Timid Semi–supervised Learning for Face Expression Analysis

Mihai Badea[a], Corneliu Florea[a,*], Andrei Racoviţeanu[a], Laura Florea[a], Constantin Vertan[a]

[a]*Image Processing and Analysis Laboratory (LAPI), University Politehnica of Bucharest, Romania,* {mbadea;aracoviteanu}@imag.pub.ro
{corneliu.florea;laura.florea;constantin.vertan}@upb.ro

Abstract

In the last years, semi–supervised learning has been proposed as a strategy with high potential for improving machine learning capabilities. Face expression recognition may highly benefit from such a technique, as accurate labeling is both difficult and costly, whereas millions of unlabeled images with human faces are available on the Internet, but without annotations. In this paper we evaluate the benefits of semi–supervised learning in the practical scenarios of face expression analysis. Our conclusion is that better performance is indeed achievable, but by methods that put a distinct emphasis on the diversity of exploring patterns in the unlabeled data domain. The evaluation is carried on multiple tasks such as detecting Action Units on EmotioNet, assessing Action Units intensity on the spontaneous DISFA database and, respectively, recognizing expressions on static images acquired in the wild, from the RAF-DB and FER+ databases. We show that, in these scenarios, a so–called timid semi–supervised learner is more robust and achieves higher performance than standard, confident semi–supervised learners.

1. Introduction

The human face[1] is a powerful mean of communication as it disseminates important cues in inter–human interactions. Due to its many practical applications, automatic face analysis has become a subject of intense investigation, and many recent results [1, 2, 3] showed the benefit of deep learning techniques.

When aiming at automatic analysis of face expression data, the samples (and thus the target problem) can be annotated using either emotion–specific labels (e.g., afraid or happy) or action units. The later were defined by the Facial Action Coding System (FACS) [4]. Action units (AU) are anatomically defined facial actions that individually, or in combinations, can describe nearly all possible facial movements or expressions. Ekman et al. [4] also introduced a set of universal expressions containing 6 classes ("anger", "fear", "disgust", "happy","sad", "surprise") and "neutral"; on occasions this set is extended to include "contempt".

A first observation is that human annotation in the case of facial expressions is hard. Bartlett et al. [5] noted that at least 100 hours of training are needed for a person to achieve minimal competency in action unit recognition. Susskind et al. [6] showed an average accuracy of detection in 6 basic expressions of 89.2% among 23 students in Psychology, which are, at least, familiar with the topic of human expression. One may expect that emotion annotation by common observers will decrease below 70% (the limit to get FACS certification) [7]. In contrast, in an experiment with 8-classes on general images from Caltech-101, Dodge and Karam [8] reported that in a tightly controlled observation sequence, the average user reached 99.3% accuracy.

Secondly, many existing databases are acquired in laboratory conditions with simulated expressions and show facial movements at fixed intensities (usually at apex). Images with expression faces in the wild or with genuine emotions and complete AU annotations are limited. Thus, the problem of face expression analysis is ideal to benefit from semi–supervised learning, as one may easily retrieve face images from the Internet, but without AU or expression labels.

Semi-supervised learning (SSL) methods are motivated by the lack of sufficient resources to create an adequately large labeled dataset, where the true power of deep learning may be unveiled [9]. The main purpose of the SSL algorithms is to improve the performance of supervised learning algorithms by using unlabeled examples.

However, to realistically test the practicality of semi-supervised learning, one needs to: (1) test in practical scenarios and (2) establish accurate baselines (i.e. close to state of the art performance). Without these two preconditions, a SSL system may look good only because it is compared to a weak reference although, in fact, the supervised part is the only one useful. In the first case, many previous works were tested in scenarios where a labeled database is considered and some of the data is considered to be unlabeled to test the SSL method. Such a methodology is disputed [9] because: (a) using the same database guarantees that there is no bias between labeled and unlabeled data; strong bias between different databases may exist [10] and it is very hard to estimate it on unlabeled data. (b) It is guaranteed that the unlabeled part contains the same classes as the labeled part; this, in most practical scenarios, cannot be guaranteed.

Paper contribution and structure. In this paper we contribute by: (1) Accurately evaluating SSL methods in practical scenarios associated with facial expression analysis; (2) Show-

*Corresponding author

[1]This preprint has not undergone peer review or any post-submission improvements or corrections. The Version of Record of this article is published in Pattern Recognition, and is available online at TBA

ing that prominent SSL methods are rather ineffective when challenged by subtle bias between datasets; (3) Showing that improved performance is reachable by exploring the unlabeled subset with ensembles of weak, but diversified teachers in a strategy named *timid SSL*.

In other words, the main contribution of the proposed paper is a solution which is able to improve the performance of the supervised methods by the introduction of unlabeled data from an additional, distinct, database, in the context of face expression analysis. Compared to prior SSL works, we focus on using separate databases for labeled and unlabeled data and, thus, we face different challenges that have a root cause in the bias between databases. To accommodate this potential bias, we emphasize the role of the diversity, implemented by exploring the unlabeled data with ensembles of weak teachers.

The paper is constructed as follows. Previous works on relevant directions are reviewed in Section 2. Theoretical components are described in Section 3. Experimental results on the effect of imbalanced data on face expression recognition and on AU analysis are detailed in Section 4. Discussions and a summary conclude the paper (Section 5).

2. Related Work

Semi-supervised learning (SSL). Semi-supervised learning can be traced back to the pioneering work of Fralick [11]. The main practical task of SSL is to improve the performance with respect to the case when the training uses only the labeled part (i.e. supervised learning), and often is reported to do so [12]. One challenge is to use only the part of the data that is the same in both domains; an example is the work of Pereira et al. [13] which exploits data variance to reduce the domain shift and is able to improve the performance with respect to the purely supervised task only. However, the addition of unlabeled data can be, occasionally, dangerous to the supervised task; Cozman and Cohen noticed [14] that whenever the modeling assumptions adopted for a particular classifier do not match the characteristics of the distribution generating the data, the SSL becomes ineffective.

In the last years, several SSL techniques reported significant success on standard benchmarks such as MNIST, SVHN, CIFAR10/100, ImageNet. Lee [15] showed that a straight–forward solution is to merely use the model trained on the labeled part to annotate the unlabeled part and further propagate in the so-called Pseudo–Label method. Hausser et al. [16] enforced bidirectional associations, in the sense of nearest neighbor, between labeled and unlabeled data as to retrieve better embedding in upper layers of deep networks. Tarvainen and Valpola [17] improved the stability of the model by enforcing a temporal exponential averaging and further regularization parameters. Miyato et al. [18] introduces a regularization based on a measure of local smoothness of the conditional label distribution given the input. Combination of Pseudo-Labeling, temporal ensembling (consensus of prediction of the unknown labels using the outputs of the network-in-training on different epochs) [19] and augmentation by considering convex combinations between pairs of images and respectively their labels (technique known as MixUp) resulted in a powerful semi-supervised solution called MixMatch [20]. Recently, Oliver et al. [9] re-evaluated several such methods and found that Mean Teacher is a serious contender on standard benchmarks. However, they also realized that if the unlabeled data does not contain all labeled classes, the performance of SSL algorithms may decrease when compared to the purely supervised solution.

Face Expression Recognition. Typical systems classify single–person expressions into discrete prototypical classes, namely anger, disgust, fear, happiness, sadness, surprise, and neutral. Among several domain reviews that detailed exhaustively the representative methods and scenarios we refer the reader to the one by Corneanu et al. [21]. Traditionally, solutions based on features and classifiers were chosen. Such work is that of Yan [22] that used 3D HOG and metric learning for nearest neighbor. More recently, transition towards deep learning took place and for instance, Liu et al. [23] blended conditional random forest and deep convolutional networks. Yet, in the last period, dominant solutions are based on deep learning, while the applications of interest are the recognition in images acquired in the wild or the analysis of genuine expressions.

Methods focusing on deep learning [24, 25, 26] train one deep network or an ensemble of deep networks and adapt the prediction on a single independent image or on a sequence. Specifically, expression recognition on static images (as, for instance, it is aimed by mobile phone consumer applications) has been addressed [27]; in this case, carefully engineered CNN-based methods seek to maximize the performance on one database and no other database (as defined in a context of SSL) is used. Multiple databases are envisaged in a set of solutions that augment performance by the inclusion of a modified center loss, as in the case of Li et al. [3]. Alternatively, mechanisms for feature selection may be included [25], for transfer learning [28] or for feature sparseness as is the case of the work by Xie et al. [26].

Other works sought to compensate the scarcity of data in terms of deep learning. In this direction, Liu et al. [29] organize the data instances in terms of hard negatives to augment the training benefits. However, in all experiments, a unique database is used (with a part artificially taken as unlabeled) and the performance is lower than reported by a fully supervised approach.

Action Units Estimation. A particular track in expression recognition is Action Unit (AU) analysis. The simplest direction of investigation is the mere detection of the AUs. Benitez *et al.* [30] reached real-time performance, but by non-deep methods. Also impressive performance has been reached by exploiting AU inter–relationships using a Bayesian Network in both estimation and detection by Wang *et al.* [31]. Making use of deep learning advances, Zhao *et al.* [1] detected AUs in conjunction with the introduction of a larger database.

Although the action units have been defined years ago [4] and their detection is possible in the standard benchmarks (e.g. Cohn-Kanade-CK, CK+, etc.), only recently databases with images in the wild got AU annotations [30]. Also, AU intensity estimation is more recent, as only newer datasets contain intensity annotations.

Most prior works for AU intensity estimation rely on supervised methods. Kaltwang *et al.* [32] generated a latent tree model (LT) by learning the dependencies among features and intensities of multiple AUs. Niu *et al.* [33] used ordinal modelling in the context of deep networks for the age estimation task; further this solution has been adapted for AU intensity estimation. Walecki *et al.* [2] combined conditional random field and copula functions (CCNN-IT) to jointly learn a deep representation and AU relationships. Tran *et al.* [34] proposed semi-parametric variational autoencoders (2DC) for the intensity estimation of multiple AUs.

Previously, supervised methods came under criticism as they may "overfit the training set when intensity annotations are not sufficient, especially for deep models" [35].

