line of work combines these two prominent approaches into the same framework. For example, MixMatch [19] enforces consistency on two perturbations (generated with MixUp [33]) and optimizes for lower entropy in the predictions. ReMixMatch [11] improves upon MixMatch by introducing two new concepts: augmentation anchoring and distribution alignment. Another hybrid method is FixMatch [10], which has gained tremendous success because of its simplicity while achieving state-of-the-art results in various domains. FixMatch learns by predicting the pseudo-labels for unlabelled images from its weak augmentation and uses them as ground truth for the hard augmentation of the same image if the confidence of the prediction is higher than a threshold. Several improvements have been made to FixMatch since its introduction, such as FlexMatch [20], which introduces the curriculum concept to adjust the threshold for different classes dynamically, and CoMatch [21], which introduces an extra contrastive loss term guided by the predicted pseudo-labels.

2.2 OOD and Unconstrained SSL

Since collecting a large amount of ID unlabelled data is difficult in practice, some recent semi-supervised methods

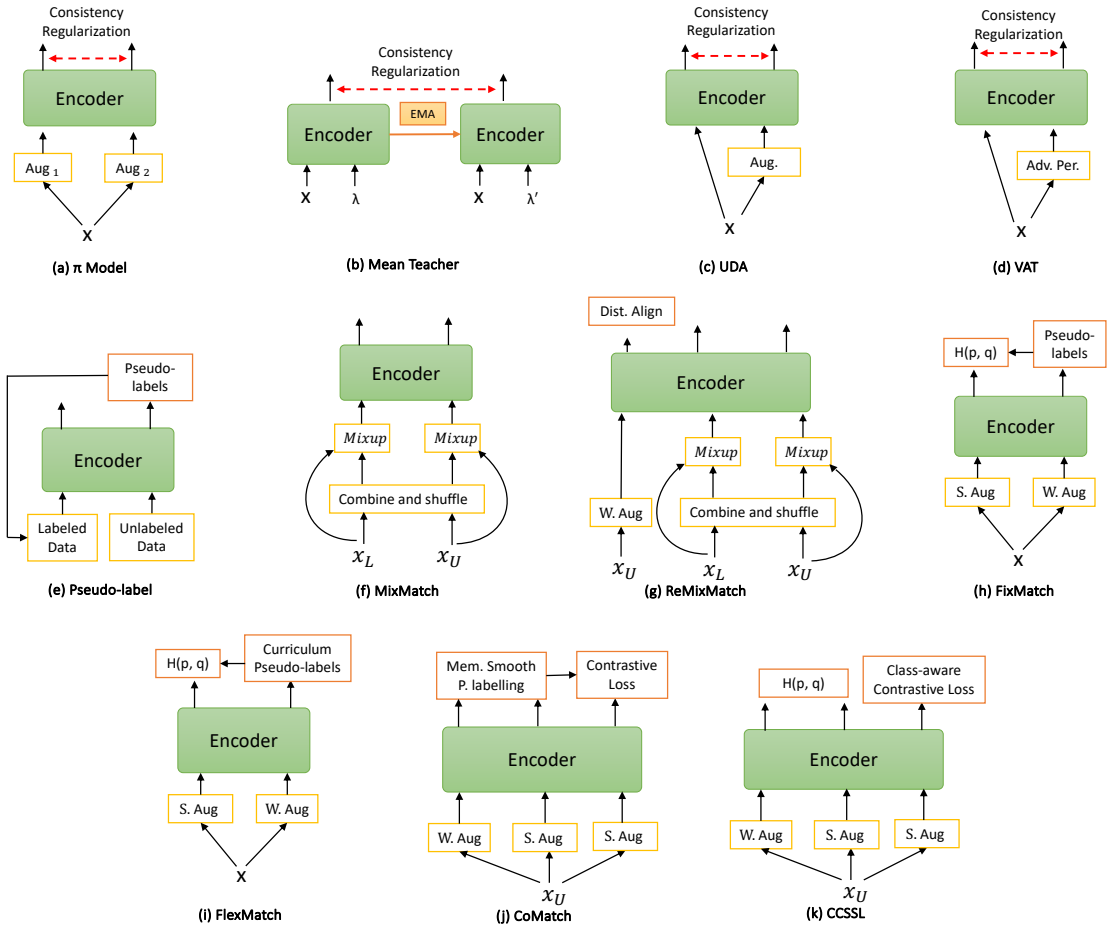


Fig. 2: Overview of the semi-supervised learning methods explored in this study. Here, Aug_i, *S. Aug*, *W. Aug* and *MixUp* refer to the ith augmentation of the input x, a strongly augmented image, a weekly augmented image, and an augmented image with *MixUp* operation. Consistency Regularization is different across methods, as defined in Eqs. 1, 2, 4, and 5 for Pi-Model, Mean Teacher, UDA, and VAT, respectively. *EMA* refers to the exponential moving average. Adv. Per. refers to adversarial perturbation. $H(p, q)$ is the cross-entropy loss. *Dist. Align* is the distribution alignment concept introduced in ReMixMatch. Curriculum pseudo-labels are generated by the concept of adaptive threshold in FlexMatch. *Mem. Smooth P. labelling* is the concept of memory-smoothed pseudo-labels introduced in CoMatch.

have focused on learning from OOD [12], [13] or unconstrained unlabelled data [14], [15]. In OOD SSL [12], [13], [22], samples in the unlabelled set belong to the same classes as the labelled set but come from a different source and, therefore, have a different data distribution. In some of the earlier works [12], the main idea was to incorporate an OOD detection module to identify and remove OOD samples from the unlabelled set, effectively focusing on learning from the ID unlabelled data. Finally, unconstrained SSL assumes that the unlabeled data are out-of-distribution relative to the labelled set. Additionally, this data is not necessarily limited to the classes present in the labelled set. CCSSL [22] proposed a method that uses class information from the labelled set along with contrastive learning to effectively learn from unconstrained unlabelled data. AuxMix [15] combined self-supervised learning with a novel entropy maximization technique to learn the representations from the unconstrained unlabelled data. UnMixMatch [14] employed hard augmentation (RandAugMix) for learning from labelled data, coupling it with contrastive learning and a rotation prediction task for learning from unconstrained unlabelled data.

2.3 Semi-supervised FER

Besides our previous work, which focused on benchmarking the most commonly used semi-supervised methods for FER [16], there are few other studies in this area. For instance, [34] investigated the use of Deep Belief Networks for semi-supervised FER and found that they produced a relative improvement over supervised methods. In [35], a Bayesian network was proposed for semi-supervised FER. More recently, [36] proposed an entropy-minimization method for semi-supervised FER, which introduced an adaptive confidence margin concept to partition the unlabelled data based on the confidence of pseudo-labels. The method was then trained on low- and high-confidence predictions separately. Furthermore, [36] explored the use of multi-modal data to learn from audio-visual signals in a semi-supervised setting. Progressive Teacher (PT) [37] introduced the concept of identifying and selecting useful samples for supervised

learning, along with a consistency loss on the unlabelled data. To address class distribution mismatch between labelled and unlabelled data, Rethink-Self-SSL [38] introduced a clustering concept that leveraged intra-cluster and inter-cluster distances to accurately identify out-of-distribution data. To mitigate the impact of false pseudo-labels on model performance, CFRN [39] introduced a feature dropout and emphasis module, enhancing its ability to discriminate between features effectively.

3 METHOD

In this section, we first discuss the problem setup for semi-supervised learning. This is followed by an overview of 11 semi-supervised methods explored in this study.

