# Iterative Distillation for Better Uncertainty Estimates in Multitask Emotion Recognition

Didan Deng[1], Liang Wu[2], Bertram E. Shi[3]
Department of Electronic and Computer Engineering,
Hong Kong University of Science and Technology, Kowloon, Hong Kong
{ddeng[1], lwuat[2]}@connect.ust.hk eebert@ust.hk[3]

## Abstract

*When recognizing emotions, subtle nuances in displays of emotion generate ambiguity or uncertainty in emotion perception. Emotion uncertainty has been previously interpreted as inter-rater disagreement among multiple annotators. In this paper, we consider a more common and challenging scenario: modeling emotion uncertainty when only single emotion labels are available. From a Bayesian perspective, we propose to use deep ensembles to capture uncertainty for multiple emotion descriptors,* i.e.*, action units, discrete expression labels and continuous descriptors. We further apply iterative self-distillation. Iterative distillation over multiple generations significantly improves performance in both emotion recognition and uncertainty estimation. Our method generates single student models that provide accurate estimates of uncertainty for in-domain samples and a student ensemble that can detect out-of-domain samples. Our experiments on emotion recognition and uncertainty estimation using the Aff-wild2 dataset demonstrate that our algorithm gives more reliable uncertainty estimates than both Temperature Scaling and Monte Carol Dropout.*

## 1. Introduction

Understanding human affective states is an essential task for many interactive systems (*e.g.*, social robots) or data mining systems (*e.g.*, user profiling). However, unlike object recognition tasks, emotion perception is strongly affected by personal bias, cultural backgrounds and contextual information (*e.g.*, environment), which increases the uncertainty of emotion perception.

To obtain a gold standard for emotion recognition, it is common to invite a number of annotators and take the most-agreed emotions as hard labels in emotion datasets [27, 33, 38, 36]. In datasets with huge number of samples [29], it is expensive to invite many annotators. Therefore, each sample is often annotated by one expert only. Single emotion labels cannot capture inter-rater disagreement.

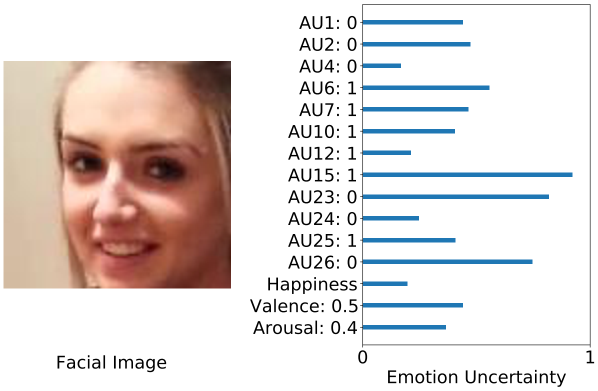

Figure 1: An example facial image with classifications generated by our model. The classifications are shown in the labels in the left. The estimated uncertainty is shown in the bar graph in the right. For example, the emotion is classified as ”Happiness”. The estimated valence and arousal scores are 0.5 and 0.4. The presence or absence of each AU is indicated by 1 or 0. The emotion uncertainty is the normalized Shannon’s entropy (*i.e.*, divided by its information length). Large uncertainty indicates low confidence. For example, AU15 ”lip corner depressor” is indicated as present, but with high uncertainty (low confidence). It is not present in the example image.

Previous work often related emotion ambiguity (uncertainty) with the variability among multiple raters’ annotations. For example, Mower *et al.* [30] characterized ambiguity using probability distributions assigned to the emotion classes. Han *et al.* [13] defined emotion uncertainty as *perception uncertainty* (*i.e.*, inter-rater disagreement). The solutions cannot be applied to emotion datasets with only single labels.

To address this problem, we adopt the Bayesian viewpoint and interpret emotion uncertainty as uncertainty in the posterior distribution over model weights. It is affected by the nosie, data distribution, and the model we choose. It does not require multiple raters' annotations, as required by the *perception uncertainty*.

In addition, we consider uncertainty simultaneously for multiple types of emotion labels (*i.e*., facial action units, basic emotions, valence and arousal), whereas past studies considered only single label types. Our intuition is that human affective states are quite complex, and should be described using a comprehensive set of emotional descriptors. Uncertainty among different emotion labels may be correlated. For example, the relationship between valence and arousal may be related to the uncertainty in perceived valence [2].

The recently released Aff-wild2 [22, 17] facilitates multitask emotion solutions [20, 24, 7]. The Aff-wild2 dataset has three types of emotion labels: facial action units, emotion categories, valence and arousal. Past emotion datasets [27, 33, 38, 21] usually have one or two types of emotion labels. However, Aff-wild2 dataset only provides single emotion labels, not multiple annotators' labels.

Using data from the Aff-wild2 dataset, we train deep ensembles with self-distillation algorithm to improve emotion recognition and uncertainty estimation. The obtained networks produce both emotions labels and the estimated uncertainty. The uncertainty is measured by Shannon's entropy computed over the probabilistic output. We give an example in Figure 1, showing the outputs of our model given a facial image input.

