Teaching with Soft Label Smoothing for Mitigating Noisy Labels in Facial Expressions

Tohar Lukov[1], Na Zhao[1], Gim Hee Lee[1], and Ser-Nam Lim[2]

[1] Department of Computer Science, National University of Singapore
tohar@u.nus.edu, {zhaona, gimhee.lee}@nus.edu.sg
[2] sernam@gmail.com

Abstract. Recent studies have highlighted the problem of noisy labels in large scale in-the-wild facial expressions datasets due to the uncertainties caused by ambiguous facial expressions, low-quality facial images, and the subjectiveness of annotators. To solve the problem of noisy labels, we propose Soft Label Smoothing (SLS), which smooths out multiple high-confidence classes in the logits by assigning them a probability based on the corresponding confidence, and at the same time assigning a fixed low probability to the low-confidence classes. Specifically, we introduce what we call the Smooth Operator Framework for Teaching (SOFT), based on a mean-teacher (MT) architecture where SLS is applied over the teacher's logits. We find that the smoothed teacher's logit provides a beneficial supervision to the student via a consistency loss – at 30% noise rate, SLS leads to 15% reduction in the error rate compared with MT. Overall, SOFT beats the state of the art at mitigating noisy labels by a significant margin for both symmetric and asymmetric noise. Our code is available at https://github.com/toharl/soft.

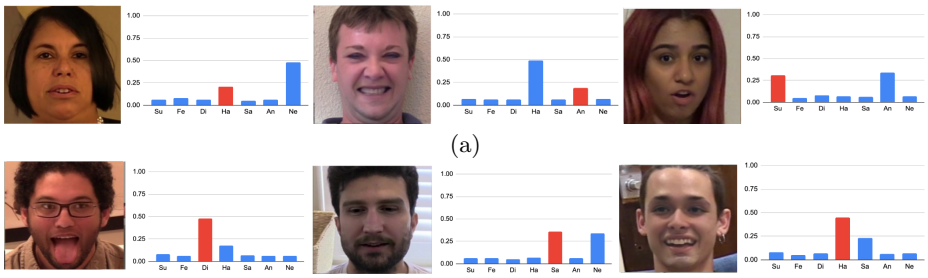
Keywords: Noisy labels, Facial expression recognition

1 Introduction

The problem of noisy labels in facial expressions datasets can be attributed to a few factors. On one hand, facial expressions can be fairly ambiguous, which leads to subjectiveness in the annotations. On the other hand, the prevalence of low-quality facial images, especially those collected from the wild, can also degrade the quality of the labels significantly. Mitigating the effect of noisy labels has thus become an important area of research in facial expression recognition (FER), where the goal is to prevent deep learning models from overfitting to the noisy labels.

Earlier studies [60,68] have shown that representing ground truth labels with label distributions (*i.e.*, multiple classes with different intensity) instead of one-hot label can help to mitigate the presence of noisy labels. Intuitively, facial expressions are often compound in nature, *e.g.*, an expression can appear both angry and sad at the same time, and a label distribution helps to capture the intricacy much better. To this end, researchers have looked into label distribution

learning (LDL) [68], label enhancement (LE) [5,44,60], and label smoothing regularization (LSR) [36,51]. Here, while the goals of LDL, LE and LSR are similar, LSR is in an analytic form, and thus is much more efficient in comparison.



(a)

(b)

Fig. 1: Label distributions generated by applying SLS to the prediction logits of the teacher. The given one-hot ground truth label is denoted in red. Su, Fe, Di, Ha, Sa, An, Ne denote Surprise, Fear, Disgust, Happiness, Sadness, Anger, Neutral, respectively.[3](a) Although the true label is ambiguous, SLS produced a label distribution that is better at describing the compound facial expression, even though the highest confidence class does not correspond to the ground truth. (b) While the highest class is predicted correctly as the ground truth class, the label distribution is better at describing the facial expression with more than one expression compared to that with only a single expression.

Inspired by these work, we propose in this paper that logits smoothing can similarly help to handle noisy labels effectively. Indeed, one of the main findings in this work is that by smoothing the logits, we are able to achieve significant performance boost in FER in the face of noisy labels. We propose Soft Label Smoothing (SLS) for logits, which smooths out multiple high-confidence classes in the logits by assigning them a probability based on the corresponding confidence, and at the same time assigning a fixed low probability to the low-confidence classes. While LSR can also be utilized for smoothing logits, it differs from SLS as it only assigns a fixed high probability to the highest confidence class. We say SLS is instance-aware as the distribution it produces varies per sample while LSR is only class-aware. Our study shows that SLS has a clear advantage over LSR.

We also further consider that the logits produced by a model for SLS to smooth should also have some tolerance to noisy labels. We are motivated by the mean-teacher (MT) architecture introduced in [52], which [39] shown can be utilized to detect noisy labels by considering the discrepancies between the

[3] Due to license restrictions, the images shown were not the actual images from RAF-DB from which the histograms were generated, but from the DFDC dataset [10] that have similar expressions

student and teacher's logits. Drawing from the success of MT in [39], we conjecture that a MT network by itself already has a stabilizing effect against noisy labels with the teacher "keeping the student in check" by "retaining" historical information as it is updated by the exponential moving average of the student's network. In this work, we employ a consistency loss that penalizes discrepancies between the student and teacher's logits, which are smoothed with SLS. With such a framework, which we refer to as the Smooth Operator Framework for Teaching (SOFT), we observe a significant boost in performance for FER, even with the challenging asymmetric (class-dependent) noise [4,40] where labels are switched to corresponding labels with the highest confusion instead of just random. In summary, the main contributions of our paper are as follow:

- We propose a novel framework, named SOFT, to mitigate label noise in FER. SOFT consists of a MT with SLS applied on the teacher's logits.
- Our simple approach does not require additional datasets, or label distributions annotations (such as in LDL), and does not cause additional computational cost during training.
- Our model produces state of the art performance for the FER task at different levels of noisy labels.