3.1 Problem Setup for Semi-supervised Learning

Let be given a small labelled set $D_l = \{(x_i^l, y_i^l)\}_{i=1}^N$, where N is the total number of samples and their corresponding class labels, and a large unlabelled set $D_u = \{(x_i^u)\}_{i=1}^M$, where M is the total number of unlabelled samples and $M \gg N$. Accordingly, semi-supervised learning aims to learn from labelled set D_l in a supervised setting while utilizing the unlabelled set D_u in an unsupervised setting to learn a better representation of the data. The performance is validated on a separate validation set $D_v = \{(x_i, y_i)\}_{i=1}^V$. There is no overlap between the sample in labelled, unlabelled, and validation sets, i.e., $D_l \cap D_u \cap D_v = \emptyset$.

3.2 Semi-Supervised Methods

3.2.1 Pi-Model

Pi-Model [17] is one of the earliest and most popular consistency regularization-based semi-supervised methods. While learning from the labelled samples in a supervised setting, Pi-model learns from the unlabelled set with a consistency regularization term. More specifically, it applies two augmentations on an unlabelled image and forces their predictions to be similar. Pi-Model also uses dropout and random max-pooling to add stochastic behaviour to the predictions. A schematic diagram of the Pi-Model is presented in Figure 2a. The consistency regularization loss of Pi-Model 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 sample transformations applied to an unlabelled sample x, f is the model, and θ represents the parameters of f.

3.2.2 Mean Teacher

Mean Teacher [18] is another consistency regularization-based semi-supervised method that is built upon Pi-model. However, Mean-teacher differs in the way it generates embeddings of two augmented samples. Instead of utilizing the online encoder (trainable encoder) to make the prediction for both images, Mean-teacher uses an exponential moving average (EMA) encoder to generate a prediction for the second image. This EMA of the student model is referred to as the teacher model, while the online encoder is called the student model. So, the Mean-teacher learns from unlabelled data by enforcing consistency between the predictions of the teacher and student models on two augmentations of the same sample through a regularization loss. A diagram demonstrating the Mean teacher method is presented in Figure 2b. The consistency regularization loss for Mean Teacher can be expressed 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 parameters of the teacher model. Finally, the weight update operation of the teacher model with an exponential moving average is formulated as:

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

where m is a momentum coefficient.

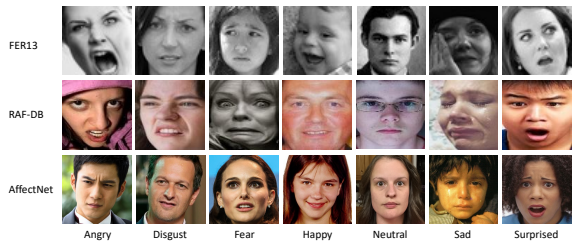


Fig. 3: Sample images from three main FER datasets: FER13, RAF-DB, and AffectNet.

3.2.3 UDA

UDA [9] is also based on the consistency regularization concept. The basic idea of UDA is similar to Pi-model but shows a large improvement in performance, only replacing the augmentation module. UDA replaces the usual augmentation module with advanced augmentation methods like AutoAugment [40], and RandAugment [31], resulting in dynamic and diverse sample augmentations. Figure 2c shows an overview of the UDA method. The consistency regularization loss of UDA can be expressed as:

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

where τ represents hard augmentations.

3.2.4 VAT

VAT [8] is similar to the Pi-model and UDA in terms of its regularization concept. However, instead of regularizing the embeddings of two augmented versions of the same sample, VAT uses adversarial perturbation as a different form of augmentation of the input sample. The overview of VAT is depicted in Figure 2d. The consistency regularization loss of VAT can be expressed as follows:

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

where γ^{adv} is the adversarial perturbation operation.

3.2.5 Pseudo-label

The Pseudo-label method, introduced in [7], is an entropy minimization-based method that presents a simple yet effective semi-supervised solution. Pseudo-label involves predicting the class probabilities for each of the unlabelled samples, which act as pseudo-labels for those images. If

the confidence of the prediction is high (low entropy), the pseudo-label of the unlabelled sample is treated as a label to train alongside labelled data. This pseudo-labelling concept has been used as a basis for several of the current state-of-the-art methods. A visual illustration of the Pseudo-label method is depicted in Figure 2e. The loss function of the Pseudo-label method is expressed 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 predicted pseudo-label for an unlabelled sample x_i^u, and λ is a coefficient to balance the weight of two loss terms.

3.2.6 MixMatch

MixMatch [19] is a semi-supervised method of the hybrid category that combines the concept of consistency regularization and entropy minimization. Similar to the entropy-based methods, MixMatch aims to generate low entropy predictions on the unlabelled data and also enforces consistency in its predictions similar to consistency-based methods. The novel component of MixMatch is the MixUp operation (an interpolation function) on labelled and unlabelled samples to generate mixed samples. Both entropy minimization and consistency regularization operations are applied to the mixed samples. The MixUp operation can be represented as:

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

where x_l and x_u are input samples from the labelled and unlabelled set, and α is the weight factor that balances the labelled and unlabelled components in the generated sample. MixMatch uses a beta distribution to randomly sample the value for alpha.

Another key innovation of MixMatch is the use of multiple augmentations on the unlabelled set and averaging the predictions on the augmented samples to generate the final prediction for that unlabelled sample. Let d_l' and d_u' be a batch of labelled and unlabelled samples after the MixUp operation. The MixMatch loss can be represented as:

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

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

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

where $H(.)$ is the cross-entropy loss, C is the total number of classes, and λ is the weight factor between the supervised and unsupervised loss terms. Figure 2f shows a visual illustration of the MixMatch method.

3.2.7 ReMixMatch

ReMixMatch [11] is a modified version of MixMatch that incorporates two new concepts: distribution alignment and augmentation anchoring. The former aims to ensure that the predictions made on the unlabelled data align with the distribution of the predictions on the labelled data. Augmentation anchoring replaces the consistency regularization of MixMatch and focuses on making the representation of strongly augmented samples similar to those of weakly augmented samples. This technique compares one weakly augmented sample against multiple strongly augmented samples. ReMixMatch also introduces a new strong augmentation method called CTAugment, which is more suitable in semi-supervised learning settings. Figure 2h provides a visual representation of the ReMixMatch.

3.2.8 FixMatch

FixMatch [10] is another hybrid semi-supervised learning method that shows impressive performance in many applications. For an unlabelled sample, FixMatch first applies weak augmentations and generates a prediction. FixMatch then considers this as a pseudo-label for a hard augmentation of the same sample if the confidence of this pseudo-label is beyond a threshold. Standard shift and flip augmentations are utilized for the weak augmentation module of FixMatch. FixMatch explores RandAugment [31] and CTAugment [11] as the hard augmentation module. The unsupervised loss of FixMatch can be represented as:

$$\mathcal{L}_u = \frac{1}{|d_u|} \sum_{x \in d_u} \mathbf{1}(max(q) \geq \tau)\ \mathcal{H}(\hat{q}, f(\mathcal{A}(x), \theta)) \tag{11}$$

where, τ is the threshold, q is the prediction on the weakly augmented sample, and $\hat{q} = arg\ max(q)$. A visual illustration of FixMatch is depicted in Figure 2h.

3.2.9 FlexMatch

FlexMatch [20] proposes an improvement over FixMatch with a curriculum learning concept for the threshold parameter. Rather than using a fixed τ for all classes, FlexMatch updates a class-specific threshold based on the learning status of that class. FlexMatch uses per-class accuracy as an indicator of the learning status of that class, which is calculated as:

$$\alpha(c) = \sum_n \mathbf{1}(max(q) \geq \tau)\ \mathbf{1}(argmax(q)) = c), \tag{12}$$

where n is the number of samples. A schematic diagram of FlexMatch is shown in Figure 2i.