Our primary contributions are as follows:

- For better uncertainty estimation performance, we propose to apply deep ensembles learned by multi-generational self-distillation. The iterative training of neural networks improves not only uncertainty estimation, but also multitask emotion recognition.
- We design Efficient Multitask Emotion Networks (EMENet) for video emotion recognition. The visual model (EMENet-V) only has $1.68M$ parameters. The visual-audio (EMENet-VA) model has $1.91M$ parameters.
- We show that single models can estimate uncertainty reliably on in-domain data, and that the ensembles can detect out-of-distribution (OOD) samples.

## 2. Related Works

### 2.1. Uncertainty in Emotion

In emotion recognition, uncertainty often refers to *perception uncertainty*, in other words, inter-rater disagreement, requiring multiple annotators. Han *et al*. [13] took the standard deviation of $K$ emotion labels given by $K$ annotators as perception uncertainty. Zhang *et al*. [37] used Kappa coefficient to represent inter-rater agreement level. Uncertainty in emotion recognition has also been used to refer to the uncertainty in probabilistic models. A work in speech emotion recognition used a probabilistic Gaussian Mixture Regression (GMR) model to get the uncertainty of samples [5]. The authors found the emotion model performs better in low-uncertainty regions than high-uncertainty regions. Dang *et al*. [6] also used probabilistic models, and applied uncertainty when fusing predictions from sub-systems of multiple modalities. These past methods relied on handcrafted features.

### 2.2. Uncertainty Estimation

Ensemble-based methods are alternatives to Bayesian methods for estimating decision uncertainty. A Deep Ensemble [25] consists of several neural networks with the same architecture, but their weights are initialized independently. From a Bayesian viewpoint, the learned weights are ”sampled” from a posterior distribution. Deep ensembles have been shown to provide uncertainty estimates robust to dataset shifts [32]. Similar to deep ensembles, the Monte Carol Dropout (MC Dropout) [11] is a Bayesian approximation method for estimating uncertainty. MC Dropout method uses dropout during both training and testing. During inference, a dropout model is sampled $T$ times, and the $T$ predictions are averaged. Temperature Scaling [12] (TS) is a post-hoc calibration method to improve uncertain estimation. It optimizes the temperature value of the softmax function on a held-out validation set. The advantage of TS is that it does not increase computation during inference, but it is prone to overfitting.

### 2.3. Knowledge Distillation

Knowledge Distillation [14] was firstly proposed by Hinton *et al*. for model compression. A special case of knowledge distillation is the self-distillation algorithm [10], where the student model has the same architecture as its teacher model. The student model usually outperforms its teacher model, as shown in [10, 35]. The multi-generational self-distillation algorithm uses the student model in the previous generation as the teacher model in the next generation. As the number of generations increases, generalization performance improves [28]. Some studies have studied the reasons behind this phenomenon. For example, Mobahi *et al*. [28] proved mathematically that self-distillation amplifies regularization in the Hilbert space. Zhang *et al*. [39] related self-distillation to label smoothing, a commonly-used technique to prevent models from being over-confident [31]. They suggested that the regularization effect of self-distillation results from instance-level label smoothing. In this work, we aim to investigate the self-distillation for improving uncertainty performance by extending single models to deep ensembles.

# 3. Methodology

## 3.1. Notations

We denote the training set as $\{X, Y\}$, where $X$ denotes the input data and $Y$ denotes the ground truth labels. The input data can be divided into two categories $\{X^{vis}, X^{aud}\}$. $X^{vis}$ represents the visual data and $X^{aud}$ represents the audio data. $X^{vis}$ contains facial images, where $X^{vis} = \{x_i^{vis} | x_i^{vis} \in \mathbb{R}^{3\times H\times H}\}_{i=1}^{N}$. The facial images are RGB images with a height (width) of $H$ pixels. $X^{aud}$ contains mel spectrograms: $X^{aud} = \{x_i^{aud} | x_i^{aud} \in \mathbb{R}^{W\times W}\}_{i=1}^{N}$. The mel spectrograms have two dimensions: the number of mel-filterbank features and the number of audio frames. They are both $W$ in our experiments.

The ground labels $Y$ can be divided into three types: $Y = \{Y^{AU} \in \mathbb{R}^{N\times 12}, Y^{EXPR} \in \mathbb{R}^{N\times 7}, Y^{VA} \in \mathbb{R}^{N\times 2}\}$. $Y^{AU}$ contains 12 facial action units labels, including AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25 and AU26. They are multi-label binary values, denoting the presence or absence of corresponding action unit. $Y^{EXPR}$ are one-hot vectors denoting 7 basic emotions: neutral, anger, disgust, fear, happiness, sadness and surprise. $Y^{VA}$ are given by continuous values representing valence and arousal in range $\{-1, 1\}$. In our experiments, we transform regression tasks into classification tasks by discretizing continuous values. We discretize the valence score or the arousal scores into 20 bins, so that the shape of $Y^{VA}$ changes to $N \times 40$.

The single model function is denoted by $f_\theta$, where $\theta$ denotes the parameters. The ensemble model with $T$ models is denoted by $F_T$, where $F_T(x) = \frac{1}{T}\sum_{t=1}^{T} \sigma\left(f_{\theta_t}(x)\right)$. $\sigma(\cdot)$ is an activation function. The output of the ensemble is the average of its members' outputs.