2 Related Work

2.1 Facial Expression Recognition

FER algorithms can be divided into two categories: handcrafted and learning-based techniques. Examples of the traditional handcrafted features based methods are SIFT [8], HOG [9], Histograms of local binary patterns [43] and Gabor wavelet coefficients [34]. Learning-based strategies [53,61] have become the majority with the development of deep learning and demonstrate high performance. Several employed two stream network to fuse face images with landmarks [67] and optical flows [50]. Some, such as [30,61,7], leverage the differences between expressive and neutral facial expressions. Recently, Ruan et al. [42] use a convolutional neural network to extract basic features which are then decomposed into a set of facial action-aware latent features that efficiently represent expression similarities across different expressions. However, all of these methods are not designed to deal with noisy labels and ambiguous facial expressions.

2.2 Learning with Noisy Labels

Deep learning with noisy labels has been extensively studied for classification tasks [47]. One line of work proposes a robust architecture by adding a noise adaptation layer [6,49,17] to the network in order to learn the label transition matrix or developing a dedicated architecture [59,22,18]. Another line of work studies regularization methods to improve the generalizability. Explicit regularization [28,58], such as dropout [48] and weight decay [31], modifies the training loss while implicit regularization [20,63] includes augmentation [45] and label

smoothing [51,36] which prevents the model from assigning a full probability to samples with corrupted labels. Reducing overfitting during training increases the robustness to noisy labels. Other methods propose noise-tolerant loss functions such as absolute mean error [16], generalized cross entropy [65] or modifying the loss value by loss correction [40,25], loss reweighting [35,56], or label refurbishment [41,46]. Another key concept is sample selection [46,37] to select the clean samples from the noisy dataset and update the network only for them, which has been shown to work well when combined with other approaches [3,46].

The problem of noisy labels in facial expressions datasets is mainly caused by ambiguous facial expressions, the subjectiveness of annotators, and low-quality facial images. Label distribution learning [15,13,12,14,33,26] and label enhancement have been proposed for mitigating ambiguity in related tasks such as head pose and facial age estimation. Zhou et al. [68] is the first to address this by learning the mapping from the expression images to the emotion distributions with their emotion distribution learning (EDL) method. However, this method assumes the availability of label distributions as ground truth which is expensive to obtain. To address the unavailable issue of label distributions, Xu et al. [60] propose label enhancement (LE) mechanism utilising one-hot label. However, the proposed approach has a high time complexity due to K-NN search, limiting the size of the dataset that can be used for training.

Zeng et al. [62] address inconsistencies in annotations of FER datasets by obtaining multiple labels for each image with human annotations and predicted pseudo labels, followed by learning a model (IPA2LT) to fit the latent truth from the inconsistent pseudo labels. Wang et al. [54] suppress uncertain samples by learning an uncertainty score and utilising a relabeling mechanism in an attempt to correct the noisy labels. However, this work does not take into account of compound expression. Moreover, these methods treat the inconsistency as noise while ignoring noisy labels caused by ambiguous facial expressions. Chen et al. [5] construct nearest neighbor graphs for label distribution learning, which requires additional datasets for related auxiliary tasks. To deal with ambiguity, She et al. [44] introduce the Distribution Mining and the pairwise Uncertainty Estimation (DMUE) approach. DMUE works by constructing multiple auxiliary branches as the number of classes in order to discover the label distributions of samples.

3 Our Approach

3.1 Background and Notation

Given a FER labelled dataset (X, Y), each sample x is annotated by a one-hot label over C classes, $y \in \{0, 1\}^C$. However, for a dataset with noisy labels, the labels can be either wrong or ambiguous. This poses a challenge for training deep neural networks as they suffer from the memorization effect, fitting to the noisy labels with their large capability, and causing performance degradation.

Label Smoothing [51]. Label Smoothing Regularization [51](LSR) has been widely used for regularization which improves generalization and calibration.

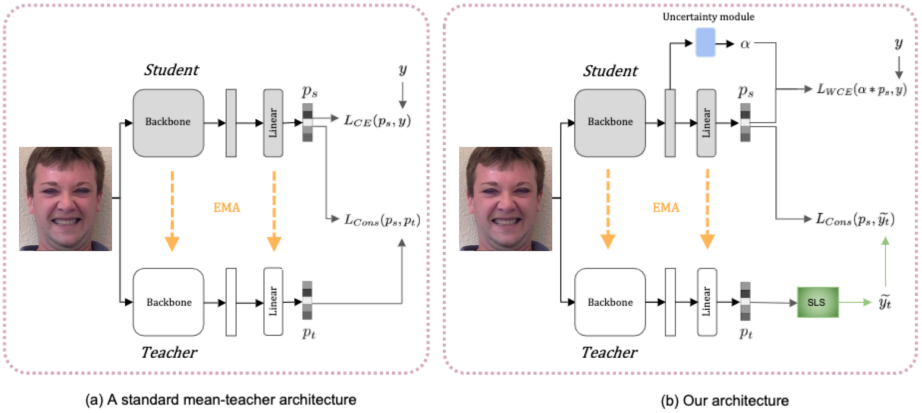


Fig. 2: Comparison between (a) standard mean-teacher architecture and (b) our architecture to learn from noisy labels in FER. Note that the given label, y, might be noisy. Our architecture is an extension of the mean-teacher architecture. We enhance the teacher's logits by applying our SLS on them. The student consists of an uncertainty module to predict the uncertainty score. The classification loss of the student is the cross entropy (CE) loss weighted by this predicted uncertainty score, which we will refer to as the weighted CE (WCE) loss following nomenclature in the literature.