3.2.10 CoMatch

CoMatch [21] introduces a graph contrastive learning concept built on Fixmatch. CoMatch jointly learns two representations of data that interact and improve with each other: class probabilities and low-dimensional embeddings. To reduce the errors in the predicted pseudo-labels, CoMatch uses the concept of memory-smoothed pseudo-labels, where label predictions are refined by considering similar data points in the embedding space. To learn better task-specific representations, CoMatch uses contrastive learning to encourage similar embeddings for samples with the same label. The Concept of CoMatch is depicted in Figure 2j.

3.2.11 CCSSL

Finally, CCSSL [22] is also built on FixMatch, and deals with the confirmation bias of pseudo-labels to improve performance on OOD unlabelled data. Unlike traditional pseudo-labelling approaches, CCSSL separates data into in-distribution and out-of-distribution categories. It then applies class-wise clustering to maintain efficient learning for

(a) Crop (b) Rotation (c) Color Jitter (d) Multiple

Fig. 4: Examples of hard augmentations.

known categories in the in-distribution data, while employing image-level contrastive learning on out-of-distribution data. Overall, CCSSL is designed as an add-on that can be easily integrated with existing pseudo-labeling methods, enhancing their effectiveness and making them more applicable in diverse real-world scenarios. The CCSSL method is illustrated in Figure 2k.

4 Experiments

In this section, we first provide details of the datasets used in this study. Following this, we present the experimental setup employed for the semi-supervised FER experiments. Finally, we present the main results for all the semi-supervised settings.

4.1 Datasets

In this study, we utilize a total of six datasets to conduct the Facial Expression Recognition (FER) experiments. For the main results, we use three datasets: FER13 [23], RAF-DB [41], and AffectNet [25]. In addition, for the experiments on unconstrained FER, we use the CelebA [28] dataset. Finally, for the FER experiments on limited data, we use the KDEF [26] and DDCF [27] datasets. Below, we provide a brief description of each dataset used in this study. Some samples from these datasets are shown in Figure 3.

FER13 [23] is a widely used dataset for FER that contains over 28.7K images of 7 emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral). The images have been collected from the internet and resized to 48×48 pixels.

RAF-DB [41] is a dataset for FER that contains around 15K images, with 12K images for training and 3K used for testing. This dataset has been annotated by 315 annotators, with each image annotated by around 40 annotators.

AffectNet [25] is a large-scale dataset for FER that contains around 284K images of 8 emotions (anger, contempt, disgust, fear, happiness, sadness, surprise, and neutral). This dataset has also been collected from the internet, and the images are of relatively high resolution.

CelebA [28] is a large-scale dataset of faces with aourd 202K images. This dataset does not necessarily contain images of expressive faces, and the images are also collected from the internet and are of relatively high resolution.

KDEF [26] is a small dataset for FER that contains images of 7 emotions, which have been captured in lab conditions. This dataset contains a total of approximately 5K images. The images are of relatively high resolution and have been captured from different angles.

DDCF [27] is another small dataset for FER that contains images of 8 emotions, which have been captured in lab conditions. There are nearly 6.5K images in this dataset. The images are of relatively high resolution and have been collected from different angles.

4.2 Implementation details

For a fair comparison between all the methods, we use the same encoder and training protocol. For the encoder, we use ResNet-50 [42]. All the experiments are reported for different numbers of labelled samples. For this, we randomly sample $N = n \times C$ images and their corresponding labels from the labelled set D_l, where C is the number of classes, and n is the number of samples per class. Specifically, we present all results for $n \in \{10, 25, 100, 250\}$. For all experiments, we report the average accuracy and standard deviations over *three* runs on different seeds. Following [10], we sample different random splits of data with different seeds.

We follow the implementation details of the original methods for method-specific parameters. For instance, we use a confidence cut-off value of 0.95 for FixMatch, a moving average weight of 0.999 for Mean-teacher, and a temperature value of 0.5 for methods that utilize sharpening distribution. For training, we use 2^{20} iterations with a batch size of 64 and SGD optimizer with a learning rate of 0.03, a momentum of 0.9, and a cosine learning rate decay scheduler.

In many recent semi-supervised methods, data augmentations are proven to be an essential component. In this context, a hard augmentation refers to a sequence of augmentations applied to a sample that results in an augmented sample that is visually distinguishable from the original input. In contrast, weak augmentation refers to a single or very low number of augmentations (≤ 2) applied to an input sample which does not change the sample drastically. Among the various hard augmentation modules mentioned in the literature, RandAugment [31] is the most commonly used one. This technique involves defining a sequence (up to 14) of augmentations that can be applied to an image, such as random crops, flips, and colour distortion. Then, a random subset of these augmentations is selected and applied to the image. We provide some examples of the augmentations used in RandAugment in Figure 4.

4.3 Semi-supervised FER with ID unlabelled data

4.3.1 Setup

This section presents the main result of 11 semi-supervised learning approaches on the FER13, RAF-DB, and AffectNet datasets for ID FER. As previously mentioned, the results are presented for 10, 25, 100, and 250 labelled samples for each emotion class. The remaining samples from each dataset are treated as unlabelled sets.

4.3.2 Performance

The main results for ID semi-supervised learning are presented in Table 1, which also includes the average accuracy across all settings (4 data splits of 3 datasets) for an overall understanding of the performance of each method. In summary, FixMatch appears to be the most successful semi-supervised method for ID data, as it outperforms other methods on 7 out of 12 settings and achieves the second-best result on the other two settings. FixMatch achieves an average accuracy of 53.41% across all settings, with a maximum standard deviation of 2.5% on FER-13 with 10 labelled samples per class. The second-best method, MixMatch, achieves an average accuracy of 47.88%, which is

TABLE 1: The performance of different semi-supervised methods with ID unlabelled data on FER13, RAF-DB, and AffectNet, when 10, 25, 100, and 250 labelled samples per class are used for training.