In our teacher-student algorithm, a teacher ensemble with $T$ models can be denoted as $F_T^{tea}$. The soft labels generated by the teacher ensemble are denote as $F_T^{tea}(X)$. The student ensemble in the $k^{th}$ generation is denoted as $F_T^{stu_k}$. The soft labels generated by this ensemble for the $k+1$ generation is $F_T^{stu_k}(X)$. We run the self-distillation algorithm for $K$ generations in total.

## 3.2. Architectures

We aim to design efficient architectures while maintaining high performance in emotion recognition. Previous studies [7, 8, 23, 19] on video emotion recognition showed that CNN-RNN architectures generally outperformed CNN architectures. This indicates that affective states have strong temporal dependencies. Therefore, we choose efficient CNN architectures as feature extractors, and use GRU layers as temporal models to integrate information over time.

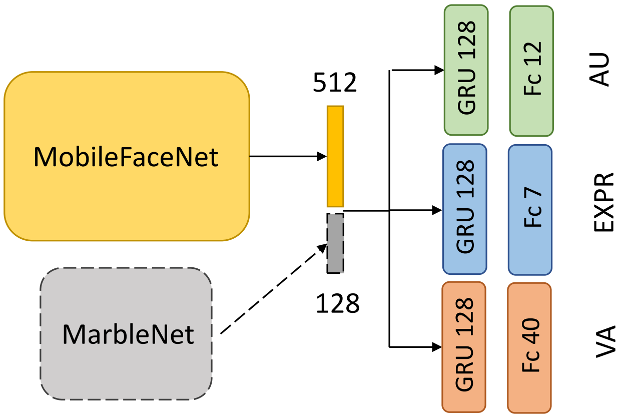

Figure 2: Our efficient model architecture. For visual modality, the feature extractor is the MobileFaceNet, and visual feature vector is a 512-dimensional vector. The weights surrounded by dashed curves are only included in the multimodal model architecture. For mutlimodal model, the MarbleNet is the audio feature extractor, and the audio feature vector is a 128-dimensional vector. The visual feature vector and the audio feature vector are concatenated before they are fed into temporal models.

For the visual modality, our model receives a sequence of facial images as input. The facial images are firstly processed by the MobileFaceNet [4]. The MobileFaceNet is a light-weighed CNN originally designed for face recognition on mobile devices. The model was pretrained on face alignment task [3]. We then finetuned the pretrained CNN weights. The feature vector is a 512-dimensional vector for each input image. The detailed architecture of the MobileFaceNet is given in [4].

For the visual-audio model, we usef a 1D-CNN to extract audio features from mel spectrograms. The audio CNN (MarbleNet) was proposed by Jia *et al.* [15] for voice activity detection. It has only $88K$ parameters. The audio feature vector extracted from one mel spectrogram is a 128-dimensional vector. The input is a $64 \times 64$ mel spectrogram, which is produced by extracting 64 mel-filterbank features from 64 audio frames (640ms). In our experiments, we sample a sequence of facial images as well as a sequence of mel spectrograms. The sample rate is same as the frame rate of the input video file. We denote the sequence length by $L$ and the batch size by $B$. The shapes of inputs to our visual-audio model are $(B, L, 3, 112, 112)$ for $X^{vis}$ and $(B, L, 64, 64)$ for $X^{aud}$. After the inputs are processed by feature extractors, the visual features and audio features are concatenated. This results in $(B, L, 640)$ feature vectors that are fed into the temporal models.

Each task has their own temporal model, which consists of one GRU layer, a ReLU activation function, and a linear output layer. The hidden sizes of all GRU layers are 128. We apply a $50\%$ random dropout on the input features to the temporal models. For the final activation function, we use Sigmoid for the AU detection, and Softmax for 7 basic emotions, and valence/arousal prediction.

### 3.3. Loss Functions

To train teacher models, we minimize the loss functions between the teacher outputs and the ground labels. We refer to these losses as *supervision losses*. To train student models, we minimize the loss functions between the student outputs and the soft labels, which are generated by the teacher models or the student models in the previous generation. We refer to these loss functions as the *distillation losses*.

***Supervision Losses***. For facial action units detection, we use a sum of class-reweighted binary cross entropy (BCE) functions. We reweight the losses using $p_c$ based on the ratios between positive samples and negative samples in the training data.

$$\mathcal{L}^{AU}(y, \tilde{y}) = \frac{1}{C}\sum_{c=1}^{C}\mathbf{BCE}(y_c, \tilde{y}_c), \tag{1}$$

$$\begin{aligned}\mathbf{BCE}(y_c, \tilde{y}_c) = -[p_c y_c \cdot \log\left(\sigma(\tilde{y}_c)\right) + \\ (1 - y_c) \cdot \log\left(1 - \sigma(\tilde{y}_c)\right)],\end{aligned} \tag{2}$$

$$p_c = \frac{\#\, negative\ samples\ in\ class\ c}{\#\, positive\ samples\ in\ class\ c}. \tag{3}$$

Variables $y$ denote ground truth labels and $\tilde{y}$ denote inferences generated by the teacher model. $\sigma(\cdot)$ in Equation 2 denotes the Sigmoid function. $C$ is the total number of action units.