A special category of solutions are the weakly supervised learning (WSL) schemes. In such a case, only some of the training data is fully labeled, while the rest is similar and related, but unlabeled. In the AU intensity case, sequences of facial movements have annotations only for the peak and the valley, while in testing all frames must be labeled. Weakly supervised learning is related to semi–supervised learning in the sense that both have labeled and unlabeled data. However, in the case of WSL, the unlabeled data is clearly correlated with the labeled data, as it comes from the same dataset. This is not the case in SSL, which we claim to be more difficult due to potential different distributions of data.

In the case of WSL, Zhao *et al.* [36] combined ordinal and Support Vector Regression to simultaneously make use of both labeled and unlabeled frames. Zhang *et al.* [35] leveraged the ordinal model for weakly supervised learning. Ruiz *et al.* [37] combined the ordinal model with multiple instance regression, while following the same task.

Concluding, while pure SSL has been around for a while and recent advances have been reported, only [38] uses unlabeled data for clustering and a more coordinate supervised learning. Closer to the SSL task are the weakly supervised solutions; still these methods assume that some data from the dataset is unlabeled, thus losing performance compared to a fully supervised framework. There are exceptions [39, 38]; for instance, Zhang *et al.* [39] starts from the labels used in the typical weak supervision framework and introduce the other frames in training, by exploiting the relation between AUs; however, the test is inside a single database, thus it does not suffer from bias. These solutions can be placed rather on the border between "semi-" and "weakly-" supervised.

3. Methodology

The majority of recent semi–supervised learning methods [15, 17, 16, 13, 18, 20], uses the same learner to predict the values on the unlabeled subset. We call these methods "*confident semi-supervised learning*" as they trust that the learner can address simultaneously both categories of data (labeled and unlabeled). However, if there is a certain difference in terms of distribution between the labeled and unlabeled data, a *timid semi-supervised learning* may be more suitable. In the latter case, the unlabeled data is explored and labeled by a diverse ensemble of weak teachers. A key concept is *diversity*, which from a statistical point of view increases the chances to offer a better label to data instances that are outliers with respect to the supervised space.

An assumption often used by the SSL methods is that class borders should go through a low density area in the data space [12]. If the distributions are different, it is likely that the unlabeled data set contains examples that are outliers with respect to the labeled examples, or even form additional, disjoint and sparse clusters. In such a case, a confident learner forces sparse examples to cluster tightly to the nearest labeled data, even though they are from another class. In the case when a predictor, built separately and upon diversity, explores a region of low density, it has a better chance to provide the correct label, that is not connected to the neighboring cluster, as given by a confident learner. The timid SSL, employing different learners to explore the space, does not use the low density assumption.

A schematic difference between the two categories, *confident vs timid*, may be followed in Figure 1.

3.1. Semi–supervised learning

In the semi-supervised learning (SSL) framework, the learner uses labeled training data $\{\mathcal{X}^l, \mathcal{Y}\} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N} \overset{\text{iid}}{\approx} p(\mathcal{X}, \mathcal{Y})$ and unlabeled training data $\{\mathcal{X}\} = \{\mathbf{x}_j\}_{j=N+1}^{N+M} \overset{\text{iid}}{\approx} p(\mathcal{X})$, and learns a predictor $f : X \to Y$, $f \in \mathcal{F}$ where $\mathcal{F}$ is the hypothesis space.

In our case, $x \in \mathcal{X}$ is an input face image showing an expression, $y \in \mathcal{Y}$ is its target label (categorical for the expression case, respectively vector of values for the AU case), $p(\mathcal{X}, \mathcal{Y})$ the unknown joint distribution and $p(\mathcal{X})$ its data marginal. The goal is to construct a predictor that assesses the future test data better than the predictor learned trained only on the labeled data set alone. In practice, it is much harder to obtain independent and identical distribution (iid) of samples inside the labeled training set and respectively across datasets (i.e. with respect to the unlabeled part) .

For the semi–supervised learning to be successful, one needs to use the unlabeled data to structure the learner f; this is often implemented as a regularization. Assuming that the predictor is defined by a set of parameters θ, we can express the problem under a Bayesian formulation:

$$f_\theta(x) = \underset{\theta}{\operatorname{argmax}}\, p(\dagger|\S, \theta) = \underset{\theta}{\operatorname{argmax}}\, \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{\sum_{y'} p(\mathbf{x}, \mathbf{y}'|\theta)} \quad (1)$$

The structuring on the unlabeled data may be written as a regularization term L_R:

$$L_R = \log p(\mathcal{X}|\theta) = \sum_{j=N+1}^{N+M} log \left(\sum_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{x}_j, \mathbf{y}|\theta) \right) \quad (2)$$

Various solutions have been proposed for L_R. Some of the most recent are listed in Table 1. One might note that some solutions [15, 17, 16] use the current learner to explore

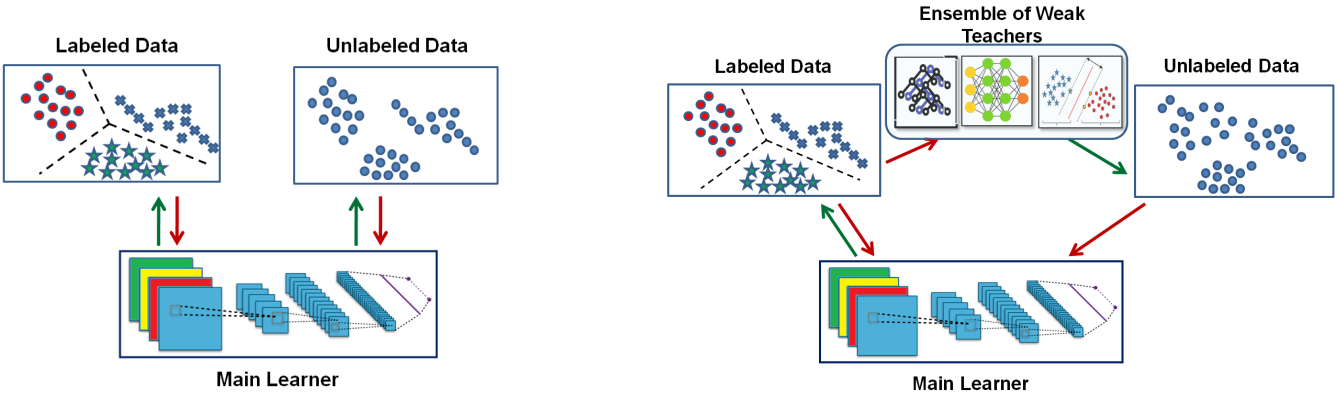


Figure 1: The schematic of the use semi-supervised confident vs timid learning for exploring the unlabeled data. Red arrows indicate information used to structure/train the learner, while green arrows indicate prediction over data. Confident learners explore unlabeled data using the current state of the main learner and, often, deploy *the low density assumption* to improve supervised training. The timid approach uses a separate and weak teacher ensemble (previously trained on annotated data) to label examples without annotations. The procedure for training the main learner is explained in section 3.4.

and label the unlabeled data $\mathbf{x}_j$, $j \in \{N+1, \ldots, N+M\}$: Pseudo–Label [15] uses the learner itself and asks for confidence, Mean Teacher [17] uses an exponential averaged version of the learner for the same task, while the Associative SSL [16] forces the learner to produce an embedding relevant for labeling with nearest neighbor. MixMatch [20] combines Pseudo-Label with Mean Teacher and other technique, thus representing a potential apex of these approaches. Overall, these methods, in which there is trust that the chosen solution is capable of exploring simultaneously both data domains, will be called in this work *confident* SSL. As it will be discussed later, these methods are reliable if the labeled and unlabeled data are indeed identically distributed.

Contrary to the mentioned solutions [15, 17, 16, 20], which use the same model to explore unlabeled data, we claim that in real tasks (where the two sets might suffer from bias) it is more efficient to learn from *different* models as, intuitively, the learner needs different paths to reach "a better optimum". Our proposal is to use *timid* learners.

In the case of timid learner, the regularization term, introduced in eq. (2), is build upon the decision of weak teachers ensembles about the unlabeled examples :

$$L_R = \frac{1}{M} \sum_{j=N+1}^{N+M} \sum_{m=1}^{C} L(V(\mathbf{x}_j); f_{\theta t}(\mathbf{x}_j)) \tag{3}$$

where $V(\mathbf{x}_j)$ is the label provided by the weak teacher and will be detailed in the next subsection, C is the number of classes, t is the iteration of the training stage, $f_{\theta t}$ is the main learner at the t-th iteration and $L(\cdot;\cdot)$ is the standard cross entropy.

3.2. Weak Teachers Ensemble

An alternative method to produce pseudo labels can be traced back to the work of Caruana et al. [40]. There, an ensemble model of classifiers is trained on the labeled data and used to pseudo-annotate the unlabeled data. However, in that work the emphasis is to construct an ensemble that is much more powerful than the basic learner; such an ensemble would be of confident learners. In a slight different direction, Krogh and Vedelsby [41] showed that given α learners, $V_\alpha(\mathbf{x})$, the ensemble can be formed by:

$$\overline{V(\mathbf{x})} = \sum_{\alpha} w_\alpha V_\alpha(\mathbf{x}). \tag{4}$$

and the ensemble generalization error, E, is:

$$\begin{aligned} E &= \overline{E} - \overline{A} = \textstyle\sum_\alpha w_\alpha E_\alpha - \sum_\alpha w_\alpha A_\alpha \\ &= \textstyle\sum_\alpha w_\alpha \left(\sum_{i=N+1}^{N+M} p(\mathbf{x_i}) \left(e^\alpha(\mathbf{x_i}) - a^\alpha(\mathbf{x_i})\right)\right) \end{aligned} \tag{5}$$

where $\overline{E}$ and $\overline{A}$ encode the ensemble learners errors and the overall ambiguities, while $e^\alpha(\mathbf{x_i})$ is the error of the α learner in predicting $\mathbf{x_i}$, and respectively $a^\alpha(\mathbf{x_i})$ is the local ambiguity. The ambiguity is computed as the variance of the learner with respect to the mean. In general, the probability of a data instance $p(\mathbf{x_i})$ cannot be accurately determined. Also, given the fact that unlabeled data misses the annotations, the accuracy over it cannot be estimated either.