	FER13				RAF-DB				AffectNet				Avg. Acc
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 [17]	37.09±3.7	40.87±2.5	50.66±1.8	56.42±1.4	39.86±3.1	50.97±2.5	63.98±1.1	71.15±0.8	24.17±4.2	25.37±3.8	31.24±3.4	32.40±2.1	43.68
Pseudo-label [7]	32.79±3.9	36.04±2.7	49.21±1.9	54.88±1.5	58.31±3.5	39.11±2.6	54.07±1.7	67.40±0.9	18.00±4.4	21.05±3.0	33.05±3.6	37.37±2.3	41.77
Mean Teacher [18]	45.21±2.6	55.14±1.8	52.17±1.6	58.06±1.3	62.05±2.9	45.17±2.3	45.57±1.8	**76.85**±0.5	19.54±3.9	20.21±3.1	20.80±2.8	44.05±1.1	45.40
VAT [8]	24.95±3.8	**55.22**±2.0	51.55±1.7	55.64±1.4	63.10±3.1	45.82±2.4	62.05±1.5	59.45±1.0	17.68±4.3	35.02±3.4	37.68±3.0	37.92±2.0	45.50
UDA [9]	46.72±2.7	49.89±1.9	50.62±1.6	60.68±1.2	46.87±3.0	**53.15**±2.4	58.86±1.6	60.82±1.0	27.42±4.1	32.16±3.2	37.25±2.8	37.64±1.8	46.84
MixMatch [19]	45.69±2.6	46.41±1.8	55.73±1.5	58.27±1.2	36.34±3.2	43.12±2.5	64.14±1.0	73.66±0.4	**30.80**±3.0	32.40±3.1	39.77±2.7	48.31±1.6	47.88
ReMixMatch [11]	41.07±2.8	43.25±1.5	44.62±1.3	57.49±1.0	37.35±3.3	42.56±2.1	42.86±1.5	61.70±0.8	29.28±3.2	33.54±2.5	41.60±1.6	46.51±1.4	43.48
FixMatch [10]	**47.88**±2.5	49.90±1.7	**59.46**±1.0	**62.20**±0.5	**63.25**±1.0	52.44±2.2	64.34±0.9	75.51±0.3	30.08±1.9	**38.31**±1.2	**46.37**±1.0	**51.25**±0.6	**53.41**
FlexMatch [20]	39.77±2.5	42.88±1.7	51.14±1.3	56.06±1.0	40.51±3.0	42.67±2.1	50.75±1.5	61.70±0.8	17.20±4.1	19.80±3.0	22.34±2.6	29.83±1.7	39.55
CoMatch [21]	40.24±2.7	49.04±1.9	54.97±1.6	59.47±1.2	40.04±3.1	52.59±2.4	**68.05**±1.1	73.46±0.6	21.23±4.2	23.54±3.2	27.45±2.8	30.31±1.9	45.03
CCSSL [22]	40.23±2.8	45.36±1.8	57.01±1.4	61.77±1.0	50.59±2.8	51.30±2.2	63.79±1.4	74.93±0.7	16.89±4.3	21.34±3.1	24.46±2.7	28.94±1.8	44.71

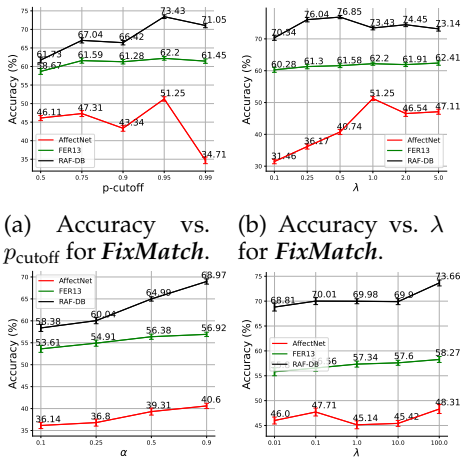
TABLE 2: Comparison between fully-supervised learning and semi-supervised learning with ID unlabelled data.

	Supervised		Semi-sup. (best)	
Dataset	**250/class**	**All data**	**250/class**	**All data**
FER13	53.58±1.1	64.57±0.9	62.20±0.5	65.15±0.5
RAF-DB	65.87±1.0	80.47±0.7	76.85±0.5	81.75±0.4
AffectNet	40.28±1.2	54.91±1.0	51.25±0.6	55.56±0.5

a considerable 5.53% drop in performance compared to FixMatch. Therefore, we can conclude that FixMatch is the most robust semi-supervised method for FER with ID unlabelled data.

4.3.3 *Sensitivity study*

All the experiments shown in the table above were conducted using default parameters for each algorithm, as reported in the original papers. In this subsection, we present a sensitivity study on the key hyper-parameters of the two best-performing methods in order to improve FER performance further. Figure 5 displays this study on the P_{cutoff} and λ values for FixMatch, as well as the α and λ values for MixMatch. In the FixMatch, the P_{cutoff} value determines the confidence threshold at which a predicted pseudo-label is considered as the label for an unlabelled image. The results shown in Figure 5a indicate that the best accuracy is achieved for a P_{cutoff} value of 0.95 for all datasets, which is consistent with the original FixMatch method. The λ value balances the weight of supervised and unsupervised loss in FixMatch. We conduct an experiment on λ (Figure 5b) and find that the optimal value varies for different datasets. While AffectNet and RAF-DB show better results for relatively smaller values of λ (1.0 and 0.5, respectively), FER13 achieves the best performance with higher values of λ, specifically 5.0. In the MixMatch method, α is the mixing coefficient used in the MixUp operation. The experiment on α (Figure 5c) indicates that higher values of α generally produce better results for all datasets, with the best performance obtained with $\alpha = 0.9$. Finally, in the experiment on the λ value of MixMatch (Figure 5d), we observe improvements for larger values of λ, with the best accuracy achieved when λ is set to *100* for all three datasets.



(a) Accuracy vs. p_{cutoff} for ***FixMatch***.
(b) Accuracy vs. λ for ***FixMatch***.
(c) Accuracy vs. α for ***MixMatch***.
(d) Accuracy vs. λ for ***MixMatch***.

Fig. 5: Sensitivity study of various parameters for two of the best semi-supervised methods on ID unlabelled data.

4.3.4 *Discussion*

Table 2 provides a summary of the best results obtained for each dataset using 250 labels per class and compares the performances with fully supervised training with an equal amount of labelled samples. Additionally, the table shows the results obtained with fully supervised learning using all the data, including their labels from the original dataset. For FER13, the semi-supervised method achieves an 8.83% improvement over fully supervised training with the same amount of labelled data (250 labels per class, for a total of 1750 out of 28K images) and is only 2.16% lower than fully supervised training with all labelled data (28K images). Similarly, for RAF-DB, the semi-supervised method obtains a 10.98% improvement over fully supervised training with an equal amount of labelled samples (250 labels per class, for a total of 1750 out of 12k images) and is only 3.62% lower than fully supervised training with all samples being labelled (12K images). Finally, for AffectNet, the semi-supervised method achieves a 10.98% improvement over the fully supervised training with the same amount of labelled data (250 labels per class, for a total of, 1750 out of 284k images) and is only 3.66% lower than the fully supervised training with all labelled data. Based on this summary, we can conclude that semi-supervised methods are able

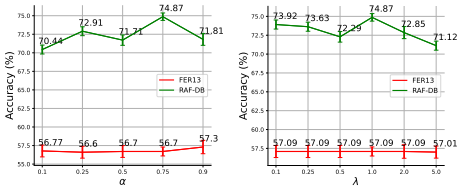
TABLE 3: The performance of different semi-supervised methods with OOD unlabelled data on FER13 and RAF-DB, when 10, 25, 100, and 250 labelled samples per class are used for training.

	FER13				RAF-DB				Avg. Acc.
Method / m	10 labels	25 labels	100 labels	250 labels	10 labels	25 labels	100 labels	250 labels	
Pi-Model [17]	24.71±3.8	29.02±2.6	43.63±1.9	53.58±1.1	38.62±3.1	39.73±2.5	46.81±1.7	66.53±0.6	42.83
Mean Teacher [18]	25.16±2.8	27.61±1.9	47.90±1.6	54.82±1.0	38.62±2.5	37.65±2.0	52.09±1.5	65.45±0.7	43.66
VAT [8]	23.95±3.1	28.49±2.0	43.37±1.7	52.98±1.1	38.62±2.6	38.85±2.1	49.58±1.6	61.60±0.9	42.18
Pseudo-label [7]	24.09±3.0	36.04±1.8	47.20±1.5	53.23±1.0	38.62±2.6	38.62±2.0	49.84±1.6	63.17±0.8	43.85
UDA [9]	27.60±2.7	38.20±1.6	52.16±1.4	56.13±1.0	39.37±2.9	41.79±2.3	60.95±1.3	65.71±0.7	47.74
MixMatch [19]	31.64±2.6	37.66±1.7	49.68±1.2	56.39±0.9	43.02±2.0	51.14±1.6	63.04±1.1	70.37±0.5	50.37
ReMixMatch [11]	**33.31**±1.9	**44.15**±1.2	53.15±1.0	**58.48**±0.8	**48.37**±1.1	**58.47**±0.9	67.47±0.5	**74.87**±0.5	**54.73**
FixMatch [10]	29.60±2.5	38.98±1.6	52.44±1.1	57.40±0.8	23.99±2.8	45.86±1.9	61.15±1.2	64.80±0.6	46.78
FlexMatch [20]	31.89±2.1	41.0±1.4	50.43±1.2	56.17±0.8	46.87±1.2	53.16±1.0	**69.59**±0.5	72.59±0.3	52.71
CCSSL [22]	30.05±2.3	39.09±1.5	**53.32**±1.1	57.77±0.7	47.82±1.1	56.52±0.9	69.33±0.5	69.07±0.3	52.87

TABLE 4: Comparison between fully-supervised learning and semi-supervised learning with OOD unlabelled data.