For basic emotion categories, we use a reweighted cross entropy (CE) function. The weights are determined by the distribution of different classes in the training set.

$$\mathcal{L}^{EXPR}(y, \tilde{y}) = \mathbf{CE}(y, \tilde{y}), \tag{4}$$

$$\mathbf{CE}(y, \tilde{y}) = -p_c\sum_{c=1}^{C} y_c \cdot \log(\tilde{y}_c). \tag{5}$$

For valence/arousal predictions, we use the Concordance Correlation Coefficient (CCC) between the scalar outputs and the ground truth labels. The CCC is defined as follows:

$$\mathbf{CCC}(y_c, \tilde{y}_c) = \frac{2\rho\sigma_y\sigma_{\tilde{y}}}{\sigma_y^2 + \sigma_{\tilde{y}}^2 + (\mu_y - \mu_{\tilde{y}})^2}, \tag{6}$$

where $y_c$ denotes ground truth labels in a batch, and $\tilde{y}_c$ denotes the scalar predictions of valence or arousal. $\rho$ is the correlation coefficient between the ground truth labels and the predictions. $\mu_y$, $\mu_{\tilde{y}}$, $\sigma_y$ and $\sigma_{\tilde{y}}$ are the means and standard deviations computed over the batch. Since our model produces a 20-dimensional softmax vector for valence/arousal, we compute the expectation values over the 20 bins in the range of $[-1, 1]$ to transform probabilistic outputs to scalar outpus.

We compute two CCCs: one for valence, and one for arousal. The supervision loss for valence and arousal prediction is:

$$\mathcal{L}^{VA}(y, \tilde{y}) = \sum_{c=1}^{2}(1 - \mathbf{CCC}(y_c, \tilde{y}_c)). \tag{7}$$

**Distillation losses**. For action units detection, we use the binary cross entropy between the soft labels and the outputs of the student models.

$$\mathcal{H}^{AU}(y^{tea}, \tilde{y}^{stu}) = \mathbf{BCE}(y^{tea}, \tilde{y}^{stu}), \tag{8}$$

$$\begin{aligned}\mathbf{BCE}(y^{tea}, \tilde{y}^{stu}) = -[p_c y^{tea} \cdot \log\left(\sigma(\tilde{y}^{stu})\right) + \\ (1 - y^{tea}) \cdot \log\left(1 - \sigma(\tilde{y}^{stu})\right)],\end{aligned} \tag{9}$$

where $p_c$ in Equation 9 is the same as $p_c$ in Equation 2. $y^{tea}$ represents the soft labels. $y^{tea} = F_T^{tea}(x)$ if it is in the first generation. $y^{tea} = F_T^{stu_k}(x)$ if it is in the $(k+1)^{th}$ generation. $\tilde{y}^{stu}$ is the output of a single student model.

For expression recognition, the distillation loss we use is the KL divergence (KLD) loss between the soft labels and the student outputs.

$$\mathcal{H}^{EXPR}(y^{tea}, \tilde{y}^{stu}) = \mathbf{KLD}(y^{tea}, \tilde{y}^{stu}). \tag{10}$$

For valence and arousal prediction, we still use negative CCC loss between the soft labels and scalar outputs of the student model.

$$\mathcal{H}^{VA}(y^{tea}, \tilde{y}^{stu}) = \sum_{c=1}^{2}(1 - \mathbf{CCC}(y^{tea}, \tilde{y}^{stu})), \tag{11}$$

where $y^{tea}$ and $\tilde{y}^{stu}$ are the scalar outputs of the teacher ensemble and the student model.

**Combination of losses**. We take a weighted sum of the losses for different tasks. For example, when training teacher models, we use a combination of supervision losses for different emotion tasks:

$$\mathcal{L} = \lambda^{AU}\mathcal{L}^{AU} + \lambda^{EXPR}\mathcal{L}^{EXPR} + \lambda^{VA}\mathcal{L}^{VA}. \tag{12}$$

When training student models, we use a combination of distillation losses for all tasks.

$$\mathcal{H} = \lambda^{AU}\mathcal{H}^{AU} + \lambda^{EXPR}\mathcal{H}^{EXPR} + \lambda^{VA}\mathcal{H}^{VA}. \tag{13}$$

When training multiple tasks, it is important to balance the weights of different losses according to the difficulty levels of different tasks. We propose a heuristic method to balance the weights. The weight of the $i^{th}$ task's loss