When applying LSR, the one-hot label is modified such that the label is mixed with a uniform mixture over all possible labels. More formally, given one-hot label $y \in \{0,1\}^C$, LSR produces $\widetilde{y} \in \mathbb{R}^{1\times C}$ and is formulated as:

$$\widetilde{y} = (\widetilde{y_1}, \widetilde{y_2}, ..., \widetilde{y_C}), \tag{1}$$

where

$$\widetilde{y_i} = \begin{cases} 1-\epsilon, & i = l \\ \frac{\epsilon}{C-1}, & \text{otherwise.} \end{cases} \tag{2}$$

Here, ϵ is a hyper-parameter that is used to smooth the distribution and l is the index of the ground truth class.

Mean-Teacher. The mean-teacher architecture [52] is originally introduced for semi-supervised learning. We adapt it to deal with noisy labels in FER. Two networks (*i.e.* student and teacher) with the same architecture are forced to output consistent predictions, despite random noise introduced to the inputs or networks. The random noise can be implemented for example by applying dropout layers or random augmentations over the input for each branch. In our training, we use the latter. As can be seen in Figure 2 (a), the inputs to the

student and teacher networks are the same. The output logits from the student network are supervised by the given labels via a classification loss (*i.e.* cross entropy) and the teacher's logits via a consistency loss, respectively.

3.2 Overview of Our Method

Our architecture, shown in Figure 2 (b), is based on a mean-teacher where the given ground truth label (y) might be noisy with an unknown noise rate. We enhance the teacher's logits by applying our soft label smoothing (SLS). The student consists of an uncertainty module that predicts an uncertainty score to compute a weighted cross entropy (WCE) loss. During testing, the uncertainty module is removed. The best performing branch, either the student or teacher, over the validation set can be used for inference.

3.3 Soft Label Smoothing (SLS)

Our SLS preserves the high-confidences of multiple (top-k) predictions and unifies the remaining low-confidence predictions. The number of the high-confidence classes, k, depends on each instance as a result. More formally, for a sample x, we denote the logits, $p(x) \in \mathbb{R}^{1\times C}$. We further obtain $q = softmax(p)$ which is a distribution vector, $||q||_1 = 1$. We then define k as the number of elements above a threshold τ, which is empirically tuned:

$$k = \sum_{i=1}^{n} [q_i > \tau]. \tag{3}$$

Here, [...] are the Iverson brackets. Our SLS is then formulated as:

$$\widetilde{y_i} = \begin{cases} \frac{q_i}{\sum_{j=1}^{C} q_j [q_j > \tau]}(1-\epsilon), & q_i > \tau, \\ \frac{\epsilon}{C-k}, & \text{otherwise.} \end{cases} \tag{4}$$

Note that for samples with $k = 1$, SLS behaves like LSR. We show in our experiments later that $k > 1$ for a significant portion of the samples at the optimal τ, thus allowing SLS to play its intended role. In our framework, we apply SLS over the teacher's logits. The weights of the teacher network [52], θ', are updated only by the exponential moving average (EMA) of the weights from the student network θ:

$$\theta' \leftarrow \omega\theta' + (1-\omega)\theta, \tag{5}$$

where $\omega \in [0, 1]$ denotes the decay rate.

3.4 Loss Function

Consistency Loss. The student is supervised by the consistency loss which is formulated as the kullback–leibler (KL) divergence between the students' logits, p_s, and the teachers' soft labels after applying our SLS, $\widetilde{y_t} = SLS(p_t)$:

$$L_{Cons} = D_{KL}(p_s \parallel \widetilde{y_t}). \tag{6}$$

Logit Weighted Cross-Entropy Loss. For a sample x_i, we denote the feature vector produced by a backbone network as $f(x_i)$. This is given to the uncertainty module [27,54] which we attach to the student network and predict an uncertainty score α_i. The uncertainty module consists of a linear layer followed by a sigmoid function σ. It is formulated as:

$$\alpha_i = \sigma(W_u^\top (f(x_i)), \tag{7}$$

where W_u denotes the parameters of the linear layer. The logit weighted cross-entropy loss is formulated as:

$$L_{WCE} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{\alpha_i W_{y_i}^\top f(x_i)}}{\sum_{j=1}^{C} e^{\alpha_i W_j^\top f(x_i)}}, \tag{8}$$

where $f(x_i)$ is the feature vector and α_i is the uncertainty score of the i-th sample labelled as the y_i-th class (*i.e.* y_i denotes the index of the annotated class). $W_j^\top f(x_i)$ is the logit of the j-th class of sample i.

Total Loss. The total loss function for training the student network is given as:

$$L_{total} = L_{WCE}(\alpha * p_s, y) + \lambda L_{Cons}(p_s, \widetilde{y_t}), \tag{9}$$

where the score α is given by the uncertainty module. λ is the weight to control the contribution of the consistency loss.