	Supervised		Semi-sup.		
Dataset	**All data**	**ID/250**	**ID/250**	**OOD/250**	**OOD/All**
FER13	64.57±0.9	53.58±1.1	62.20±0.5	58.48±0.8	70.40±0.4
RAF-DB	80.47±0.7	65.87±1.0	76.85±0.5	74.87±0.5	83.73±0.2

to achieve a significant improvement over fully supervised training with the same amount of labelled data, and can achieve comparable performance to fully supervised training on large amounts of labelled samples. Finally, we show the results for using all the labelled data along with ID unlabelled data. While the unlabelled set is the same size as the labelled set, training in a semi-supervised setting provides improvement over fully-supervised learning.



(a) Accuracy vs. α for ***ReMixMatch***.
(b) Accuracy vs. λ for ***ReMixMatch***.
(c) Accuracy vs. P_{cutoff} for ***CCSSL***.
(d) Accuracy vs. λ for ***CCSSL***.

Fig. 6: Sensitivity study of various parameters for two of the best semi-supervised methods on OOD unlabelled data.

4.4 Semi-supervised FER with OOD unlabelled data

4.4.1 Setup

In this section, we present the results of OOD semi-supervised learning. In this setting, the unlabelled data consists of images belonging to the same expression categories as the labelled data but originating from different sources and therefore having a different distribution. Specifically, we use a pre-defined number of samples (10, 25, 100, or 250 labelled samples per class) from the training set of FER-13 and RAF-DB as the labelled data and the complete training set (images without labels) of AffectNet as unlabelled (OOD) data. The accuracy is reported on the validation set of the FER-13 and RAF-DB datasets, respectively.

4.4.2 Performance

The performance of different semi-supervised methods with OOD unlabelled data are presented in Table 3. We can draw two key observations from this table. Firstly, we can observe a significant drop in the performance of all methods compared to when ID unlabelled data are used. These findings are consistent with previous works in the OOD semi-supervised literature [43], [44]. For instance, in our study, we find that FixMatch achieves a 62.20% accuracy on FER-13 (250 labels) with ID unlabelled data, but this drops to 57.4% when OOD data are utilized. The performance drop is even more substantial with smaller labelled set sizes. For example, when only 10 samples per class are available on FER-13, the FixMatch performance drops from 47.88% to only 29.6%. The second observation is that the best-performing method is different for OOD semi-supervised learning. While FixMatch and MixMatch were the top two methods for ID semi-supervised learning, ReMixMatch, and CCSSL are the top two methods for OOD semi-supervised learning. Both of these methods perform significantly better than FixMatch.

4.4.3 Sensitivity study

Next, we conduct a sensitivity analysis on the two best-performing methods, ReMixMatch and CCSSL, for OOD semi-supervised learning. Since these methods were originally designed for ID semi-supervised learning in general computer vision applications, it is crucial to explore their performance for different hyperparameters under the OOD setting. Figure 6 presents the results of this study. We examine the impact of α and λ for ReMixMatch and P_{cutoff} and λ for CCSSL. In the experiment with ReMixMatch's α values (Figure 6a), we discover that the best results for both datasets are achieved with a value of 0.75, which is also the default in the original ReMixMatch paper. We also find that the optimal value for λ was 1.0 for both datasets. In the experiments on CCSSL's P_{cutoff} value (Figure 6c), we

TABLE 5: The performance of different semi-supervised methods with unconstrained unlabelled data on FER13, RAF-DB, and AffectNet, when 10, 25, 100, and 250 labelled samples per class are used for training.

	FER13				RAF-DB				AffectNet				Avg. Acc
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	
Pi-Model [17]	22.28±3.8	27.03±2.5	44.96±1.5	54.79±1.0	38.62±3.1	38.62±3.0	52.18±1.5	65.74±0.6	18.00±4.3	18.74±3.2	19.66±2.8	35.57±1.7	36.35
Mean Teacher [18]	24.88±3.0	29.77±2.0	43.66±1.6	53.22±1.1	36.18±3.1	38.62±2.9	50.59±1.6	66.49±0.6	17.69±4.3	19.11±3.1	21.34±2.7	38.34±1.4	36.66
VAT [8]	21.94±3.0	30.05±2.0	43.56±1.6	52.42±1.1	38.62±3.1	41.17±2.5	51.76±1.6	62.32±0.8	17.86±3.3	19.23±3.1	20.89±2.7	26.34±1.8	35.51
Pseudo-label [7]	23.36±3.0	33.51±1.8	47.70±1.5	53.57±1.0	38.62±3.1	37.45±2.0	49.54±1.6	63.40±0.8	17.29±3.3	19.54±3.0	21.06±2.6	24.03±1.7	35.76
UDA [9]	24.88±2.9	33.99±1.8	49.40±1.4	52.97±1.0	27.05±3.0	43.94±1.9	52.41±1.5	62.87±0.7	16.91±3.9	17.60±3.0	28.31±1.9	33.14±1.6	36.96
MixMatch [19]	31.43±2.7	39.44±1.7	51.99±1.2	56.26±0.9	42.37±1.1	50.52±0.9	60.01±0.5	68.84±0.4	25.14±3.0	26.26±2.0	33.40±1.6	39.23±1.2	43.74
ReMixMatch [11]	32.49±2.3	**42.05**±1.4	**52.83**±1.1	**58.04**±0.8	**49.15**±1.0	**54.60**±0.8	**64.50**±0.5	**70.70**±0.5	**26.29**±2.7	**29.23**±1.8	**37.31**±1.4	**40.40**±1.0	**46.47**
FixMatch [10]	26.58±2.6	35.16±1.7	49.29±1.3	54.30±1.0	33.41±2.9	39.99±1.9	52.93±1.4	60.59±0.7	15.77±3.9	17.23±3.0	28.46±1.9	31.23±1.6	37.08
FlexMatch [20]	25.42±2.7	37.21±1.7	50.21±1.3	54.83±1.0	36.86±2.8	49.45±1.8	59.62±1.2	64.05±0.7	24.40±3.3	28.06±1.8	32.26±1.7	36.74±1.2	41.59
CoMatch [21]	**33.01**±2.6	40.90±1.6	50.67±1.3	55.83±1.0	27.95±3.0	32.53±2.0	64.11±1.1	68.17±0.6	23.40±3.4	27.50±2.0	32.33±1.8	35.79±1.3	41.02
CCSSL [22]	24.69±2.8	39.51±1.7	51.14±1.3	55.63±1.0	31.32±2.9	47.26±1.8	55.25±1.3	61.90±0.7	16.51±2.3	25.11±1.9	25.00±1.7	33.71±1.3	38.92

TABLE 6: Comparison between fully-supervised learning and semi-supervised learning with unconstrained unlabelled data.