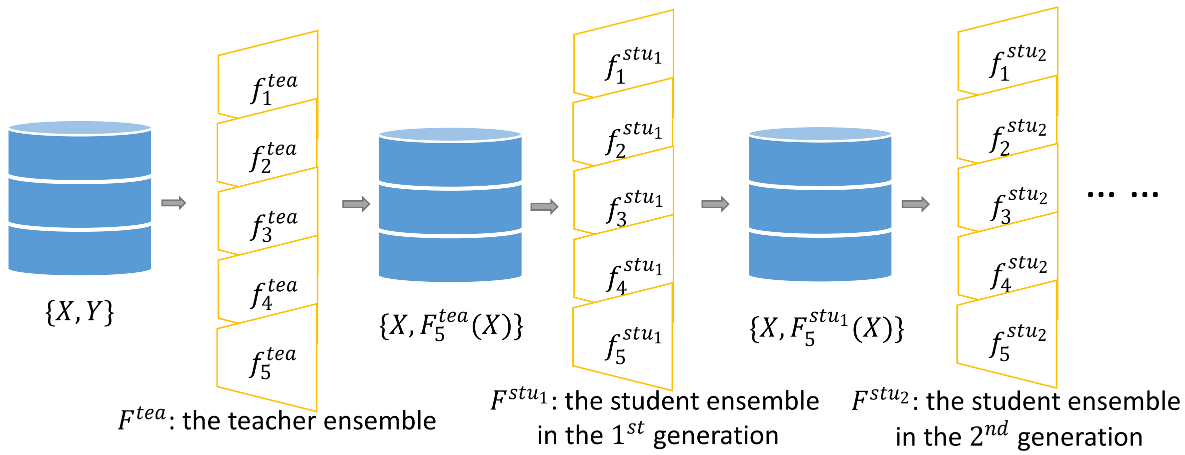

Figure 3: The diagram of our teacher-student algorithm.

depends on the number of epochs with no performance improvement on the validation set. If it is larger, we assign larger weights to this task's loss to increase its influence on the gradients. Appendix B shows the pseudo code for this heuristic method and ablation studies.

### 3.4. Algorithm

Our algorithm is a special case of the self-distillation algorithm. In the original self-distillation algorithm [10, 28], the teacher model and the student model are single models with the same architecture. In our algorithm, we propose to use deep ensembles for the following benefits:

1. The deep ensembles can be naturally trained on a distributed system. The parallel computing facilities parallel training of each local model, which saves training time.

2. The soft labels provided by the teacher ensemble contains more reliable uncertainty information than that provided by a single teacher model.

3. We can use one single model or a few models in our student ensemble to perform emotion tasks, which brings more flexibility when it comes to computation cost.

In Figure 3, we illustrate the teacher-student algorithm for deep ensembles. $\{X, Y\}$ is the original dataset. Most of instances in $\{X, Y\}$ only have one type of emotion labels, while other two types of emotion labels are missing. The $t^{th}$ teacher model $f_t^{tea}$ learns to fill in the missing labels. In the training batch of $\{X, Y\}$, we sample an equal number of instances for three tasks, and then compute the loss in Equation 12. Note that $\{f_t^{tea}\}_{t=1}^{T}$ are all trained on the same dataset $\{X, Y\}$, but start with different random initialization. After training $\{f_t^{tea}\}_{t=1}^{T}$ parallelly, we take average over their predictions on the training data. This generates the soft labels for the student models in the first generation. The soft labels are denoted as $F_T^{tea}(X)$.

In the first generation, student models $f_t^{stu_1}$ are trained on $\{X, F_T^{tea}(X)\}$. After all student models are trained, we use the student ensemble to generate the soft labels for the next generation. We iterative the teacher-student training in order to find the best number of generations.

## 4. Experiments

### 4.1. Dataset

We only used the video data from the Aff-wild2 [22] dataset. The Aff-wild2 dataset has three subsets, one for each emotion task. In each subset, the data distributions are quite unbalanced. Appendix A shows the data distributions for the three subsets. The data distributions determine the class weights $p_c$ in Equation 2 and 5. Appendix A also gives the choices of $p_c$.

### 4.2. Hyper-parameters

We used the Adam [16] optimizer. The learning rate was initialized as $1e^{-3}$. For visual model training, we trained the model for 10 epochs, and decreased the learning rate by a factor of 10 after every 3 epochs. For multimodal training, we trained the models for 15 epochs and decreased the learning rate by a factor of 10 after every 4 epochs.

### 4.3. Metrics

**Emotion metrics**. We used the same evaluation metrics as suggested in [18]. For facial AU detection, the evaluation metric is $0.5 \cdot \text{F1} + 0.5 \cdot \text{Acc}$, where F1 denotes the unweighted F1 score for all 12 AUs, and Acc denotes the total accuracy. For expression classification, we used $0.67 \cdot \text{F1} + 0.33 \cdot \text{Acc}$ as the metric, where F1 denotes the unweighted F1 score for 7 classes, and Acc is the total accuracy. For valence and arousal, we evaluated them with CCC.

**Uncertainty metric**. Same to [25], we evaluated the

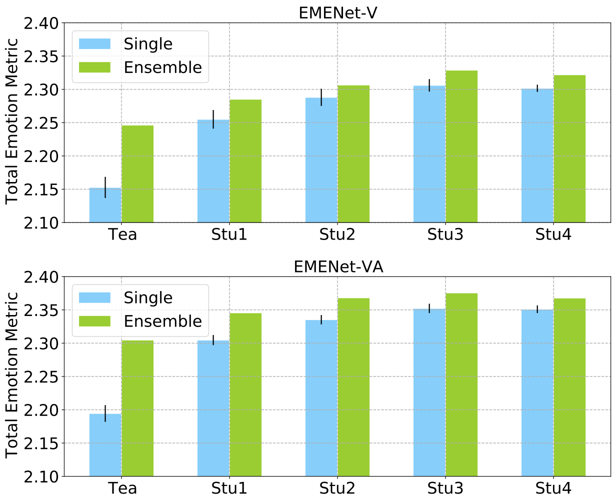

Figure 4: The total emotion metrics for the visual model (EMENet-V) and the visual-audio model (EMENet-VA), on the validation set of the Aff-wild2. For the single models' results, we average the total emotion metric over five runs, and the standard deviations are shown with error bars. "Tea" stands for the teacher model (ensemble). "Stu1" stands for the student model (ensemble) in the first generation.