4 Experiments

In this section, we first describe the three datasets we used in our experiments. We then present the ablation studies we conducted to demonstrate the efficacy of each component in our approach. Next, we present results that demonstrate the robustness of our approach against noisy labels by injecting varying amount of "synthetic" noise into the training dataset. Finally, we provide comparative results with several state-of-the-art approaches on the three datasets.

4.1 Datasets

RAF-DB [32] includes 30,000 face images that have been tagged with basic or compound expressions by 40 experienced human annotators. Only images with seven expressions (neutral, happiness, surprise, sadness, anger, disgust, fear) are utilised in our experiment, resulting in 12,271 images for training and 3,068 images for testing. For measurement, the overall sample accuracy is adopted.

AffectNet [38] is a very large dataset with both category and Valence-Arousal classifications. It comprises almost one million images retrieved from the Internet by querying expression-related keywords in three search engines, 450,000 of which are manually annotated with eight expression labels ('contempt' is annotated in addition to the expressions in RAF-DB). We train and test either on these eight emotions or only seven emotions (without 'contempt'), which we denote by AffectNet-8 and AffectNet-7 respectively. For measurement, the mean class accuracy on the validation set is employed following [55,54,44].

FERPlus [2] is an extension to FER2013 [19], including 28,709 training images and 3,589 testing images resized to 48x48 grayscale pixels. Ten crowd-sourced annotators assign each image to one of eight categories as in Affectnet. The most popular vote category is chosen as the label for each image as in [55,54,44].

Since the noisy labels in these datasets are unknown, we follow [54,44] to inject synthetic label noise to them, in order to assess the denoising ability of our method. Specifically, we flip each original label y to label y' by a label transition matrix T, where $T_{ij} = Pr[y' = j|y = i]$. We use two types of noise, symmetric (*i.e.* uniform) [54,44] and asymmetric (*i.e.* class-dependent) [4,40]. For symmetric noise, a label is randomly switched to another label to simulate noisy labels. On the other hand, asymmetric noise switches a label to the label that it is most often confused with (which can be identified from a confusion matrix). Figure 5(b) illustrates asymmetric and symmetric noise transition matrices for 30% noise rate, respectively.

Table 1: Mean Accuracy of the different components of our model. U denotes the use of an uncertainty module and SLS denotes our smoothing function over the teacher logits in a MT. Acc denotes using a pretrained model on the facial dataset, MS-Celeb-1M. Acc_i stands for training from a pretrained model on ImageNet.

MT	U	SLS	Acc_i	Acc
×	×	×	71.80	75.12
√	×	×	79.43	83.89
×	√	×	73.82	82.75
×	×	√	80.93	85.33
√	√	×	80.31	85.13
√	×	√	82.56	86.82
√	√	√	**83.18**	**86.94**

4.2 Implementation Details

We first detect and align input face images using MTCNN [64]. We then resize them to 224 × 224 pixels, followed by augmenting them with random cropping and horizontal flipping. As for our backbone network, we adopt the commonly

used ResNet-18 [24] in FER. For a fair comparison, the backbone network is pre-trained on MS-Celeb-1M with the standard routine [55,54,44]. The student network is trained by an Adam optimizer. We first pre-train the student with SLS for 6 epochs, and then add the uncertainty module for another 80 epochs with a learning rate of 0.0001 and a batch size of 64. The parameters of the teacher network is updated using EMA (see Eqn 5), and the weight decay is set to 0.999 following the original MT paper [52]. The loss weight λ (see Eqn 9) is set to 10. We use the better performing branch that is evaluated over the validation set (*i.e.* the student without the uncertainty module or the teacher) for inference.

4.3 Ablation Studies

SLS Effects. In the following, we examine the effects of SLS. All experiments are performed on RAF-DB with 30% injected symmetric noise unless specified otherwise. Additional experiments, including for asymmetric noise and other noise ratios, are provided in the supplementary material.

Mitigating label noise. We confirm empirically that SLS is beneficial for mitigating label noise. In Table 1, we isolate the effect of SLS and observe that SLS alone leads to an additional error reduction of 15% when added to a MT. To assess the effect of SLS for mitigating noisy labels, we show that SLS not only improves the predictions on clean samples, but also corrects the predictions of the noisy samples. To demonstrate that, in Table 2, following [36], we report performance with and without SLS on the noisy and clean parts of the training data. The noisy part refers to the 30% of the training labels that are randomly selected and switched in a symmetric noise setting. Compared to performance without SLS, under "Noisy Correct" in the table, SLS causes a substantial number of the predictions in the noisy part to be corrected to the original correct labels, while at the same time under "Noisy Noise", SLS reduces by 52%, the noisy part from being predicted with the injected wrong labels.

Table 2: Performance (Acc) of our model trained with 30% symmetric noise on different parts of the training data. The training data is separated into clean and noisy parts (see Subsection 4.3). By applying SLS, not only is the test and train accuracy higher, the accuracy on the noisy part is higher ("Noisy Correct") as well, with the original correct labels predicted instead of the injected noisy labels. Moreover, under "Noisy Noise", which denotes the number of samples in the noisy part predicted with the injected wrong labels, SLS was able to reduce the number significantly.