We emphasize that the ambiguity (given as a positive measure) is subtracted from the overall error and larger amounts are needed for improved performance.

In other words, eq. (5) shows that the total error is the difference between individual errors and the ensemble diversity. Thus, one may achieve greater performance either by (a) increasing the learner performance or (b) by increasing the diversity[2] of the methods. The first way of increasing the performance, given a truly unlabeled dataset, is uncertain how to be accomplished as no reference measure is available.

[2] In this work, *diversity* is used in a pattern recognition/machine learning sense and it does not refer to anatomical/psychological categories of the persons represented in the image data.

Table 1: Semi-supervised solutions and their corresponding regularization term.

Method	**Regularization**, L_R	**Notes**
Proposed	$\frac{1}{M}\sum_{j=N+1}^{N+M}\sum_{m=1}^{C} L(V(\mathbf{x}_j); f_{\theta t}(\mathbf{x}_j));$	$V(\mathbf{x}_j)$ – label provided by outer learner t - iteration; C- no. of classes; $f_{\theta t}$ learner at t-th iteration $L(\cdot)$ - cross entropy
Pseudo–Label [15]	$\alpha(t)\frac{1}{M}\sum_{j=N+1}^{N+M}\sum_{m=1}^{C} L(y_m^j; f_{\theta t}^m(\mathbf{x}_j));$	$y_j = \begin{cases} 1 & j = \text{argmax}_j\, f_\theta(\mathbf{x}_j) \\ 0 & otherwise \end{cases}$ α-balancing term
Mean Teacher [17]	$\alpha(t)\frac{1}{M}\sum_{j=N+1}^{N+M} d(f_{\theta t}(x_j), f_{\overline{\theta}_t}(x_j));$	$d(a,b) = \|a^2 - b^2\|$ t - iteration; α-balancing term $\overline{\theta_t} = \beta\overline{\theta_{t-1}} + (1-\beta)\theta_t$
Association [16]	$\frac{1}{N}\sum_{i=1}^{N} log\sum_{j=N+1}^{N+M}\frac{e^{d^C(ij)}}{\sum_{j'=N+1}^{N+M} e^{d^C(ij')}}$ + $\frac{1}{M}\sum_{j=N+1}^{M} log\sum_{1=1}^{N}\frac{e^{d^c(ij)}}{\sum_{j'=N+1}^{N+M} e^{d^c(ij')}}$	$d^c(i,j) =< f_\theta(\mathbf{x}_i), f_\theta(\mathbf{x}_j) >$ d^C – similarity measure ; $<,>$ – scalar product

On unlabeled data it is not possible to compute accuracy (classification/regression errors). Other solutions simply assume that given the power of a specific solution, in general, it will also prevail on the current problem in the sense that it will provide a smaller error. This idea, while having merits in many practical solutions, is not absolutely true, due to the "no free lunch" theorem: for each solution, no matter how strong, there will always be a topology where it will fail. In these circumstances, we aim to increase the diversity in order to compensate for the impossibility of computing the error. By increasing the diversity, we aim to more safely explore spaces with a different distribution. The diversity was implemented by using various combinations of classifiers and hand-crafted features.

Another strong point is that ambiguity needs only unlabeled data to be evaluated [41] and it does not impose any restriction on the labeled part of the data.

Optimization in other solutions [15, 17, 16, 20] does not consider diversity or variance of the classifier. It focuses on minimizing the prediction error and follows this goal by transduction, which is assuming that a strong learner will always have a small error.

In our proposal, the ensemble diversity is achieved by training on the supervised part of the data various systems based on diverse features, diverse learners and on various bootstrapped subsets of the data. The feature descriptors are Histogram of Oriented Gradients - HOG and Local Binary Pattern - LPB. The learners are Support Vector Machine - SVM, Random Forest - RF, Gradient Boosted Machine - GBM and Multi-Layer Perceptron - MLP. A summary of the weak learners is presented in Table 2. All learners have the same importance ($w_\alpha = 1$) in eq. (4).

The ensemble of weak learners will act as a weak teachers ensemble (WTE) on the unlabeled data for the main learner, forming the so-called *timid SSL*. While one may choose the individual learners to be deep architectures due to their prowess, they require more resources to train and are not so easy to be diversified in the sense defined by Eq. 5.

Table 2: Weak Learners from the ensemble used to explore the unlabeled space. Different experts are obtained using the same features and learners by bootstrapping three times at 60% the training set.

Feature	**Learners**	**No. of experts**
LPB	SVM ; RF ; GBM ; MLP	4×3
HOG	SVM ; RF ; GBM ; MLP	4×3
LPB+HOG	SVM ; RF ; GBM	3×3

3.3. Labeling the Unlabeled Part

Another perspective on semi-supervised learning is that often, in the learning process, one produces labels for the unlabeled data and uses them to improve the accuracy of the learner. The learner itself [15, 17, 20], a nearest neighbor based on the learner embedding [16] or an outer system (in our case) can be used to produce labels. The labeling quality varies between perfect, where it will matter if the labeled and unlabeled sets are indeed identical distributed, yet occupying different parts of the space, and respectively, totally noisy; using them, it will proliferate the noise, thus hurting the learner.

Given the nature of the face expression problem, a likely view is that it is improbable to obtain two databases that have been acquired differently to be without any bias. Thus, the databases are prone to be non-identically distributed. We emphasize again, that for high dimensional data $\mathbf{x}_i$, such as images, it very hard to accurately estimate data distribution alone $p(\mathcal{X})$, or with labels $p(\mathcal{X}, \mathcal{Y})$ and to be able to quantify the potential bias.

3.4. Training and testing procedure

The actual procedure is as follows:

Training. **Input:** a pair of datasets addressing the same aspect of the face expression problem. One dataset has labels, one has not.

Procedure:

1. Train a set of diverse weak teachers on the annotated dataset. The set is detailed in tables 2 (type) and 5 (baseline performance for a specific database). Diversity is ensured by the use of various features, classifiers and, respectively, by bootstrapping of the training set.
2. Use the weak teachers ensemble to annotate the unlabeled data. In the case of a classification problem, the decision is based on plurality, while for a regression problem, on the mean of the experts prediction. We compute the regularization term listed in table 1. At this moment, both datasets have labels.
3. Use both datasets to train the main learner (deep convolutional neural network - CNN) using a standard optimization algorithm (in our case Stochastic Gradient Descent). Iterations alternate between the two datasets; thus it is activated either main loss or the semi–supervised regularization L_R.

Output: Trained CNN, denoted by $f_\theta()$ in eq. (3).

Testing. **Input:** an image with a face.

Procedure: Use the trained CNN to predict the label of the given image. The label is scalar for the expression recognition case and multivariate in the case of action units.

4. Evaluation

4.1. Databases and Scenarios

We tested in two directions: expression recognition in static images in the wild and action unit estimation. In each case, we considered two scenarios associated with a database and a task.

Face expressions in the wild. For this scenario, we tested on the FER+[27] and Real-world Affective Face Database (RAF-DB) [3] database. Both databases contain images retrieved after relevant searches on the Internet, followed by human filtering and annotation.

FER+ is derived from FER2013 and it contains 28709 training images, 3589 validation (public test) and another 3589 (private) test images, in the wild. FER images have 48×48 pixels, are gray-scale and contain only the face. Barsoum et al. [27] noted the high noise in the original FER 2013 labels and performed some "cleaning", by removing the images with missing faces and providing user set labels. The labels have been obtained by aggregating the opinion of 10 non-specialist annotators. Example images are in Figure 2, (a,b). While FER+ is more reliable with respect to the quality of labels, we still report results on FER2013.

RAF-DB [3] contains facial images in the wild. Original images are color and large enough such that the cropped face often requires downsizing to 224×224 pixels. The database is annotated by at least 40 trained annotators per image and divided into 12271 training images and 3078 testing images. It is labeled for seven basic emotions. Example images are in Figure 2, (e,f).

The unlabeled data for both experiments is the first subset of the MegaFace database [42], which contains approximately 311.000 images with faces. The images have been randomly selected from the Internet. The faces were cropped from the images based on the bounding box provided by the widely used MTCNN face detector. Each image contains a face acquired in an unrestricted background (i.e "in the wild") that has an expression. However, there is no information about the expression; one may assume that plurality is with neutral, contempt (looking at the camera) and happy (smiling) face, but nothing more. Example images are in Figure 2 (c,d).

Action Units. We also performed two tests: AU binary detection on images in the wild and AU intensity estimation on images in laboratory conditions.

The detection scenario is run entirely on the EmotioNet dataset [30]. It contains 1M images collected from the Internet, being unconstrained. Out of these, 50,000 images were manually annotated with binary labels with multiple AUs. Overall, 7 AUs appear in more than 5% of the images. The original paper [30] divided the annotated set into 25,000 images train/test partitions and used F1 score as main metric. The training partition is used as labeled for the SSL, while as unlabeled data we selected 500k from the remainder 950k images.

For the Action Unit intensity estimation experiment, we rely on the DISFA [43] database as source for the labeled training data and testing. Images from the Extended Cohn-Kanade – CK+ [44] are taken as unlabeled data. The DISFA dataset contains video recordings of 27 subjects spontaneously reacting to YouTube videos, totaling more than 200K frames. Our experiments were performed with a subject independent setting (dividing data into training and testing partitions) in the same manner as prior art did [36, 34, 2, 35, 37]. The database has AU's intensity annotated by experts.

The CK+ database consists of 593 sequences of posed facial actions from 123 subjects.