in-domain uncertainty estimation performance using the negative log-likelihood (NLL) for classification tasks (*i.e.*, EXPR and AU) and root mean square error (RMSE) for regression tasks (*i.e.*, valence and arousal). Lower NLL or RMSE means better uncertainty estimation performance. For out-of-domain uncertainty performance, we created an OOD detection task by importing non-facial images, and evaluated the binary classification performance using ROC (receiver operating characteristic) curves and AUC (area under the ROC curve) scores.

# 5. Results

## 5.1. Task Performance

| Experiments | T | AU | EXPR | VA | | Total Emotion |
|---|---|---|---|---|---|---|
| | | | | Valence | Arousal | |
| w/o re. | 1 | 0.6773 | 0.5128 | 0.3830 | 0.5268 | 2.0999 |
| w/o re. | 5 | 0.6858 | 0.5354 | 0.4099 | 0.5537 | 2.1848 |
| w/o re. | 10 | 0.6843 | 0.5449 | 0.4105 | 0.5577 | 2.1974 |
| w/ re. | 1 | 0.6632 | 0.5541 | 0.4202 | 0.5192 | 2.1527 |
| w/ re. | 5 | 0.6808 | 0.5779 | 0.4423 | 0.5455 | **2.2465** |

Table 1: Validation results with the teacher models using visual modality only. "w/ re." means we apply class reweighting for EXPR and AU. $T$ is the number of models in an ensemble. $T = 1$ means it is a single model. Total emotion metric is the sum of all metrics of the three emotion tasks.

| Methods | # Gen. | # Param. | AU | EXPR | VA | |
|---|---|---|---|---|---|---|
| | | | | | Valence | Arousal |
| EMENet-V | 3 | 1.68M | 0.6320 | 0.4639 | 0.4942 | 0.4355 |
| EMENet-V | 3 | 8.4M | 0.6328 | 0.4704 | 0.5104 | 0.4419 |
| EMENet-VA | 2 | 1.91M | 0.6418 | **0.5046** | **0.5355** | 0.4442 |
| EMENet-VA | 1 | 9.55M | **0.6528** | 0.5041 | 0.5326 | **0.4537** |

Table 2: The emotion metrics on the test set of the Aff-wild2 dataset. "# Gen." denotes the number of generations. "# Param." denotes the number of parameters. The enmsebles have five times larger number of parameters than single models.

**Computation cost**. We designed two model architectures for the visual modality and visual-audio modalities respectively. We refer to them as EMENet-V and EMENet-VA. The number of parameters and FLOPs for EMENet-V are $1.68M$ and $228M$. For EMENet-VA, they are $1.91M$ and $234M$. The FLOPs are the number of floating-point operations when the visual input is one RGB image (112x112) and audio input is one spectrogram (64x64).

**Class reweighting**. We show the effect of class reweighting in Table 1. After applying class reweighting, we found the EXPR metric for single models was improved significantly, where the F1 score increased by $12.6\%$, and the accuracy of EXPR increased by $1.7\%$. Although the AU metric degraded after using class reweighting, its F1 score increased by $12.5\%$. The AU metric degraded because its accuracy dropped from $0.8947$ to $0.8249$. We think this is due to the highly unbalanced data distribution in the AU subset.

**Ensemble size**. We changed the ensemble size when training teacher models without class reweighting. The results are reported in Table 1. From single models ($T = 1$) to ensemble models ($T = 5$), the total emotion metric increased by $4\%$. However, from $T = 10$ to $T = 5$, the total emotion metric only increased by $0.58\%$. We kept the ensemble size $T = 5$ for the rest of experiments because of its relative efficiency.

**Teacher-Student training**. We trained our teacher models and student models using our proposed algorithm (Figure 3). The total emotion metrics for both the teachers and students in multiple generations are shown in Figure 4. Our first finding is that ensembles always outperform single models on the total emotion metric. As we increased the number of generations, the performance gap between the single models and ensembles became smaller. This is probably because the variability between models becomes smaller and smaller after more and more generations of self-distillation.

Our second finding on the results of EMENet-V is that the emotion performance does not increase monotonically as the number of generations increases. This is consistent with [28], where they interpreted increasing generations of