SLS	Test	Train	Noisy Correct	Noisy Noise
×	85.13	87.81	64.25	28.33
√	**86.94**	**89.67**	**75.78**	**13.57**

Design of SLS. We now present our findings on the three main design components of SLS, namely: 1) instance-awareness, 2) non-zero low confidences, and 3) applying on the teacher's logits. To assess the contribution of each one of these

Table 3: The effects of different design components for SLS. In both tables the first row is a MT with uncertainty module without smoothing. Table (a) shows the three main design choices in SLS: 1) Applied over the teacher's logits, 2) Instance-aware smoothing, 3) Non zero low-confidences. Table (b) shows a comparison of the branch to which SLS is applied on (i.e, the student's logits, the teacher's logits or both). See Subsection 4.3 for more details.

(a) Different Smooth Methods

Smooth method	Teacher	Ins-aware	Non-zero	Acc
×	×	×	×	85.13
LSR	×	×	×	85.43
LSR*	√	×	×	85.85
SLS(0)	√	√	×	85.30
SLS	√	√	√	**86.94**

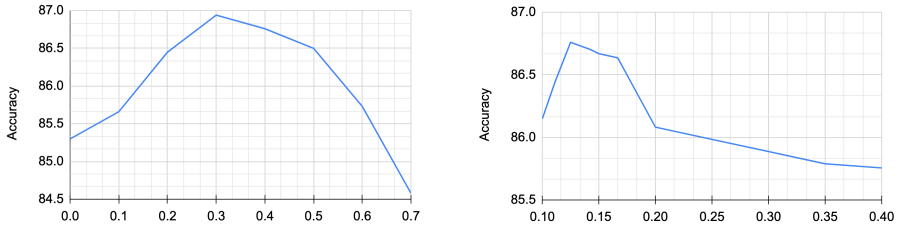
(b) Different Targets

Student	Teacher	Acc
×	×	85.13
√	×	85.23
×	√	**86.94**
√	√	85.36

components, in Table 3a, we present ablation studies by performing experiments with different smoothing methods. Table 3b compares the performance between where SLS is applied on, namely, the student's logits, the teacher's logits or both of them. The first row in both tables is a MT with uncertainty module without applying smoothing at all. We now discuss our findings here in details:

1. **Instance-awareness.** SLS is an instance-aware smoothing mechanism. For each sample, it utilizes the original confidence of the multiple high-confidence predictions. In Table 3a, we isolate this effect by comparing SLS with LSR*. LSR* denotes a version of LSR, where instead of a one-hot label, the input is now the logits, which is first transformed into a one-hot label by taking the top-1 prediction. LSR* is not considered as instance-aware since it ignores the original intensities of each sample, instead it produces the same smoothed label for all the samples with the same top-1 predicted class. We note that both SLS and LSR* are applied on the teacher logits. We observe that SLS performs better than LSR* by 1% which shows the advantage of being instance-aware.
2. **Non-zero low confidences.** The instance-awareness of SLS is derived from the utilization of multiple high confidence classes in its calculations. We are also curious on the effect of the way SLS handles low confidence classes. In Table 3a, we compare SLS with SLS(0). Referring to Eqn. 4, the latter is SLS with $\epsilon = 0$ which zeros out the low confidence classes. SLS outperforms the accuracy of SLS(0) by 1.6%, demonstrating the contribution of the non zero mechanism.
3. **The teacher benefit.** We explore the benefit of applying SLS over the teacher's logits in a MT vs in a vanilla network in Table 1, rows 4 (SLS) and 6 (MT+SLS). We train a vanilla network consisting of a backbone and classification layer. We apply SLS over the logits and supervise the network with a consistency loss between the produced logits and the logits after applying

SLS. The results demonstrates that indeed MT increases the performance by 1.5% over the vanilla network. We also explore the benefit of applying SLS over the teacher's logits vs the student's logits in Table 3b. In Table 3b, we can see that indeed applying SLS to the teacher's logits is more beneficial to the performance (with an increase of 1.7%) when compared to applying it on the student's logits or both. This is an interesting result since one would naturally think that applying to both the teacher's and student's logits could produce better results. Our conjecture is that this is due to the student being more prone to the noisy labels since it is the classification branch that ingests them, while the teacher is a more stable branch. As a result, applying SLS on the student could in fact accentuate the noisy labels.



(a) ϵ (b) τ

Fig. 3: Mean Accuracy(%) on the RAF-DB dataset with 30% symmetric noise for (a) varying smoothing parameter ϵ, and (b) varying τ.

Evaluation of varying ϵ. Figure 3(a) shows the Mean Accuracy(%) of varying smoothing parameter ϵ, on RAF-DB with 30% symmetric noise. Performance peaks at $\epsilon = 0.3$, after which performance degrades with higher values of ϵ until it is worse than the baseline at $\epsilon = 0$. For these high ϵ values, the label distribution produced by SLS is no longer meaningful.