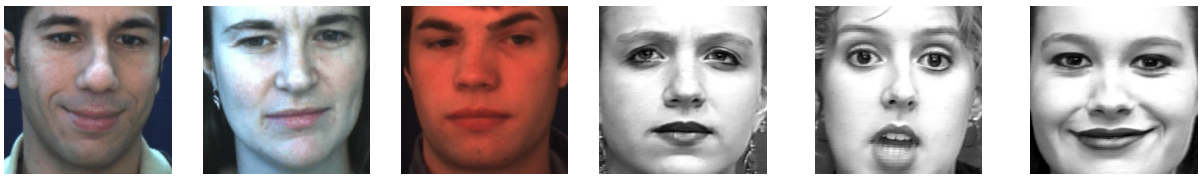
Although CK+ has weak labels, the test is relevant because the problem of AU estimation is very domain specific and some differences between DISFA and CK+ exist: DISFA is spontaneous (thus has the trait aimed by a practical application), while CK+ is posed. Using posed expression is the normal way to produce new data with AU annotations. Also DISFA is newer and contains RGB information, CK+ contains many videos older than 15 years that have a weak image quality and provides only gray-scale information. An illustration of example images is in Figure 3. Overall, the scenario is indeed practical: having to analyze genuine expressions, one seeks to extend the dataset with posed expressions, in a laboratory setup; indirectly there will be some definite bias encoded in the nature of the images.

4.2. Implementation

The deep learning architectures used (e.g. AlexNet, VGG-16, ResNet, DenseNet) are standard, with batch normalization and L1 sparsity regularization over weights. For the timid SSL, we added the term obtained from the weak teachers and used it

(a) (b) (c) (d) (e) (f)

Figure 2: Face crop images from FER+ database (a,b) from MegaFace (c,d) and, respectively, from RAF-DB (e,f).



(a) AU12:3 (b) AU4:2, AU6:1, AU9:2 (c) AU1:1, AU2:1 (d) AU4:1 (e) AU1:3, AU2:3, AU5:3, AU25:3 (f) AU12:3

Figure 3: Face crop images from DISFA database (a-c) and respectively from Cohn-Kanade+ (d-f). We note the marked AU and reported intensities. On the CK+ database, the intensities are extrapolated from the final frame, assumed to be the most intense.

for regularization. The training was done employing the stochastic gradient descent (SGD) algorithm with a learning rate adapted to the architecture and to the problem for 150 epochs. For the expression recognition scenarios, the networks were trained with cross-entropy, while for AU, the mean squared error was used. For the small architecture (i.e. AlexNet) the learning rate starts at 10^{-3} and decreases with a ratio of 10 every 50 epochs; for larger ones, it starts at 10^{-4}. Both PyTorch and TensorFlow were used; several scenarios were run on both libraries and results were compatible.

For the confident SSL solutions, various combinations of hyper–parameters have been tried and only the best performance is reported. In cases when we used the public code, we sought good value for learning rates and other hyper–parameters, while leaving intact (if not specified otherwise) the weight of the contribution of the unlabeled data.

4.3. Face Expression Recognition

In the context of face expression recognition two metrics have been previously used to establish the performance of a method:

- Accuracy (marked as Acc. in Table 4), also named total average, is the percentage of correctly predicted labels:

$$Avg = \frac{1}{N_{img}} \sum_{j=1}^{N_{img}} (y^p(j) == y(j)) \tag{6}$$

  where $y(j)$ is the label of the j-th image, while $y^p(j)$ is its prediction. This metric is more widely used in face expression recognition experiments.

- Average accuracy, marked as "Avg. Acc." in Table 4, is the average of per–class–accuracy or the average of the diagonal values of the confusion matrix:

$$Avg.Acc = \frac{1}{7} \sum_{i=1}^{N_{Em}=7} Avg(i) \tag{7}$$

$$Avg(i) = \frac{\sum_{j=1}^{N_{img}} (y^p(j) == y(j)) \wedge (y(j) == i))}{\sum_{j=1}^{N_{img}} ((y(j) == i))} \tag{8}$$

  This metric is widely used in experiments associated with the RAF-DB database [3].

Results achieved on the expression recognition experiment are presented in Table 3, while experimenting on small gray images (FER+) and in Table 4 while experimenting on normal color images (RAF-DB).

First, to establish a baseline, we report the performance of several architectures trained solely on the labeled data (supervised). We also cite previously published, carefully engineered prior art methods [3, 28, 24, 25] and respectively [27, 25].

As said, all databases (RAF-DB, FER+/FER2013 and MegaFace) contain images randomly acquired from the Internet, thus without obvious bias between the datasets. However, most of the confident semi-supervised learners failed to produce any improvement, even hurting the performance. The Pseudo–Label solution was able to improve only when we set α to be constant and stopped using the unlabeled part at half of the training stage. These results, although slightly disappointing, are consistent with previous reports [9]. There, while testing on CIFAR10 and ImageNet, it has been found that if the unlabeled database has a different distribution than the labeled one, then the transfer of information hurts. This suggests some conclusions. First, there are uneven distributions (although not evident) of the two parts of the dataset. We have used our best

Table 3: Recognition rates (accuracy) within the 8-class problem on the FER+ and the 7-class on the FER2013 database. Results with reference, but marked by (*) have been obtained by us using author code and the ones marked by (**) by re–implementing. The Pseudo–Label solution was able to improve only when we set α to be constant and stopped using the unlabeled part at half of the training stage

Method			FER+	FER2013
SUPERVISED		AlexNet [25]	n/a	61.1
		AlexNet	78.08	68.2
		Densenet 121	79.12	66.66
		DenseNet 201	80.05	69.70
		FSN – AlexNet [25]	n/a	67.6
		VGG – Sparse [26]	n/a	70.08
		ResNet – Sparse [26]	n/a	71.90
		VGG – Majority vote [27]	83.85	–
		VGG – Probabilistic label [27]	84.99	–
SSL	CONFIDENT	AlexNet - Pseudo–Label [15](**)	80.82	69.62
		DenseNet 121 - Pseudo–Label [15](**)	82.25	70.15
		AlexNet - Mean Teacher [17] (*)	46.87	44.41
		DenseNet 121 - Mean Teacher [17] (*)	47.35	44.38
		AlexNet - Association [16] (*)	73.15	62.21
		DenseNet 201 - MixMatch [20] (*)	82.18	–
	TIMID	AlexNet+ WTE - Majority vote	83.52	71.3
		DenseNet 121 - WTE - Majority vote	83.85	70.66
		DenseNet 201 - WTE - Majority vote	84.87	71.45
		ResNet 50 - WTE - Majority vote	***85.45***	***72.54***

performer to predict the expressions on all images from the datasets used in the FER+ experiment; the resulting histograms are in Figure 5 and are obviously different. Second, the confident semi-supervised learners are less suitable for practical applications, in which the distribution cannot be controlled.

Results from Tables 3, 4 clearly show that the proposed solution (marked by Timid SSL), which uses Weak Teachers Ensemble, provides a significant $(4-5\%)$ and universal performance increase, largely surpassing the performance increase of any other SSL.

Only the weakening of Pseudo–Label and respectively the ensemble of weak teachers (based on diversity) was able to improve the performance of standard architectures until the point where they become competitive with the skillful engineered method reported in prior works [3, 28, 24, 25, 27]. In recent works [9, 20] it has been found that Mean Teacher is a better performer than Pseudo–Label, while our experiments showed the opposite. However the experiments differ by (1) small resolution images with easily distinguishable objects there [9, 20] as images from CIFRA and SVHN were used in contrast to the large resolution images with subtle differences in pixels from face, here; (2) reduced or no bias between labeled and unlabeled database there (as both are extracted from a unique database) versus potential bias, here, as two databases have been employed. In our view, the main reason for the performance in version is the bias in distributions between supervised and unsupervised datasets, fact discussed in section 4.6. Mean Teacher is more keen on preserving the direction in which the parameters are adjusted. Given the bias, the Mean Teacher will follow less optimal directions; in contrast Pseudo–Label is more adaptable to the data. Furthermore, the recent and strong solutions of MixMatch [20] encountered convergence problems; it often lead to predicting a unique value equal to the best represented class. Imposing strong assumptions about database distribution, such as being precisely the same, in the case of labeled and unlabeled set may hurt performance. We are delving deeper in the analysis of database bias in subsection 4.6. On RAF-DB, our solution managed to outperform previously reported results.

To provide a better insight into the functionality of the ensemble of weak learners, we report their individual performance on FER2013/FER+ database in Table 5. As one can see, their performance is much lower than the overall system and lower than any solution based on deep learning. We consider that the superior performance is owing to diversity, as enhanced by eq. (5) and discussed in section 3.3.

Confusion matrices for the best solutions proposed by us may be followed in Table 6 for the FER+ database and, respectively, in Table 7 for RAF-DB. Visual examples with both positive and negative results may be followed in Figure 4.

4.4. Action Unit Detection and Intensity Estimation

We have treated the problem of action unit detection/estimation as a multiple instance regression, training a single architecture to report the intensity on the annotated AUs. In this case, the set