| Methods | AU1 | AU2 | AU4 | AU6 | AU7 | AU10 | AU12 | AU15 | AU23 | AU24 | AU25 | AU26 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tea | 0.407 | 0.348 | 0.424 | 0.435 | 0.477 | 0.437 | 0.380 | 0.389 | 0.464 | 0.459 | 0.491 | 0.442 | 0.430 |
| Stu1 | 0.366 | 0.326 | **0.354** | **0.387** | 0.471 | 0.427 | 0.379 | 0.322 | 0.388 | 0.328 | **0.482** | 0.413 | 0.387 |
| Stu2 | 0.358 | 0.319 | 0.355 | 0.388 | **0.468** | 0.424 | 0.383 | 0.314 | 0.367 | 0.303 | 0.487 | 0.400 | **0.381** |
| Stu3 | 0.351 | 0.312 | 0.369 | 0.395 | 0.479 | 0.435 | 0.400 | **0.310** | 0.353 | 0.287 | 0.488 | 0.398 | **0.381** |
| Stu4 | **0.338** | **0.301** | 0.377 | 0.401 | 0.482 | 0.440 | 0.421 | 0.322 | **0.347** | **0.271** | 0.490 | **0.386** | **0.381** |
| TS [12] | 0.405 | 0.345 | 0.420 | 0.435 | 0.476 | 0.435 | 0.379 | 0.388 | 0.462 | 0.449 | 0.491 | 0.441 | 0.427 |
| MC [11] | 0.408 | 0.350 | 0.455 | 0.429 | 0.472 | **0.422** | **0.362** | 0.357 | 0.402 | 0.499 | 0.485 | 0.414 | 0.421 |

Table 3: The NLL values for 12 action units, which are evaluated on the validation set of the Aff-wild set. We compare our single teacher models and single student models with other methods, *i.e.*, TS (temperature scaling [12]) and MC (Monte-Carol Dropout [11]). The model architecture used in this comparison is EMENet-V.

| | EXPR | Valence | Arousal |
|---|---|---|---|
| Methods | NLL | RMSE | RMSE |
| Tea | 1.052 | 0.400 | 0.256 |
| Stu1 | 0.911 | 0.379 | 0.234 |
| Stu2 | **0.905** | 0.377 | **0.231** |
| Stu3 | 0.918 | 0.373 | 0.232 |
| Stu4 | 0.957 | **0.370** | 0.233 |
| TS [12] | 0.998 | - | - |
| MC [11] | 1.071 | 0.398 | 0.251 |

Table 4: The NLL values for EXPR recognition and the RMSE values for valence and arousal prediction. Metrics are evaluated on the validation set. TS optimizes temperature for lower NLL on a held-out validation set, which is not beneficial for RMSE in regression tasks. Therefore, we only compare our models with TS for EXPR task.

self-distillation as amplifying regularization. The best number of generations for EMENet-V was three. More generations added too much regularization, leading to poorer performance on the validation set. We found the same phenomenon in the results of EMENet-VA. The best number of generations was also three. We evaluated our models on the test set of the Aff-wild2. The results are listed in Table 2. The visual-audio models always have better performance than visual models, although with slightly larger computation cost.

### 5.2. Uncertainty Performance

There are two types of uncertainty we are interest in: the aleatoric uncertainty and the epistemic uncertainty. The aleatoric uncertainty arises from the natural complexities of the underlying distribution, such as class overlap, label noise, input data noise, *etc*. The epistemic uncertainty arises from a lack of knowledge about the best model parameters. It can be explained away given enough data in that region. The latter uncertainty is an indicator of out-of-distribution (OOD) samples.

We capture the two types of uncertainty with deep ensembles. Following [9] and [26], we compute the two types of uncertainty as follows:

$$\underbrace{\mathcal{H}(E_{p(\theta|\mathcal{D})}\left[P(y|x,\theta)\right])}_{total\ uncertainty} = \underbrace{E_{p(\theta|\mathcal{D})}\left[\mathcal{H}(P(y|x,\theta)\right]}_{aleatoric\ uncertainty} + \underbrace{\mathcal{MI}\left[P(y),\theta|x,\mathcal{D}\right]}_{epistemic\ uncertainty}. \quad (14)$$

$P(y)$ is the probabilistic distribution output of a single model. $\mathcal{H}$ is the Shannon's entropy. $p(\theta|\mathcal{D})$ is the posterior distribution of the model weights $\theta$ which is trained on dataset $\mathcal{D}$.

**In-domain uncertainty**. The aleatoric uncertainty is an indicator of noise inherent in in-domain data. It is computed as the average entropy of single models' outputs in an ensemble, as shown in Equation 14. To evaluate the performance of in-domain uncertainty, we computed the average NLL (or RMSE) for single models to measure the similarity between predicted probability distributions and the true probability distributions.

To evaluate AU uncertainty estimation performance, we computed the NLL values on the AU validation set. Besides the NLL values of our single teacher models and single student models, the NLLs values of Temperature scaling (TS) [12] and Monte-Carol Dropout (MC Dropout) [11] were also computed. For a fair comparison, we adopted the test-time cross-validation in [1] to compute the NLL in TS. The optimal temperature was optimized on a randomly-split half of the validation set. The NLL was then evaluated on the other half of the validation set. For MC Dropout, we averaged the probability outputs of ten forward passes, where the model weights were sampled randomly for every forward pass from the dropout.

Table 3 shows the NLL values of single models, where the model architecture is EMENet-V. The validation performance for the EMENet-VA architecture is given in Appendix C. Table 3 shows that the single student models in later generations (*i.e.*, 2, 3 and 4) have better uncertainty estimation performance than TS and MC Dropout. When it comes to the averaged NLL values, our method outperforms

TS by $10.8\%$ and MC Dropout by $9.5\%$.