	Supervised		Semi-sup.		
Dataset	**All data**	**ID/250**	**ID/250**	**Unc./250**	**Unc./All**
FER13	64.57±0.9	53.58±1.1	62.20±0.5	58.04±0.8	70.50±0.4
RAF-DB	80.47±0.7	65.87±1.0	76.85±0.5	70.70±0.5	82.75±0.2
AffectNet	54.91±1.0	40.28±1.2	51.25±0.6	40.40±1.0	57.65±0.6

observe that lower values of this parameter yields better results. Specifically, RAF-DB produces the best result with a value of *0.5*, while FER13 achieves the best result with a value of *0.75*. Finally, for λ values (Fig 6d), we obtain the best accuracy for different values for RAF-DB and FER13 datasets, with *1.0* and *0.25* respectively.

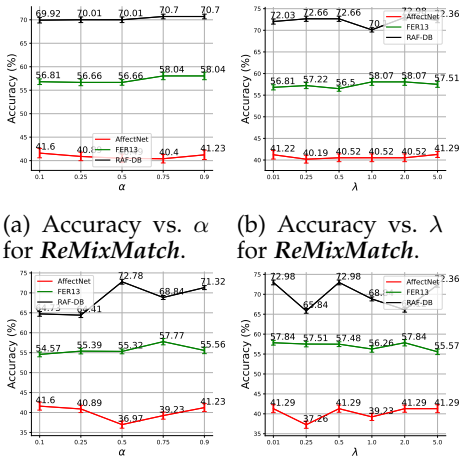
4.4.4 Discussion

Table 4 summarizes the results of semi-supervised FER using OOD unlabelled data and compares them to both fully-supervised and semi-supervised methods with ID unlabelled data. All results are shown for 250 labelled samples per class. We observe a drop in performance of 1.98% and 3.72% for RAF-DB and FER13 datasets, respectively, compared to ID semi-supervised learning. However, the performance of semi-supervised learning with OOD unlabelled samples is still better than fully-supervised learning by 9.0% and 4.9% for the two datasets, respectively. Therefore, we can conclude that learning with the presence of OOD data is still a better choice than relying on a fully supervised setting alone when presented with limited labelled data. Finally, we find that using all the labelled data along with the OOD unlabelled data provides considerable improvement (5.83% and 3.26% for FER-13 and RAF-DB) over fully supervised learning with all the labelled data.

4.5 Semi-supervised FER with *unconstrained* unlabelled data

4.5.1 Setup

In this section, we present the results of semi-supervised FER using unconstrained unlabelled data, which is considered the most challenging setting for semi-supervised learning. Here, the unlabelled data are obtained from a different



(a) Accuracy vs. α for ***ReMixMatch***. (b) Accuracy vs. λ for ***ReMixMatch***.

(c) Accuracy vs. α for ***MixMatch***. (d) Accuracy vs. λ for ***MixMatch***.

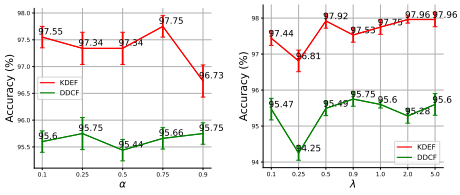
Fig. 7: Sensitivity study of various parameters for two of the best semi-supervised methods on unconstrained unlabelled data.

source than the labelled data and do not necessarily contain images of known classes. In unconstrained semi-supervised learning, we use a pre-defined number of samples (10, 25, 100, or 250 labelled samples per class) from the training set of FER-13, RAF-DB, and AffectNet as the labelled data and complete training set of AffectNet (do not contain labels) as unlabelled data. The accuracy is reported on the validation set of the FER-13, RAF-DB, and AffectNet, respectively.

4.5.2 Performance

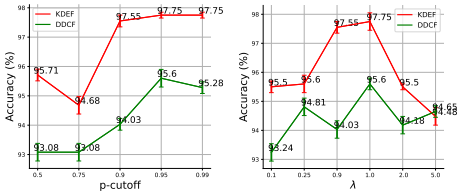
Table 5 shows the results of different semi-supervised methods with unconstrained unlabelled data. We make the following two observations from this table. Firstly, The best-performing method on unconstrained unlabelled data is similar to that of the OOD setting. Again, ReMixMatch achieves the best average results compared to the other methods. However, the second-best method is different from the OOD setting (CCSSL), and is now MixMatch. The performance of the rest of the methods is significantly lower than these two methods. Secondly, the performance of semi-supervised methods is much lower than training with ID data but not too far away from OOD data. For example, the performance of ReMixMatch on FER-13 with

just 10 labelled samples on unconstrained unlabelled data is 8.58% lower than training with ID labelled data, but only 0.82% lower than on OOD data. This indicates that **for FER, unconstrained unlabelled data can result in competitive semi-supervised performance to OOD samples.** This is an important finding since it is more convenient, practical, and economical to collect large unconstrained unlabelled data than ID data or even OOD data of expressive faces.



(a) Acc. vs. α for ***ReMixMatch***.

(b) Acc. vs. λ for ***ReMixMatch***.



(c) Acc. vs. P_{cutoff} for ***FixMatch***.

(d) Acc. vs. λ for ***FixMatch***.

Fig. 8: Sensitivity study of various parameters for two of the best semi-supervised methods on small unlabelled data.

4.5.3 Sensitivity Study

Next, we present a sensitivity study on the main hyper-parameters of the two best methods for unconstrained semi-supervised learning (ReMixMatch and MixMatch). The results are presented in Figure 7.

The first experiment on ReMixMatch's α values shows that different datasets have different optimal values. Specifically, the optimal values for RAF-DB, FER13, and AffectNet are 0.5, 0.75, and 0.1, respectively (Figure 7a). Similarly, optimal values for λ experiments are different for each dataset (Figure 7b). In this case, the optimal values for RAF-DB, FER13, and AffectNet are 2.0, 1.0, and 2.0, respectively. MixMatch experiments follow a similar pattern of having different optimal values. For α, the best values are 0.5, 0.75, and 0.1 (Figure 7c), and for λ (Figure 7d), the optimal values are 0.5, 2.0, and 0.5 for RAF-DB, FER13, and AffectNet, respectively. In conclusion, hyper-parameters are generally specific to each dataset when learning from unconstrained unlabelled data.

4.5.4 Discussion

Table 6 summarizes the results of learning from unconstrained unlabelled data. The key takeaway from this table is that although semi-supervised learning with unconstrained data achieves lower performance compared to ID semi-supervised learning, it still outperforms fully supervised learning when an equal number of labelled samples are used. For instance, FER13 achieves an accuracy of 53.58% with fully supervised learning, while the same dataset reaches 58.04% accuracy through the utilization of semi-supervised learning with unconstrained unlabelled data. We also report the results for using the whole dataset as the labelled data, and unconstrained unlabelled data. The results from this study show that using semi-supervised learning with unconstrained unlabelled data considerably improves performance over fully-supervised learning. Here, the improvements are 5.93%, 2.28%, and 2.73% for FER-13, RAF-DB, and AffectNet, respectively.

4.6 Semi-supervised FER with small ID unlabelled data

4.6.1 Setup

In this section, we discuss the results of semi-supervised learning with small ID unlabelled data. While gathering substantial quantities of unlabelled ID data can pose challenges, obtaining a smaller set of unlabelled ID data might be more feasible in some cases. Consequently, we also conduct experiments with small but ID unlabelled data for completeness. To this end, we utilize small datasets collected in lab environments. More specifically, we conduct the experiments on KDEF [26] and DDCF [27] datasets, and present the results for a similar number of labelled samples as in previous experiments, with the remaining samples being used as the unlabelled set. Performances are reported on the validation set of the corresponding datasets.