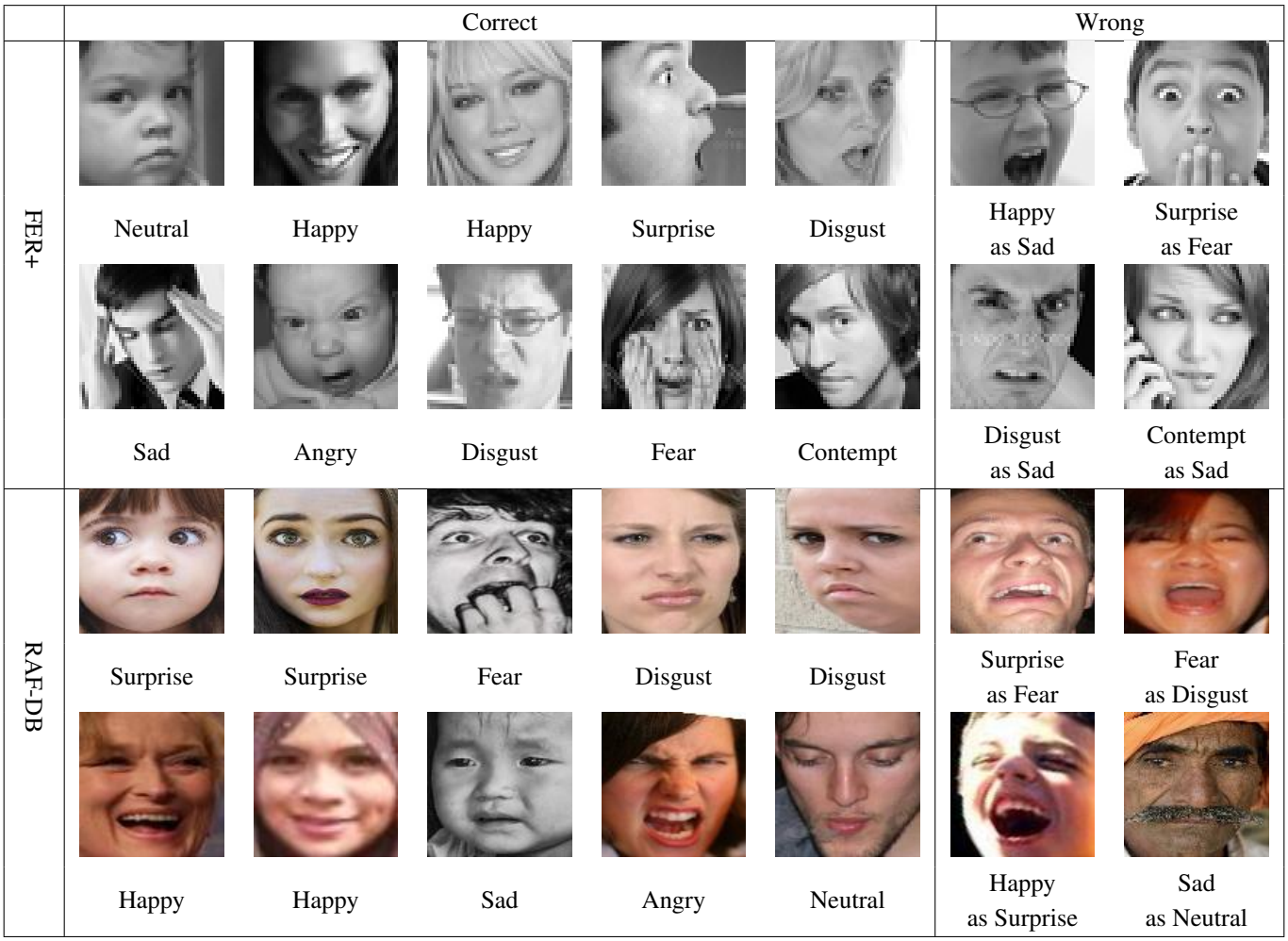
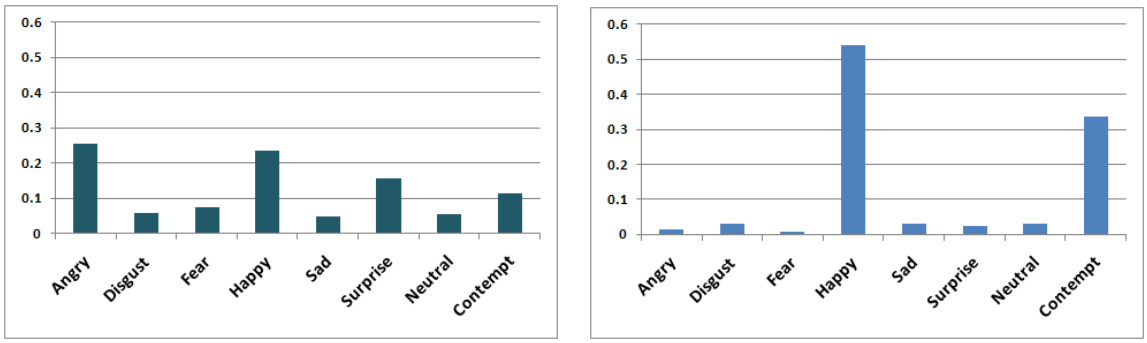


Figure 4: Examples of images from the FER+ database (first two rows) and respectively from RAF-DB database.

Table 4: Performance within the 7-class problem on the RAF-DB database. FSN - feature selection network, FSM frame-to-sequence method and WTE (Weak Teachers Ensemble) marks our timid learner. With bold we marked the best result.

Method / Metric			Avg. Acc.	Acc.
SUPERVISED		AlexNet - [3]	55.60	68.90
		VGG-16 [3]	58.22	70.53
		DenseNet 201	71.22	81.50
		DLP-CNN [3]	74.20	84.13
		ResNet-18 [28]	–	80.00
		FSM [24]	65.52	72.21
		FSN [25]	72.46	81.10
SSL	CONFIDENT	AlexNet - Pseudo–Label [15](**)	60.4	73.21
		VGG-16 - Pseudo–Label [15](**)	64.20	77.12
		AlexNet - Mean Teacher [17] (*)	54.45	60.10
		VGG-16 - Mean Teacher [17] (*)	57.67	68.56
		DenseNet 201 - Mean Teacher [17] (*)	55.35	68.10
		DenseNet 201 - MixMatch [20] (*)	59.12	73.21
	TIMID	AlexNet - WTE	66.6	78.18
		VGG-16 - WTE	**78.64**	**85.41**
		DenseNet 201 - WTE	75.98	83.15



FER+ **MegaFace**

Figure 5: Histogram of the expressions in the FER+ database and respectively MegaFace, as it has been estimated by the best performer (timid SSL over VGG-16).

of investigated architectures is limited to AlexNet and VGG–16. These architectures are able to obtain more than competitive performance.

In this case, the objective evaluation is carried based on the following correlation coefficients:

1. Pearson Correlation Coefficient:

$$PCC = \frac{A}{B}$$
$$A = N_{img}\sum_{j=1}^{N_{img}} y^p(j)y(j) - \sum_{j=1}^{N_{img}} y^p(j)\sum_{j=1}^{N_{img}y(j)}$$
$$B = \sqrt{N_{img}\left(\sum_{j=1}^{N_{img}}(y^p(j))^2\right) - \left(\sum_{j=1}^{N_{img}} y(j)\right)^2}$$
$$\cdot\sqrt{N_{img}\left(\sum_{j=1}^{N_{img}}(y(j))^2\right) - \left(\sum_{j=1}^{N_{img}} y^p(j)\right)^2} \quad (9)$$

where $y(j)$ is the label of the j-th image, while $y^p(j)$ is its prediction.

2. Intra-class Correlation Coefficient (ICC):

$$\begin{aligned} ICC &= \frac{1}{(N_{img}-1)\sigma^2}\sum_{j=1}^{N_{img}}(y^p(j)-\overline{y})(y(j)-\overline{y}); \\ \overline{y} &= \frac{1}{2N_{img}}\sum_{j=1}^{N_{img}}(y^p(j)+(y(j)); \\ \sigma^2 &= \frac{1}{2N_{img}-1}\left(\sum_{j=1}^{N_{img}}(y^p(j)+(y(j))^2 + \right. \\ &\quad \left.\sum_{j=1}^{N_{img}}(y(j)+(y(j))^2\right. \end{aligned} \quad (10)$$

Since the confident semi-supervised methods have been designed for classification, adjustments were needed to make them learn multiple simultaneous variables. In the Association solution [16], the distribution of classes was completely canceled, while for Pseudo–Label we enforced to report discrete intensity and train using the difference.

Detection on images in the wild from the EmotioNet database

Table 5: Detailed performance (accuracy) of the weak experts on the FER+/FER2013 database. Models using the same classifier and features used 60% randomly selected training data instances (bootstrapping).

No.	Feature	Classifier	Acc.-FER2013	Acc.-FER+
1	HOG_1	SVM	48.90	60.10
2	HOG_2	SVM	51.30	61.30
3	HOG_3	SVM	50.60	62.70
4	LBP_1	SVM	33.70	43.50
5	LBP_2	SVM	32.10	44
6	LBP_3	SVM	33.40	42.40
7	$HOG+LBP_1$	SVM	49.10	62.60
8	$HOG+LBP_2$	SVM	47.50	61.20
9	$HOG+LBP_3$	SVM	48.30	63.20
10	HOG_1	RF	43.60	55.70
11	HOG_2	RF	44.50	56.00
12	HOG_3	RF	43.80	56.30
13	LBP_1	RF	35.00	47.20
14	LBP_2	RF	35.60	46.50
15	LBP_3	RF	34.80	47.00
16	$HOG+LBP_1$	RF	39.30	54.30
17	$HOG+LBP_2$	RF	41.50	53.20
18	$HOG+LBP_3$	RF	41.30	53.20
19	HOG_1	GBM	40.40	46.00
20	HOG_2	GBM	41.50	46.50
21	HOG_3	GBM	40.80	47.20
22	LBP_1	GBM	32.70	36.80
23	LBP_2	GBM	32.60	37.50
24	LBP_3	GBM	33.30	37.70
25	$HOG+LBP_1$	GBM	39.00	45.00
26	$HOG+LBP_2$	GBM	38.70	45.80
27	$HOG+LBP_3$	GBM	38.20	44.40
28	HOG_1	MLP	43.20	54.30
29	HOG_2	MLP	43.80	56.10
30	HOG_3	MLP	42.60	55.70
31	LBP_1	MLP	32.10	45.70
32	LBP_2	MLP	33.70	44.50
33	LBP_3	MLP	32.80	45.00

was achieved by thresholding the predicted intensity; thus, the predicted labels became binary.

The results on the EmotioNet are shown in Table 8. The metric and the procedure follow the works introducing the database [30, 38]. There, the standard measure used for evaluation is the F1 score, which is defined for the action unit AU_i as follows:

$$F1_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{11}$$

Here, $Precision_i$ is the fraction of the automatic annotations of AU_i that are correctly recognized (i.e., number of correct recognitions of AU_i divided by the number of images with detected AU_i), and $Recall_i$ is the number of correct recognitions of AU_i over the actual number of images with AU_i.

In this situation too, the confident SSLs fail to improve the performance, although in this case the dataset was collected in a unitary manner. Furthermore, in this case, testing against prior art is broader, as [38] also proposed a SSL method which uses clustering over a base learner defined by several prior methods: DRML - [1], TSVM - Transductive SVM [45], GFK - [46].