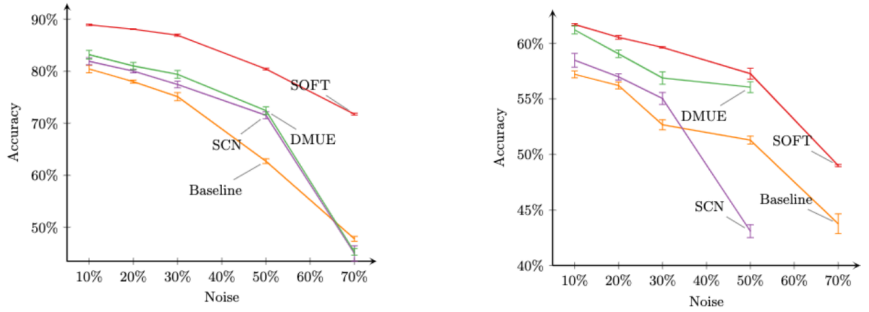
Evaluation of varying τ. Figure 3(b) shows the Mean Accuracy (%) of varying τ on RAF-DB with 30% symmetric noise. Referring to Eqn 3, for each sample, the value of τ and the original confidences q, influence the number of high confidences k. Here, we initialize all experiments with a MT trained for 6 epochs (without SLS). We also log the number of samples in the training data with more than one high confidence ($k > 1$) at the beginning of training with SLS. Performance peaks at $\tau = \frac{1}{8}$ for which we observe that 65% of samples have $k > 1$. Then, performance degrades with higher values of τ and lesser samples with $k > 1$ until τ reaches 0.4, by when all samples have only one high-confidence class.

Qualitative results. After our model is trained with SLS in an MT architecture, we extract the logits of the teacher corresponding to different training images. We observe that the label distributions are consistent with how human would categorize the facial expressions. In Figure 1(b), we present examples for label distribution where the ground truth corresponds to the highest confidence

class in the logits. In Figure 1(a), on the other hand, even though the highest confidence class is different from the ground truth, the former is actually plausible and might be even better than the ground truth label.

4.4 Comparison to the State of the Art

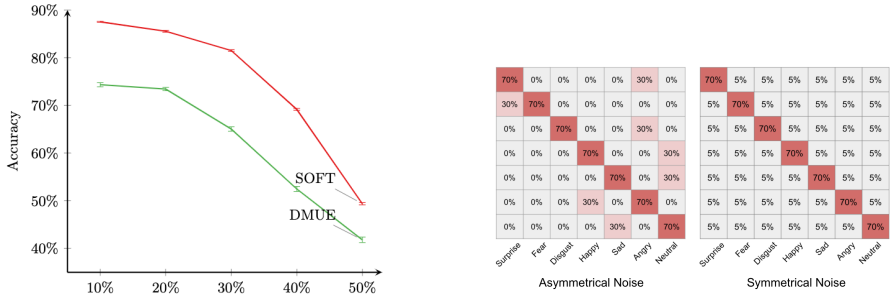
Our ablations, particularly Table 1 and 3a, show that SLS in MT produces the best performance. In addition, an uncertainty module added to the student branch provides an additional modest improvement. In this section onwards, we will refer to SOFT as one that includes SLS and an uncertainty module in a MT network.



(a) RAF-DB (b) AffectNet-8

Fig. 4: Accuracy on RAF-DB and AffectNet-8 with injected symmetric noise. SOFT consistently beats the other methods and performs well at noise rates as high as 70%. For AffectNet-8, the released code for DMUE and SCN does not work at 70% noise rate. A detailed table can be found in the Appendix.

Evaluation on Synthetic Symmetric Noise. We quantitatively evaluate the robustness of SOFT against mislabelled annotations on RAF-DB and AffectNet. We inject varying amount of symmetric noise to the training set. We compare SOFT's performance with the state-of-the-art noise-tolerant FER methods, SCN [54] and DMUE [44], as shown in Figure 4. For a fair comparison, we follow the same experimental settings such that all methods are pre-trained on MS-Celeb-1M with ResNet-18 as the backbone. Similar to the other methods' experimental settings, the baseline network is build with the same ResNet-18 backbone and fully-connected layer for classification. After three repetitions of each experiment, the mean accuracy and standard deviation on the testing set are presented. SOFT consistently beats the vanilla baseline and the other two state-of-the-art methods, SCN and DMUE. The improvement from SOFT becomes more significant as the noise ratio increases. We outperform the best performing method, DMUE, by 5.7%, 7%, 7.5% and 26.5% on RAF-DB with 10%, 20%, 30% and 70% noise rates, respectively. Also, we outperform DMUE by 0.1%, 1.9%, 2.8% on AffectNet-8 with 10%, 20%, 30% noise rates, respectively. With the noise rate as high as 70%, we improve the accuracy by 5.2% over the vanilla baseline on AffectNet-8. Please see a detailed table in the Appendix.



(a) Performance (b) Noise transition matrix (30%)

Fig. 5: (a) Accuracy on RAF-DB with injected asymmetric noise. SOFT consistently and significantly beats the state-of-the-art, DMUE. (b) Illustration of asymmetric and symmetric noise transition matrix for 30% noise rate as an example. We flip each original label y to label y' by the label transition matrix T, where $T_{ij} = Pr[y' = j|y = i]$.

Evaluation on Synthetic Asymmetric Noise. In Figure 5(a), we also show experiments on asymmetric (or class-dependent) noise [4,40], for which true labels are more likely to be flipped to a specific label. For example, 'Surprise' is most likely to be confused with 'Anger'. An example for asymmetric noise transition is given by the left matrix in Figure 5(b). This type of noise is a better representation of real-world corruption and ambiguity, but has not been investigated by previous methods for handling noisy labels in FER. We first obtain a confusion matrix after training a vanilla network. Subesequently, we use the top-1 mis-labeled class to construct the noise transition matrix for each class. As shown in Figure 5(a), SOFT consistently and significantly outperforms the state-of-the-art, DMUE, by 13%, 12%, 16% on RAF-DB with 10%, 20%, 30% noise rates, respectively. A detailed table reporting the comparison can be found in the Appendix. Beyond a 30% noise rate, the performance decreases significantly for both methods. Intuitively, at high asymmetric noise ratios, for example at 50% and assuming an evenly distributed training set, clean samples for a given label will likely only be present 50% of the time, with its sole wrong-label counterpart being the other 50%. This makes it very challenging for any models to learn the correct pattern. This phenomenon is not as severe in a symmetric noise setting, since the corresponding wrong labels are evenly distributed among the rest of the labels, so that the clean samples for a given class still dominate for that label, until we hit noise ratio of about 85%.