4.6.2 Performance

Table 7 shows the main results of different semi-supervised methods in this setting. ReMixMatch again performs the best across all methods. With ID data in this setting, FixMatch again performs well and shows the second-best average accuracy. The average accuracy for ReMixMatch and FixMatch are 73.09% and 71.76%, respectively.

4.6.3 Sensitivity study

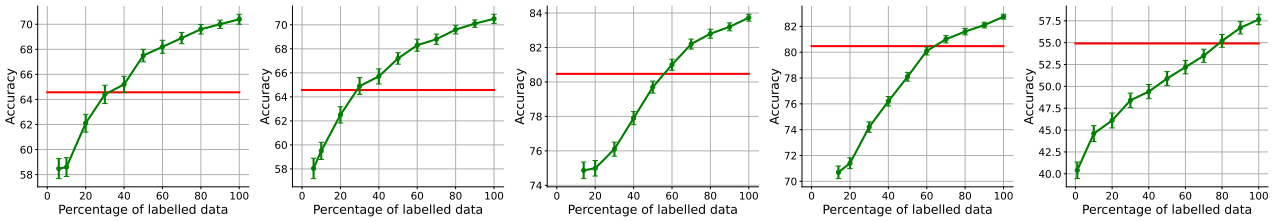
We also analyze the sensitivity of the two best-performing methods, ReMixMatch and FixMatch, in this setting. Figure 8 illustrates the results of this study. The experiment on the α parameter of ReMixMatch (Figure 8a) shows that higher values lead to better accuracy. The optimal values are found to be 0.9 and 0.75 for DDCF and KDEF datasets, respectively. Similarly, for the λ parameter (Figure 8b), the best results are obtained for 0.9 and 2.0 for DDCF and KDEF datasets. On the other hand, the experiments on FixMatch reveal that the default parameters of the original FixMatch yield the best accuracy for both datasets. Specifically, a P_{cutoff} value of 0.95 (Figure 8c) and a λ value of 1.0 (Figure 8d) achieve the best performances.

4.7 General discussions

In this section, we provide a discussion on some of our key observations. More specifically, we present a comparison between different unlabelled settings of SSL, a comparison to supervised learning, and a few insights for improving SSL for FER.

TABLE 7: The performance of different semi-supervised methods with small unlabelled data on KDEF and DDCF, when 10, 25, 100, and 250 labelled samples per class are used for training.

	KDEF				DDCF				Avg. Acc
Method / m	10 labels	25 labels	100 labels	250 labels	10 labels	25 labels	100 labels	250 labels	
Pi-Model [17]	31.90±2.0	58.90±1.1	85.28±0.7	94.27±0.3	19.81±2.6	30.82±1.8	77.20±0.9	88.99±0.4	60.90
Mean Teacher [18]	29.24±2.1	50.31±1.3	82.82±0.8	91.62±0.4	18.40±2.7	30.03±1.9	74.37±1.0	89.94±0.4	58.34
VAT [8]	24.54±2.2	44.79±1.4	79.75±0.9	91.21±0.7	23.58±2.5	40.57±1.8	72.01±1.1	88.99±0.5	58.18
Pseudo-label [7]	31.08±2.1	53.37±1.3	80.98±0.8	93.25±0.4	29.40±2.5	41.19±1.8	77.83±1.0	86.79±0.7	61.74
UDA [9]	**36.81**±1.9	46.42±1.2	88.14±0.6	97.55±0.2	30.35±2.5	80.82±0.9	86.79±0.7	93.71±0.4	70.07
MixMatch [19]	28.83±2.0	63.39±1.1	91.21±0.6	93.87±0.4	14.78±2.6	84.28±0.7	89.15±0.5	93.40±0.3	69.86
ReMixMatch [11]	29.86±1.9	**67.08**±1.0	**93.46**±0.4	**97.75**±0.2	25.79±2.5	**85.53**±0.7	**89.62**±0.5	**95.60**±0.3	**73.09**
FixMatch [10]	33.54±2.0	54.60±1.2	89.16±0.6	97.55±0.3	38.52±2.6	79.40±0.9	86.64±0.8	94.65±0.3	**71.76**
FlexMatch [20]	27.61±2.1	40.08±1.3	93.05±0.5	96.32±0.3	39.62±2.6	24.37±1.9	86.01±0.7	93.71±0.6	62.60
CoMatch [21]	21.08±2.2	52.12±1.3	85.05±0.7	92.16±0.4	**44.94**±2.6	63.52±1.8	87.18±0.7	90.70±0.5	67.09
CCSSL [22]	19.43±2.2	37.83±1.4	86.50±0.9	94.27±0.5	14.15±2.7	55.50±1.9	86.01±0.8	94.65±0.4	61.04



(a) FER13 with OOD. (b) FER13 with Unc. (c) RAF-DB with OOD. (d) RAF-DB with Unc. (e) AffectNet with Unc.

Fig. 9: Performance for different percentages of labelled data. Red line indicates fully-supervised learning with all samples.

4.7.1 Comparing different unlabelled settings

To better understand the overall behaviour of SSL using different categories of unlabeled data (ID vs. OOD vs. unconstrained), we present the performance of the *best-performing* method for each setting on FER-13 in Table 8. The results demonstrate that expectedly, the highest performance (62.20%) can be achieved with ID unlabelled data when a limited number of labelled samples (250) are available. We also observe that with limited OOD and unconstrained unlabelled data, similar performances can be achieved, underperforming the ID setting. Interestingly, the results show that when the amount of unlabelled data is considerably scaled, unconstrained unlabelled data are in fact more beneficial for SSL, in comparison to ID unlabelled data. Moreover, the similarity between OOD and unconstrained performance persists. These results indicate that using free-living data irrespective of their potential data or class distributions is a viable and effective approach for SSL when sufficiently large amounts of unlabelled data can be collected.

4.7.2 Comparison to supervised learning

To understand the impact of the amount of labelled data on SSL, we perform a detailed analysis by using different percentages of labelled data. As seen in Figure 9, we expectedly observe that more labelled data help the model in learning better representations. However, the figure shows that the added benefit of using more labelled samples decreases as more and more labelled samples are incorporated. Additionally, in Figure 9, we present the performance of fully-supervised learning (shown in the figures with a red line) in comparison to SSL. An interesting observation from this analysis is that irrespective of the type of unlabelled data used, SSL always outperforms fully supervised learning given a sufficient number of labelled samples. This further demonstrates the effectiveness of SSL with unconstrained/OOD data and the viability of reducing reliance on labels given the availability of unlabelled data.

4.7.3 Insights for improving SSL

While applying existing SSL methods to FER, we noticed that directly transferring their default setups for this purpose does not yield optimal performance. To improve the performance of existing methods, we conduct a comprehensive study to identify FER-specific best practices for various aspects of the SSL methods. Furthermore, we find the different methods perform differently in different semi-supervised settings studied in the paper. Overall, our study reveals the following key insights into semi-supervised learning for FER. First, when learning from more challenging learning scenarios, such as OOD, unconstrained, and small unlabelled data, the unsupervised loss plays a more critical role. Thus, increasing the value of the unlabelled loss factor (λ) improves performance. Secondly, considering the important role of unlabeled data and acknowledging that large volumes of such data can greatly enhance performance, we foresee a new perspective toward developing SSL for FER. Specifically, we anticipate that integrating *strong* unsupervised methods with small/modest supervised frameworks can result in robust and generalized frameworks, leading to the creation of more scalable SSL techniques in the area.