Furthermore, the solution introducing the EmotioNet database [30] can be seen as a strong learner and we report an AlexNet trained with the unsupervised part labeled by this method. These results, marked by (AlexNet with [30]) allow comparison when the focus is on performance, instead of diversity, as defined by eq. (5). Overall, the top performance is for the proposed timid SSL, proving that for unlabeled datasets, diversity explores better than sheer strength.

The results for AU intensity estimation on the DISFA set using Intraclasss Correlation Coefficient (ICC) are presented in Table 9. In this case, the mere use of the WTE predicted labels on CK+ images proved less efficient due to the bias. To gain an improvement and to stop spreading bad labels, we used the WTE annotated CK+ images only in the first 10 epochs (forming an Early Weak teachers - EWT). In such a case, using the unlabeled part, we were able to improve the performance with 3% w.r.t the baseline.

Regarding the performance of confident semi-supervised learning methods, we see that again, they degrade the performance. A likely cause is related to the difference between images: posed vs genuine expression, different distribution of intensities, different quality of images. Another observation is that all these semi-supervised methods being confident, they did use the unlabeled set in all the epochs of training.

In Table 10 we detail the results based on the Pearson Correlation Coefficient (PCC) metric and reference prior works that reported in the same framework. One [32] is supervised, while the other two [36, 35] are weakly supervised as they focus only on the neutral phase and apex. Using more labeled data (which is available) in a simple architecture is sufficient to significantly outperform them. As it was the case when evaluating based on ICC (listed in eq. (10)) , a VGG-16 trained with our method achieves top performance.

This scenario is the only one from the four ones presented in the this paper in which the unlabeled part from the semi–supervised framework has labels. We have experimented with

Table 6: Confusion matrix for the 8 expressions problem on the FER+ database obtained with timid SSL over VGG-16 .

	Ang.	**Disg.**	**Fear**	**Hap.**	**Sad**	**Sur.**	**Neut.**	**Cont.**
Ang.	0.79	0.01	0.02	0.05	0.04	0.03	0.06	0.01
Disg.	0	0.71	0	0	0.12	0	0.12	0.06
Fear	0.04	0	0.64	0.07	0.04	0.11	0.11	0
Hap.	0.02	0	0	0.87	0.03	0.01	0.05	0
Sad	0.03	0	0.03	0.04	0.71	0.01	0.16	0.01
Sur.	0.03	0	0.05	0.04	0.01	0.82	0.05	0
Neut.	0.02	0	0	0.03	0.07	0.02	0.85	0.01
Cont.	0	0	0	0	0.11	0	0.22	0.67

Table 7: Confusion matrix for the 7 expression problem on the RAF-DB database timid SSL over ResNet-50

Expres	**Sur.**	**Fear.**	**Disg.**	**Hap.**	**Sad.**	**Ang.**	**Neut.**
Sur	0.835	0.049	0.020	0.026	0.006	0.012	0.052
Fear.	0.075	0.755	0	0	0.113	0.057	0
Disg.	0.030	0.010	0.626	0.061	0.091	0.04	0.141
Hap.	0.004	0.002	0.015	0.934	0.019	0.009	0.016
Sad.	0.009	0.018	0.037	0.018	0.826	0.011	0.081
Ang.	0.024	0.018	0.109	0.042	0.042	0.758	0.006
Neut.	0.027	0.004	0.050	0.061	0.073	0.013	0.771

a network in which instead of a learner that explores the unlabeled part, the ground truth labels are used. This would be the ideal case and compared to other SSL solutions achieved very good performance. Such a solution incorporates the bias between the dataset and represents an upper bound of SSL performance for the same architecture.

Due to its popularity, DISFA allows a detailed comparison with prior art. The performance of a carefully trained AlexNet, helped by the WTE applied on the unlabeled data is significantly better than any of the weakly supervised method, by a margin of 8%. Better performance is reported by larger and skilfully engineered deep learning methods trained in a fully supervised manner [34]. Using an architecture from the same category (VGG-16) and augmenting with our solution top performance was reached.

4.5. Framework ablation

The main scenario investigated assumes to use an annotated database entirely as source of labeled data and a different database as source of unlabeled images. In this subsection we investigate the performance of the proposed solution when the number available labeled images is gradually smaller. This would correspond to a scenario less investigated before, where the availability of annotations is severely limited. The achieved performance on the two expression databases, namely FER+ and RAF-DB may be followed in tables 11 and 12.

When only a low number of labels available it hurts the performance of the weak teacher ensemble and, thus, it reduces the precision of the proposed solution. In these cases other solutions report better performance. When the amount of labels becomes sufficient, out solution regains its superiority.

4.6. Performance and training dataset bias

One claim of the proposed method is that the timid SSL performs better on datasets with different distributions than "confident" SSL which is bound to have similar distribution between the labeled data and unlabeled data. To quantify this aspect we have devised the following experiment.

In a formal manner, in this experiment we have partially evaluated the behavior of the proposed solution when the differences appear between the labeled and unlabeled set. The distribution of a set is $p(\mathcal{X}, \mathcal{Y})$ and $p(\mathcal{X}), p(\mathcal{Y})$ are only the marginals. It is generally accepted that is too hard to evaluate $p(\mathcal{X})$ accurately. Hence, we have imposed that $p(\mathbf{y}_j), j = 1 \dots N$ and respectively $p(\mathbf{y}_j), j = N+1 \dots N+M$ to be different. The differences in the label marginal forced differences in $p(\mathcal{X}, \mathcal{Y})$ too. The difference is quantified by the Kullback Leibler divergence.

For the actual evaluation, we have considered images from RAF-DB. In the original database, 3068 images were chosen to be in the test set. We have chosen as training set for all following experiments a set of 3068 images from the original RAF-DB set having precisely the same distribution of labels as the test set. In this case, the architecture was AlexNet and the baseline performance obtained by purely supervised training is 71.8%.

Table 8: F1 score while detecting Action Units on the EmotioNet database. DRML stands for Deep region and multi-label, GFK for Geodesic flow kernel, TSVM for Transductive SVM.

	Method/ AU	1	4	5	6	12	25	26	Avg
SUPERVIS.	AlexNet [38]	.24	.35	.40	.73	.87	.89	.46	.561
	AlexNet- ours	.32	.57	.29	.71	.77	.84	.50	.572
	VGG-16 - ours	.45	.64	.42	.73	.79	.85	.53	.627
	DRML [38]	.25	.36	.40	.75	.87	.89	.46	.569
SSL	AlexNet [38]	.25	.35	.39	.75	.87	.89	.47	.570
	DRML [38]	.26	.36	.40	.79	.88	.89	49	.581
	GFK [38]	.19	.31	.32	.74	.85	.86	.39	.522
	TSVM [38]	.24	.32	.40	.76	.87	.88	.47	.564
	AlexNet - Assoc. [16] (*)	.33	.58	.37	.72	.78	.85	.47	.584
	AlexNet - Pseudo–Label [15] (**)	.31	.56	.28	.71	.76	.83	.50	.564
	AlexNet - [30]	.69	.60	.40	.72	.77	.82	.50	.597
	AlexNet - WTE	.37	.60	.40	.73	.78	.86	.51	.604
	VGG-16 - WTE	.47	.65	.41	.74	.79	.85	.52	***.635***

We have considered a series of unlabeled sets extracted from the remainder of the original RAF-DB training set. When extracting the datasets, we have computed the Kullback-Leibler divergence between the labels of this experiment training set and the labels of the extracted set. In the training, the images from the extracted set were considered unlabeled, so to form the semi-supervised framework. The KL divergence is 0 when the two sets have the same distribution of labels and it has larger values when the bias increases. The unlabeled data has been extracted on a random basis from available images. For each method in at least one case, one of the labels category was completely absent.

To ensure bias, for experiments characterized by KL divergence larger than 0, the number of unlabeled images increased. More specifically, while for $KL = 0$ a set of 3068 images were considered, around 6000 images were used when we obtained $KL = 1.3$.

For this experiment, we have compared the proposed solution with Pseudo–Label [15], as in the previous tests it showed the best performance. The comparative performance can be seen in Figure 7. The performance of the Pseudo–Label slightly decreases while the bias increases. In the same time, the timid SSL stayed constant.

The slight increase of the performance for the timid SSL solution has been connected with a larger unlabeled set. The same behavior has not been encountered for the other considered solution.

4.7. Impact of the diversity

Another problem analyzed in this paper is how much diversity is required to improve the performance. In other words, we seek to quantify the variation of the performance of the solution with respect to the number weak experts used. To investigate this aspect, we have considered the DenseNet 201 architecture on the FER+ experiment. The results may be followed in Table 13. The experts have been randomly selected and we have ran three times the training/testing procedure for each case. The reported accuracy is the average of the three attempts. As one can see, the minimum amount to obtain improvement with respect to the case when the training was purely supervised is around 50% (using 16 experts out of 33). Too few experts lead to a decrease in performance. The increase in accuracy is in direct relation with the number of number of experts and thus with the diversity.

5. Discussion

Due to the difficulty of annotation, face expression related tasks are the almost ideal candidate for the use of the semi-supervised learning. An often encountered, practical, scenario is to use an entire database for the labeled part and another database, consistent with the first but with different content, as the unlabeled part. Our proposed strategy, named *timid SSL* was based on diversity. It showed consistently improved results when the two datasets may have a bias, as verified in several different scenarios. Two such scenarios regard face expression recognition in the wild (i.e. one on large color images - RAF-DB and one on small gray ones - FER+), and two regard the AU estimation (i.e. detection on images in the wild - EmotioNet and intensity estimation on spontaneous sequences - DISFA).