Comparison on Benchmarks. In the previous experiments, we demonstrated the noise-tolerance ability of SOFT. We now verify that this does not cause performance degradation for the original FER datasets (i.e., without any injected

Table 4: Comparison of FER state-of-the-art accuracy (without synthetic noise). $^+$ denotes both AffectNet and RAF-DB are used as the training set. * denotes using extra label distribution instead of one-hot label. Refer to Sec. 4.4 on the preprocessing procedure that we utilized for each of these datasets.

(a) AffectNet-7

Method	Noise-tolerant	Acc
LDL-ALSG$^+$ [5]	Y	59.35
CAKE [29]	N	61.70
DDA-Loss [11]	N	62.34
EfficientFace [66]	N	63.70
DAN [57]	N	65.69
SOFT	Y	**66.13**

(b) AffectNet-8

Method	Noise-tolerant	Acc
IPA2LT$^+$ [62]	Y	55.71
RAN [55]	N	59.50
EfficientFace [66]	N	59.89
SCN [54]	Y	60.23
DAN [57]	N	62.09
DMUE [44]	Y	**62.84**
SOFT	Y	62.69

(c) RAF-DB

Method	Noise-tolerant	Acc
LDL-ALSG$^+$ [5]	Y	85.53
IPA2LT$^+$ [62]	Y	86.77
SCN [54]	Y	87.03
SCN$^+$ [54]	Y	88.14
DMUE [44]	Y	88.76
DAN [57]	N	89.70
SOFT	Y	**90.42**

(d) FERPlus

Method	Noise-tolerant	Acc
PLD* [2]	N	85.10
SeNet50* [1]	N	88.80
SCN [54]	Y	88.01
RAN [55]	N	88.55
DMUE [44]	Y	**88.64**
SOFT	Y	88.60

noise). Table 4 compares SOFT to the state-of-the-art FER methods on AffectNet (7 or 8 classes), RAF-DB and FERPlus datasets. Among all the FER methods, LDL-ALSG, IPA2LT, SCN and DMUE are the only noise-tolerant FER methods, among which DMUE achieves the best results. SOFT outperforms these recent state-of-the-art methods on RAF-DB, Affectnet-7 (with 90.42% and 66.13%, respectively) and is comparable with the other methods on AffectNet-8 and FERPlus (with 62.69% and 88.6%, respectively).

5 Conclusion

We have introduced in this paper a novel solution to deal with noisy labels in FER - SOFT, which incorporates a soft label smoothing technique (SLS) into the Mean-Teacher paradigm. Through extensive ablations and benchmarks that we presented in this paper, we show that there is strong empirical support for SOFT. What is unexplored in this paper, however, is the applicability of SOFT to other use cases outside of FER, as well as the effect of SLS on other "student-teacher like" architectures, including contrastive and self-supervised learning frameworks [23,21], both of which also compare pairs of samples during training. We are hopeful that the findings about SOFT would be valuable in helping researchers pursue these directions.

References

1. Albanie, S., Nagrani, A., Vedaldi, A., Zisserman, A.: Emotion recognition in speech using cross-modal transfer in the wild. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 292–301 (2018)
2. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 279–283 (2016)
3. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. Advances in Neural Information Processing Systems **32** (2019)
4. Blanchard, G., Flaska, M., Handy, G., Pozzi, S., Scott, C.: Classification with asymmetric label noise: Consistency and maximal denoising. Electronic Journal of Statistics **10**(2), 2780–2824 (2016)
5. Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., Rui, Y.: Label distribution learning on auxiliary label space graphs for facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13984–13993 (2020)
6. Chen, X., Gupta, A.: Webly supervised learning of convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1431–1439 (2015)
7. Chen, Y., Wang, J., Chen, S., Shi, Z., Cai, J.: Facial motion prior networks for facial expression recognition. In: 2019 IEEE Visual Communications and Image Processing (VCIP). pp. 1–4. IEEE (2019)
8. Cheung, W., Hamarneh, G.: n-sift: n-dimensional scale invariant feature transform. IEEE Transactions on Image Processing **18**(9), 2012–2021 (2009)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 1, pp. 886–893. Ieee (2005)
10. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020)
11. Farzaneh, A.H., Qi, X.: Discriminant distribution-agnostic loss for facial expression recognition in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 406–407 (2020)
12. Gao, B.B., Xing, C., Xie, C.W., Wu, J., Geng, X.: Deep label distribution learning with label ambiguity. IEEE Transactions on Image Processing **26**(6), 2825–2838 (2017)
13. Geng, X.: Label distribution learning. IEEE Transactions on Knowledge and Data Engineering **28**(7), 1734–1748 (2016)
14. Geng, X., Qian, X., Huo, Z., Zhang, Y.: Head pose estimation based on multivariate label distribution. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
15. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. IEEE transactions on pattern analysis and machine intelligence **35**(10), 2401–2412 (2013)
16. Ghosh, A., Kumar, H., Sastry, P.: Robust loss functions under label noise for deep neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)