TABLE 8: Comparison across different settings on FER-13.

Setup	Accuracy (%) 250 labels	All
No unlabelled data	53.58$\pm$1.1	64.57$\pm$0.9
SSL (ID)	**62.20**$\pm$0.5	65.15$\pm$0.5
SSL (OOD)	58.48$\pm$0.8	70.40$\pm$0.8
SSL (Unconstrained)	58.04$\pm$0.8	**70.50**$\pm$0.4

5 Conclusion

This research offers a comprehensive analysis of 11 semi-supervised methods for FER. The study evaluates the performance of these methods in various unlabelled data scenarios, including ID, OOD, unconstrained, and very small unlabelled sets. Our primary finding is that FixMatch is the most effective semi-supervised method for learning from ID unlabelled data. However, for all other real-world scenarios (OOD, unconstrained, and small set), ReMixMatch consistently outperforms other semi-supervised methods. Another noteworthy finding is that semi-supervised learning from any data scenario produces better results in comparison to fully-supervised learning from the same number of labelled samples. When learning from an ID unlabelled set, semi-supervised methods can produce a performance improvement of up to 11% over the fully-supervised method. Although compared to ID, performance is generally lower for both OOD and unconstrained unlabelled data, these methods still outperform fully-supervised learning. Interestingly, the unconstrained setting, despite being significantly more challenging than ID or OOD, underperforms OOD by only a small margin. This is a significant observation since collecting large amounts of unconstrained unlabelled data is considerably easier and more practical than collecting ID or even OOD (but constrained) data. Overall, we anticipate that this research will serve as a useful guide for further investigation into semi-supervised learning in the context of FER, as well as other domains. One limitation of this study is that we primarily focused on facial expression datasets with static images of macro-expression. Exploring the effectiveness of semi-supervised learning for dynamic or micro expressions remains an interesting area for future research. While our study highlights the promise of unconstrained unlabelled data on the overall performance of semi-supervised FER, further investigation is needed to understand the impact of data quality and potential biases within such datasets.

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] M. Kolahdouzi, A. Sepas-Moghaddam, and A. Etemad, "Facetoponet: Facial expression recognition using face topology learning," *IEEE Transactions on Artificial Intelligence*, 2022.

[2] ——, "Face trees for expression recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2021, pp. 1–5.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *International Conference on Learning Representations*, 2020.

[12] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura, "Classification-reconstruction learning for open-set recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4016–4025.

[13] L.-Z. Guo, Z.-Y. Zhang, Y. Jiang, Y.-F. Li, and Z.-H. Zhou, "Safe deep semi-supervised learning for unseen-class unlabeled data," in *International Conference on Machine Learning*, 2020, pp. 3897–3906.

[14] S. Roy and A. Etemad, "Scaling up semi-supervised learning with unconstrained unlabelled data," in *AAAI Conference on Artificial Intelligence*, 2024.

[15] A. Banitalebi-Dehkordi, P. Gujjar, and Y. Zhang, "Auxmix: semi-supervised learning with unconstrained unlabeled data," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3999–4006.

[16] S. Roy and A. Etemad, "Analysis of semi-supervised methods for facial expression recognition," in *IEEE International Conference on Affective Computing and Intelligent Interaction*, 2022, pp. 1–8.

[17] 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.

[18] 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.

[19] 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.

[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.

[21] J. Li, C. Xiong, and S. C. Hoi, "Comatch: Semi-supervised learning with contrastive graph regularization," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9475–9484.

[22] F. Yang, K. Wu, S. Zhang, G. Jiang, Y. Liu, F. Zheng, W. Zhang, C. Wang, and L. Zeng, "Class-aware contrastive semi-supervised learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 421–14 430.

[23] 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*, 2013, pp. 117–124.
[24] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2852–2861.
[25] 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.
[26] D. Lundqvist, A. Flykt, and A. Öhman, "The karolinska directed emotional faces (kdef)," *CD ROM from Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet*, vol. 91, no. 630, pp. 2–2, 1998.
[27] K. A. Dalrymple, J. Gomez, and B. Duchaine, "The dartmouth database of children's faces: Acquisition and validation of a new face stimulus set," *PloS One*, vol. 8, no. 11, p. e79131, 2013.
[28] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision*, December 2015.
[29] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.
[30] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 557–11 568.
[31] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703.
[32] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," in *International Conference on Learning Representations*, 2020.
[33] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.
[34] A. R. Kurup, M. Ajith, and M. M. Ramón, "Semi-supervised facial expression recognition using reduced spatial features and deep belief networks," *Neurocomputing*, vol. 367, pp. 188–197, 2019.
[35] I. Cohen, N. Sebe, F. G. Cozman, and T. S. Huang, "Semi-supervised learning for facial expression recognition," in *5th ACM SIGMM international workshop on Multimedia information retrieval*, 2003, pp. 17–22.
[36] H. Li, N. Wang, X. Yang, X. Wang, and X. Gao, "Towards semi-supervised deep facial expression recognition with an adaptive confidence margin," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4166–4175.
[37] J. Jiang and W. Deng, "Boosting facial expression recognition by a semi-supervised progressive teacher," *IEEE Transactions on Affective Computing*, 2021.
[38] B. Fang, X. Li, G. Han, and J. He, "Rethinking pseudo-labeling for semi-supervised facial expression recognition with contrastive self-supervised learning," *IEEE Access*, 2023.
[39] H. Sun, C. Pi, and W. Xie, "Semi-supervised facial expression recognition by exploring false pseudo-labels," in *IEEE International Conference on Multimedia and Expo*, 2023, pp. 234–239.
[40] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 113–123.
[41] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2852–2861.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[43] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," *Advances in Neural Information Processing Systems*, 2018.
[44] J.-C. Su, Z. Cheng, and S. Maji, "A realistic evaluation of semi-supervised learning for fine-grained classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 966–12 975.

Shuvendu Roy is a Ph.D. student at Department of Electrical and Computer Engineering, and Ingenuity Labs Research Institute, at Queen's University in Canada. He received his B.Sc. in Computer Science and Engineering from Khulna University of Engineering & Technology, Bangladesh. His current research is focused on computer vision with deep learning, as well as self-supervised and semi-supervised learning.

Ali Etemad is an Associate Professor at the Department of Electrical and Computer Engineering, Queen's University. He holds an endowed professorship of Mitchell Professor in AI for Human Sensing & Understanding. He leads the Ambient Intelligence and Interactive Machines (Aiim) lab. He received his M.A.Sc. and Ph.D. degrees in Electrical and Computer Engineering from Carleton University, Ottawa, Canada, in 2009 and 2014, respectively. His main areas of research are machine learning and deep learning focused on human-centered applications with wearables, smart devices, and smart environments. Prior to joining Queen's, he held several industrial positions as lead scientist. He has published over 160 papers in top venues in the area, is a co-inventor of 10 patents, and has given over 25 invited talks at different venues. Dr. Etemad is an Associate Editor for IEEE Transactions on Affective Computing and IEEE Transactions on Artificial Intelligence. He has served as a PC member/reviewer, and has held organizing roles at various venues. He has received a number of awards including Supervisor of the Year Award (at Queen's), Instructor of the Year Award (at Queen's), and several Best Paper Awards (e.g., at ACM ICMI'23). Dr. Etemad's lab and research program have been funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada, Ontario Centers of Excellence (OCE), Canadian Foundation for Innovation (CFI), Mitacs, and other organizations, as well as the private sector.