Other recently introduced SSL methods [15, 17, 16, 20] make use of the same learner to explore and label the unlabeled part, falling in the category of confident SSL methods. They have showed impressive performance in scenarios where the labeled and unlabeled data are from the same distribution and the labeled data is scarce. Intuitively, they use the unlabeled data to cement the path chosen by the optimizer due to the labeled set. Our findings, consistent with previous evaluation [9], showed that such a strategy is not successful when data from different databases is used, as it, probably, originates in other distributions. Here, confident SSLs cannot use the new data to explore

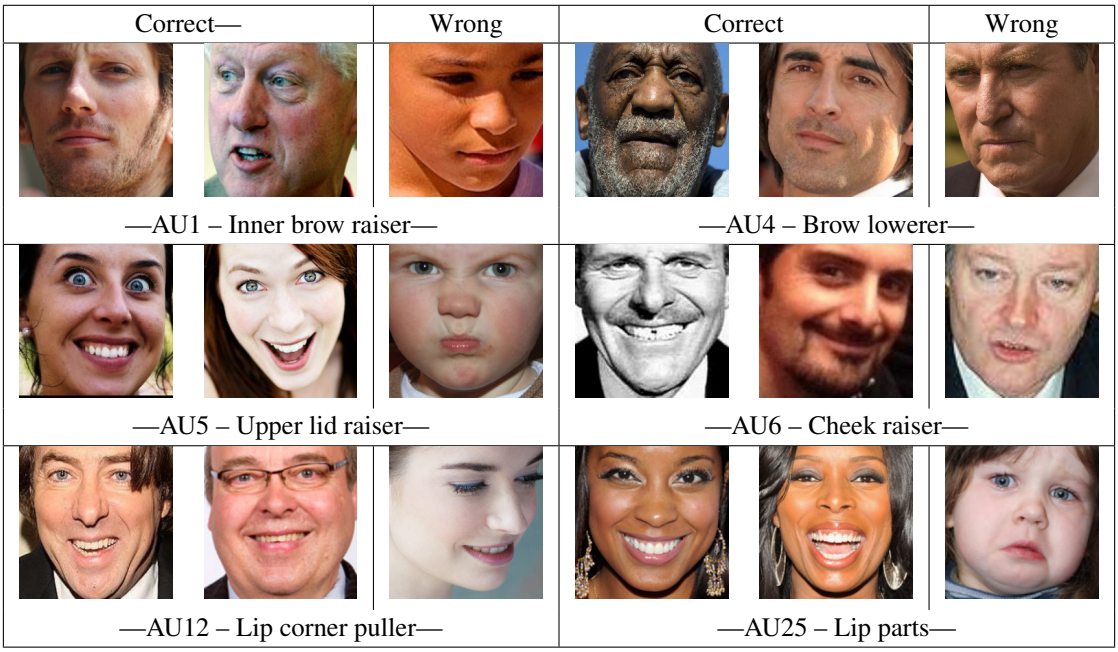


Figure 6: Examples of images from the EmotioNet database. Examples marked with ”Correct” are true positive, while those marked with ”wrong” are missed detections (false positives) with respect to named AU.

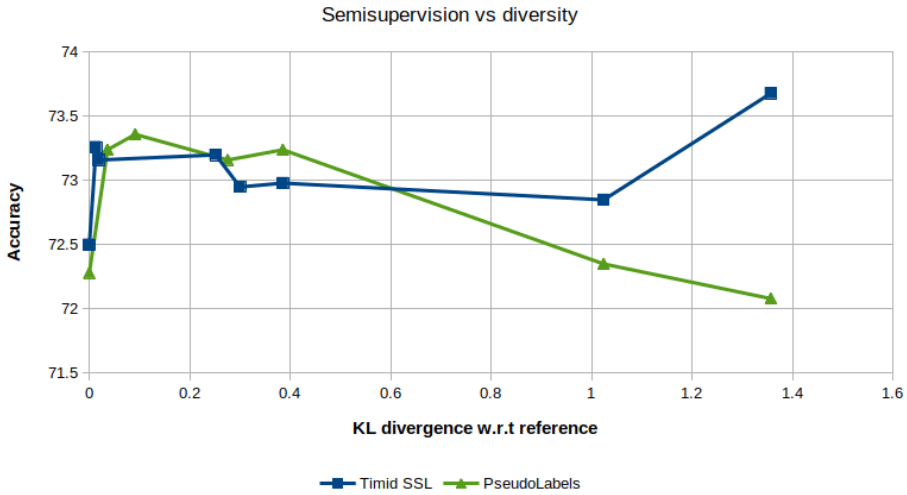


Figure 7: The performance of the proposed timid SSL when compared to the PseudoLabels solution when different distributions between the labeled and unlabeled set are used in the semi-supervised framework. Baseline performance is 71.8%.

Table 9: Intra–class Correlation Coefficient for Action Unit Intensity estimation on the DISFA dataset. SSL - marks the semi-supervised learning solutions. OSVR stands for ordinal SVR, BN for Bayesian Network, CCNN – copula conditional neural network, DC – Deep Coder, DOR – Dynamic-Ordinal-Regression, BORMIR - Bilateral Ordinal Relevance Multi-Instance Regression.

	Method/AU	1	2	4	5	6	9	12	15	17	20	25	26	Avg.
SUPERVISED	AlexNet - baseline	.25	.13	.35	.52	.51	.43	.80	.03	.45	.10	.79	.57	0.411
	VGG-16 - baseline	.5	.27	.68	.55	.57	.52	.75	.17	.38	.22	.84	.53	0.495
	OSVR [36]	.16	.12	.43	.06	.62	.54	.82	.43	.37	.28	.77	.53	0.418
	LT [32]	.22	.02	.04	.10	.23	.04	.43	.04	.02	-.03	.29	.14	0.129
	OR-CNN [33]	.03	.07	.01	0	.29	.08	.67	.13	.27	0	.59	.33	0.195
	BN -base [31]	.23	.43	.37	.17	.45	.39	.63	.28	.44	.13	.68	.13	0.361
	BN – full [31]	.25	.28	.82	.10	.23	.44	.86	.46	.69	.01	.85	.14	0.43
	CCNN [2]	.18	.15	.61	.07	.65	.55	.82	.44	.37	.28	.77	.54	0.445
	VGG16-2DC [34]	.70	.55	.69	.05	.59	.57	.88	.32	.10	.08	.90	.50	0.50
	S-DOR [37]	.40	.47	.28	.35	.45	.11	.78	.20	.14	.09	.81	.32	0.37
WSL	BORMIR [35]	.2	.25	.30	.17	.38	.18	.58	.16	.23	.09	.71	.15	0.283
	OSVR [36]	.21	.04	.25	.15	.23	.15	.31	.12	.07	.09	.62	.09	0.194
	MI-DOR [37]	.40	.47	.28	.35	.45	.11	.78	.20	.14	.09	.81	.32	0.37
SSL	KBSS [39]	.23	.11	.48	.25	.50	.25	.71	.22	.25	.06	.83	.41	0.36
	Proposed AlexNet - EWT	.53	.2	.65	.28	.58	.64	.83	.02	.3	.03	.76	.61	0.453
	Proposed VGG-16 - EWT	.52	.26	.69	.57	.55	.59	.75	.29	.41	.23	.88	.55	***0.525***

new regions of the space, because, they will, likely, collapse the unlabeled data on the gradient path defined by the labeled one.

More precisely, following our experiments, one observes that the performance consistently improves as more complex network architectures are considered (ResNet > DenseNet201 > VGG-16 > DenseNet 121 > AlexNet). This trend does hold for confident SSLs based training. In our evaluation, Pseudo–Label lead to better accuracy than MixMatch, Association and respectively Mean Teacher, in parallel with complexity correlation.

We have designed a particular experiment, with images from RAF-DB, aimed to quantify robustness with respect to bias between labeled and unlabeled sets. There, the timid SSL showed stationary performance with respect to the bias and increased with respect to the number of images in the unlabeled set. This behavior was in contrast with the other tested solution, which showed to be affected by changes in the distribution of labels.

In tasks related to face expression, one may successfully use confident SSLs in a scenario where unitary acquisition is possible, yet annotation, due to costs, is not. There, the distributions are the same and performance improvement may be possible.

Although, initially, in all experiments, the labeled and unlabeled datasets seem compatible, experimental probing showed differences. To cope with them, we resorted to *timid* SSL, where a diverse ensemble of weak teachers is used to structure and explore the data space. Of vital importance is to use in the ensemble learners that behave differently from the main learner and in such a manner new parts of the data space can be explored by the gradient optimization. Experiment on EmotioNet showed that diversity behaves better than mere performance while exploring the unlabeled part.

Often, in SSL testing, experiments designed to be cross database are actually done inside a single database, for reasons of simplicity. While in this way the i.i.d. hypothesis is met, this diverts from the aim of performance improvement.

In experiments where multiple data sources should have been envisaged, systems that function rather far from the optimum points may produce skewed conclusions and it is uncertain if the improvement is due to the design of the system or due to the recovery of the true optimum. A standard architecture, carefully trained, and combined with timid SSL is more robust to database bias and leads to better accuracy.

Acknowledgement

This work was supported by the Ministry of Innovation and Research, UEFISCDI, partially by the project SPIA-VA, agreement 2SOL/2017, grant PN-III-P2-2.1-SOL-2016-02-0002 and partially by the project TRANSLATE, TE 66/2020, PN-III-P1-1.1-TE-2019-0543. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

[1] K. Zhao, W.-S. Chu, H. Zhang, Deep region and multi-label learning for facial action unit detection, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3391–3399. 1, 2, 11

Table 10: Pearson Correlation Coefficient for Action Unit Intensity estimation on the DISFA dataset using semi–supervised learning. In the top part, we report the performance when manual annotated labels (ground truth) from CK+ are taken (marked with AlexNet + CK+ reference).