17. Goldberger, J., Ben-Reuven, E.: Training deep neural-networks using a noise adaptation layer (2016)
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
19. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: A report on three machine learning contests. In: International conference on neural information processing. pp. 117–124. Springer (2013)
20. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
21. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
22. Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I., Zhang, Y., Sugiyama, M.: Masking: A new perspective of noisy supervision. Advances in neural information processing systems **31** (2018)
23. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
25. Hendrycks, D., Mazeika, M., Wilson, D., Gimpel, K.: Using trusted data to train deep networks on labels corrupted by severe noise. Advances in neural information processing systems **31** (2018)
26. Hou, P., Geng, X., Zhang, M.L.: Multi-label manifold learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 30 (2016)
27. Hu, W., Huang, Y., Zhang, F., Li, R.: Noise-tolerant paradigm for training face recognition cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11887–11896 (2019)
28. Jenni, S., Favaro, P.: Deep bilevel learning. In: Proceedings of the European conference on computer vision (ECCV). pp. 618–633 (2018)
29. Kervadec, C., Vielzeuf, V., Pateux, S., Lechervy, A., Jurie, F.: Cake: Compact and accurate k-dimensional representation of emotion. arXiv preprint arXiv:1807.11215 (2018)
30. Kim, Y., Yoo, B., Kwak, Y., Choi, C., Kim, J.: Deep generative-contrastive networks for facial expression recognition. arXiv preprint arXiv:1703.07140 (2017)
31. Krogh, A., Hertz, J.: Ba simple weight decay can improve generalization,ˆ in advances in neural information processing systems, volume 4, je moody, sj hanson, and rp lippmann (1992)
32. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2852–2861 (2017)
33. Li, Y.K., Zhang, M.L., Geng, X.: Leveraging implicit relative labeling-importance information for effective multi-label learning. In: 2015 IEEE International Conference on Data Mining. pp. 251–260. IEEE (2015)
34. Liu, C., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image processing **11**(4), 467–476 (2002)

35. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. IEEE Transactions on pattern analysis and machine intelligence **38**(3), 447–461 (2015)
36. Lukasik, M., Bhojanapalli, S., Menon, A., Kumar, S.: Does label smoothing mitigate label noise? In: International Conference on Machine Learning. pp. 6448–6458. PMLR (2020)
37. Malach, E., Shalev-Shwartz, S.: Decoupling" when to update" from" how to update". Advances in Neural Information Processing Systems **30** (2017)
38. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
39. Nguyen, D.T., Mummadi, C.K., Ngo, T.P.N., Nguyen, T.H.P., Beggel, L., Brox, T.: Self: Learning to filter noisy labels with self-ensembling. arXiv preprint arXiv:1910.01842 (2019)
40. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1944–1952 (2017)
41. Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596 (2014)
42. Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., Wang, H.: Feature decomposition and reconstruction learning for effective facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7660–7669 (2021)
43. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and vision Computing **27**(6), 803–816 (2009)
44. She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., Mei, T.: Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6248–6257 (2021)
45. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data **6**(1), 1–48 (2019)
46. Song, H., Kim, M., Lee, J.G.: Selfie: Refurbishing unclean samples for robust deep learning. In: International Conference on Machine Learning. pp. 5907–5915. PMLR (2019)
47. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. arXiv preprint arXiv:2007.08199 (2020)
48. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
49. Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., Fergus, R.: Training convolutional networks with noisy labels. arXiv preprint arXiv:1406.2080 (2014)
50. Sun, N., Li, Q., Huan, R., Liu, J., Han, G.: Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recognition Letters **119**, 49–61 (2019)
51. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
52. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017)

53. Wang, C., Wang, S., Liang, G.: Identity-and pose-robust facial expression recognition through adversarial feature learning. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 238–246 (2019)
54. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6897–6906 (2020)
55. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing **29**, 4057–4069 (2020)
56. Wang, R., Liu, T., Tao, D.: Multiclass learning with partially corrupted labels. IEEE transactions on neural networks and learning systems **29**(6), 2568–2580 (2017)
57. Wen, Z., Lin, W., Wang, T., Xu, G.: Distract your attention: Multi-head cross attention network for facial expression recognition. arXiv preprint arXiv:2109.07270 (2021)
58. Xia, X., Liu, T., Han, B., Gong, C., Wang, N., Ge, Z., Chang, Y.: Robust early-learning: Hindering the memorization of noisy labels. In: International conference on learning representations (2020)
59. Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2691–2699 (2015)
60. Xu, N., Liu, Y.P., Geng, X.: Label enhancement for label distribution learning. IEEE Transactions on Knowledge and Data Engineering (2019)
61. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2168–2177 (2018)
62. Zeng, J., Shan, S., Chen, X.: Facial expression recognition with inconsistently annotated datasets. In: Proceedings of the European conference on computer vision (ECCV). pp. 222–237 (2018)
63. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
64. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
65. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems **31** (2018)
66. Zhao, Z., Liu, Q., Zhou, F.: Robust lightweight facial expression recognition network with label distribution training. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 3510–3519 (2021)
67. Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., Metaxas, D.N.: Learning active facial patches for expression analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2562–2569. IEEE (2012)
68. Zhou, Y., Xue, H., Geng, X.: Emotion distribution recognition from facial expressions. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 1247–1250 (2015)