	Method/AU	1	2	4	5	6	9	12	15	17	20	25	AU26	Avg.
SUPERVIS.	AlexNet - only DISFA	.55	.38	.61	.50	.47	.48	.81	.01	.27	.06	.81	.78	0.476
	VGG-16 - only DISFA	.54	.34	.73	.56	.58	.56	.75	.17	.43	.24	.88	.58	0.530
	AlexNet with CK and reference	.65	.41	.72	.58	.62	.69	.85	.04	.42	.04	.83	.68	0.545
WSL	LT [32]	.28	.02	.08	.11	.31	.07	.52	.10	.04	-.03	.34	.23	0.173
	OSVR [36]	.24	.07	.33	.21	.26	.20	.34	.13	.08	.11	.66	.12	0.231
	BORMIR [35]	.26	.33	.36	.25	.45	.24	.63	.29	.32	.21	.74	.19	0.353
SSL	AlexNet - Pseudo–Label [15](**)	.31	.27	.51	.24	.37	.34	.65	.02	.29	0	.66	.47	0.351
	AlexNet - Mean Teacher [17](*)	.28	.17	.51	.23	.18	.39	.59	.03	.16	.06	.55	.33	0.254
	AlexNet - Association [16] (*)	.24	.06	.52	.4	.47	.43	.63	0	.15	.05	.55	.45	0.371
	Proposed AlexNet - WTE(CK+) all	.40	.14	.62	.40	.47	.48	.74	.04	.39	.13	.77	.70	0.441
	Proposed AlexNet - WTE(CK+) early	.46	.43	.61	.48	.53	.52	.82	0	.38	.11	.82	.79	0.496
	Proposed VGG-16 - WTE(CK+) early	.58	.32	.74	.55	.59	.6	.77	.35	.45	.26	.91	.59	**0.560**

Table 11: Accuracy obtained with proposed solution and respectively prior art on the FER+ dataset when the number of available image with labels is decreased to the amount mentioned in the top row. The unlabeled data remains as it was in the experiments listed in table 3. The CNN architecture used is DenseNet 201.

Solution/Labels	**400**	**4000**	**10000**	**All**
Timid SSL - WTE	40.58	58.09	70.23	84.87
Mean Teacher [17]	42.39	57.36	66.45	68.56
Pseudo-Label [15]	41.92	58.09	68.93	83.21
MixMatch [20]	41.40	58.21	67.28	82.18

[2] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller, M. Pantic, Deep structured learning for facial action unit intensity estimation, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5709–5718. 1, 3, 6, 15

[3] S. Li, W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, IEEE Trans. on Image Processing 28 (1) (2019) 356–370. 1, 2, 6, 7, 8, 10

[4] P. Ekman, W. Friesen, J. Hager, Facial action coding system: Research nexus, Network Research Information, Salt Lake City, 2010. 1, 2

[5] M. S. Bartlett, J. C. Hager, P. Ekman, T. J. Sejnowski, Measuring facial expressions by computer image analysis, Psychophysiology 36 (2) (1999) 253–263. 1

[6] J. Susskind, G. Littlewort, M. Bartlett, J. Movellan, A. Anderson, Human and computer recognition of facial expressions of emotion, Neuropsychologia 45 (1) (2007) 152–162. 1

[7] P. Ekman, E. L. Rosenberg, What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the FACS, Oxford Scholarship, 2005. 1

[8] S. Dodge, L. Karam, Can the early human visual system compete with deep neural networks?, in: Proceedings of the IEEE Int. Conf. on Computer Vision, 2017, pp. 2798–2804. 1

[9] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, I. J. Goodfellow, Realistic evaluation of deep semi-supervised learning algorithms, in: Proc. of Neural Information Processing(NIPS), 2018, pp. 3239–3250. 1, 2, 7, 8, 13

[10] A. Torralba, A. A. Efros, Unbiased look at dataset bias, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1521–1528. 1

[11] S. Fralick, Learning to recognize patterns without a teacher, IEEE Trans. on Information Theory 13 (1) (1967) 57–64. 2

[12] O. Chapelle, B. Scholkopf, A. Zien, Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews], IEEE Trans. on Neural Networks 20 (3) (2009) 542–542. 2, 3

[13] L. A. Pereira, R. da Silva Torres, Semi-supervised transfer subspace for domain adaptation, Pattern Recognition 75 (2018) 235 – 249. 2, 3

[14] F. G. Cozman, I. Cohen, Unlabeled data can degrade classification performance of generative classifiers, in: Proc. of Association for the Advancement of Artificial Intelligence (AAAI) Conf., 2002, pp. 327–331. 2

[15] D.-H. Lee, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: Proc. of Int. Conf. on Machine Learning (ICML) Workshops, 2013, pp. 1–6. 2, 3, 4, 5, 8, 10, 13, 16, 17

[16] P. Haeusser, A. Mordvintsev, D. Cremers, Learning by association-a versatile semi-supervised training method for neural networks, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2017, pp. 89–98. 2, 3, 4, 5, 8, 10, 13, 16

[17] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Proc. of Neural Information Processing Systems(NIPS), 2017, pp. 1195–1204. 2, 3, 4, 5, 8, 10, 13, 16, 17

[18] T. Miyato, S. ichi Maeda, S. Ishii, M. Koyama, Virtual adversarial training: a regularization method for supervised and semi-supervised learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (8) (2018) 1979–1993. 2, 3

[19] S. Laine, T. Aila, Temporal ensembling for semi-supervised learning, in: Fifth International Conference on Learning Representations (ICLR), 2017, pp. 1–13. 2

[20] D. Berthelot, N. Carlini, I. J. Goodfellow, N. Papernot, A. Oliver, C. Raffel, Mixmatch: A holistic approach to semi-supervised learning, in: Proc. of Neural Information Processing Systems (NIPS), 2019, pp. 5050–5060. 2, 3, 4, 5, 8, 10, 13, 16, 17

[21] C. A. Corneanu, M. O. Simón, J. F. Cohn, S. E. Guerrero, Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications, IEEE Trans. on Pattern Analysis and Machine Intelligence(T. PAMI) 38 (8) (2016) 1548–1568. 2

[22] H. Yan, Collaborative discriminative multi-metric learning for facial expression recognition in video, Pattern Recognition 75 (2018) 33 – 40. 2

[23] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, Z. Luo, Conditional convolu-

Table 12: Accuracy obtained with proposed solution based on the RAF-DB dataset with varying number of labels. The unlabeled data remains as in experiments listed in table 4. The CNN architecture used is DenseNet 201.

Solution /Labels	**400**	**1000**	**4000**	**All**
Timid SSL - WTE	41.41	48.86	61.60	83.15
Mean Teacher [17]	42.43	48.55	56.85	68.10
Pseudo-Label [15]	44.60	49.78	59.23	76.82
MixMatch [20]	41.68	50.21	61.48	73.21

Table 13: Performance (accuracy) FER+ database when various numbers of weak experts have been employed. The architecture is DenseNet 201.

Experts	**All - 100%**	**66%**	**48%**	**33%**	**0% - None**
Accuracy	83.15	82.84	81.64	80.98	81.50

tion neural network enhanced random forest for facial expression recognition, Pattern Recognition 84 (2018) 251 – 261. 2

[24] C.-M. Kuo, S.-H. Lai, M. Sarkis, A compact deep learning model for robust facial expression recognition, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 2121–2129. 2, 7, 8, 10

[25] S. Zhao, H. Cai, H. Liu, J. Zhang, S. Chen, Feature selection mechanism in CNNs for facial expression recognition, in: Proc. of British Machine Vision Conference (BMVC), 2018, pp. 1–12. 2, 7, 8, 10

[26] W. Xie, X. Jia, L. Shen, M. Yang, Sparse deep feature learning for facial expression recognition, Pattern Recognition 96 (2019) 106–966. 2, 8

[27] E. Barsoum, C. Zhang, C. C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in: Proc. of Int. Conf. on Multimedia Interfaces (ICMI), 2016, pp. 279–283. 2, 6, 7, 8

[28] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, F. Jurie, An occam's razor view on learning audiovisual emotion recognition with small training sets, in: Proc. of Int. Conf. on Multimedia Interfaces (ICMI), 2018, pp. 589–593. 2, 7, 8, 10

[29] X. Liu, B. V. Kumar, P. Jia, J. You, Hard negative generation for identity-disentangled facial expression recognition, Pattern Recognition 88 (2019) 1 – 12. 2

[30] C. Fabian Benitez-Quiroz, R. Srinivasan, A. M. Martinez, Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5562–5570. 2, 6, 11, 13

[31] S. Wang, J. Yang, Z. Gao, Q. Ji, Feature and label relation modeling for multiple-facial action unit classification and intensity estimation, Pattern Recognition 65 (2017) 71 – 81. 2, 15

[32] S. Kaltwang, S. Todorovic, M. Pantic, Latent trees for estimating intensity of facial action units, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2015, pp. 296–304. 3, 11, 15, 16

[33] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output cnn for age estimation, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4920–4928. 3, 15

[34] D. L. Tran, R. Walecki, O. Rudovic, S. Eleftheriadis, B. Schuller, M. Pantic, Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding, in: Proc. of Int. Conf. on Computer Vision (ICCV), 2017, pp. 3209–3218. 3, 6, 12, 15

[35] Y. Zhang, R. Zhao, W. Dong, B.-G. Hu, Q. Ji, Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2314–2323. 3, 6, 11, 15, 16

[36] R. Zhao, Q. Gan, S. Wang, Q. Ji, Facial expression intensity estimation using ordinal information, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3466–3474. 3, 6, 11, 15, 16

[37] A. Ruiz, O. Rudovic, X. Binefa, M. Pantic, Multi-instance dynamic ordinal random fields for weakly supervised facial behavior analysis, IEEE Trans. on Image Processing 27 (8) (2018) 3969–3982. 3, 6, 15

[38] K. Zhao, W.-S. Chu, A. M. Martinez, Learning facial action units from web images with scalable weakly supervised clustering, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2090–2099. 3, 11, 13

[39] Y. Zhang, W. Dong, B.-G. Hu, Q. Ji, Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7034–7043. 3, 15

[40] R. Caruana, A. Niculescu-Mizil, G. Crew, A. Ksikes, Ensemble selection from libraries of models, in: Proc. of Int. Conf. on Machine Learning, 2004, p. 18. 4

[41] A. Krogh, J. Vedelsby, Neural network ensembles, cross validation, and active learning, in: Proc. of Neural Information Processing Systems (NIPS), 1995, pp. 231–238. 4, 5

[42] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, E. Brossard, The megaface benchmark: 1 million faces for recognition at scale, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4873–4882. 6

[43] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, J. F. Cohn, Disfa: A spontaneous facial action intensity database, IEEE Trans. on Affective Computing 4 (2) (2013) 151–160. 6

[44] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR) Workshops, 2010, pp. 94–101. 6

[45] T. Joachims, Transductive inference for text classification using support vector machines, in: Proc. of Int. Conf. on Machine Learning (ICML), 1999, pp. 200–209. 11

[46] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in: Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2066–2073. 11