In Table 4, we list the uncertainty metrics for facial expressions (EXPR), valence and arousal. We find that the single student models have better uncertainty performance than TS and MC Dropout for both tasks. Our algorithm improves the EXPR NLL by $10.3\%$ when compared with TS and $15.5\%$ when compared with MC Dropout. As for valence/arousal RMSE scores, our method outperforms MC Dropout by $7.0\%$/$8.0\%$.

**Out-of-domain Uncertainty**. Epistemic uncertainty is an indicator of data insufficiency in certain regions of the input space. We expect that out-of-domain samples will be associated with large epistemic uncertainty. The epistemic uncertainty is computed by subtracting the aleatoric uncertainty from the total uncertainty.

To evaluate the performance of the epistemic uncertainty computed by our models for detecting out-of-domain samples, we chose the Fashion-MNIST [34] training set as OOD samples. It contains 60,000 gray-scale images with objects like bags, shirts, trousers, *etc*. The Aff-wild2 validation set has over 500,000 images. By mixing the Aff-wild2 validation set with the the Fashion-MNIST training set, we created an OOD detection task. We used the ensembles to compute the epistemic uncertainty of each sample. Since we performed multitask emotion prediction, the ensemble model produced an epistemic uncertainty value for each task. We averaged the epistemic uncertainties over all emotion tasks, and used a threshold to classify samples into in-domain and out-of-domain samples. We plotted the ROC curves showing the OOD performance across all thresholds. The ROC curves computed on the averaged epistemic uncertainty are shown in Figure 5. Each ROC curve corresponds to the ensemble in one generation. The AUC scores are also computed to show the OOD performance. Higher AUC scores indicate more accurate differentiation between in-domain and out-of-domain samples.

From Figure 5, we had the best AUC score for the student ensemble in the second generation. The AUC scores show a unimodal relationship with the number of generations. As we distill more and more, the uncertainty performance gradually increases to its best value, then decreases.

**Summary**. We find that self-distillation improves both types of uncertainty. The epistemic uncertainty produced by our ensembles can detect out-of-domain samples accurately. The best AUC score was $0.898$. When estimating in-domain uncertainty, ensembling is not required. If the computation resources are limited, it is acceptable to use the probability outputs of a single model to compute the entropy: as an approximation to the aleatoric uncertainty computed from an ensemble's output (Equation 14).

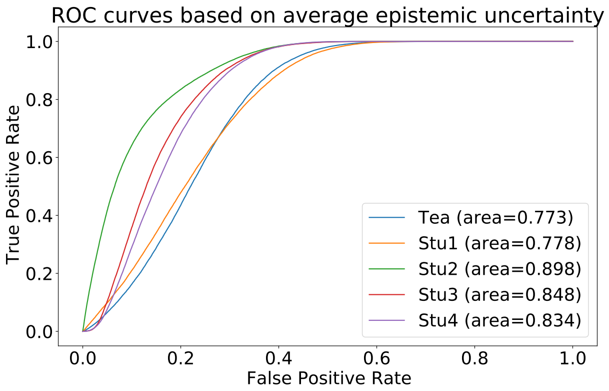

Figure 5: The ROC curves and AUC scores for the OOD detection task we created (Fashion-MNIST training set are OOD samples). Each curve corresponds to the ensemble in a certain generation. The model architecture for the results is EMENet-V.

## 6. Conclusions

In this paper, we propose to apply deep ensemble models learned by a multi-generational self-distillation algorithm to improve emotion uncertainty estimation. Our designed model architectures are efficient, and can be potentially applied in mobile devices. Our experimental results show that our algorithm can improve both the emotion metrics and uncertainty metrics as the number of generations increases. The uncertainty estimates given by our models are reliable indicators of in-domain and out-of-domain samples. In the future, we will study the regularization effect of the self-distillation algorithm, and seek better regularization methods to replace the time-consuming progress of self-distillation.

## 7. Acknowledgements

This work was supported in part by the Hong Kong Research Grants Council under grant number 16213617.

## References

[1] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*, 2020. 7

[2] CJ Brainerd. The emotional-ambiguity hypothesis: A large-scale test. *Psychological science*, 29(10):1706–1715, 2018. 2

[3] Cunjian Chen. Pytorch face landmark: A fast and accurate facial landmark detector. https://github.com/cunjian/pytorch_face_landmark, 2021. 3

[4] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verifica-

tion on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018. 3

[5] Ting Dang, Vidhyasaharan Sethu, Julien Epps, and Eliathamby Ambikairajah. An investigation of emotion prediction uncertainty using gaussian mixture regression. In *INTERSPEECH*, pages 1248–1252, 2017. 2

[6] Ting Dang, Brian Stasak, Zhaocheng Huang, Sadari Jayawardena, Mia Atcheson, Munawar Hayat, Phu Le, Vidhyasaharan Sethu, Roland Goecke, and Julien Epps. Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in avec 2017. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 27–35, 2017. 2

[7] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020. 2, 3

[8] Didan Deng, Zhaokang Chen, Yuqian Zhou, and Bertram Shi. Mimamo net: Integrating micro-and macro-motion for video emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2621–2628, 2020. 3

[9] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR, 2018. 7

[10] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018. 2, 5

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2, 7, 12

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 2, 7, 12

[13] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 890–897, 2017. 1, 2

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[15] Fei Jia, Somshubra Majumdar, and Boris Ginsburg. Marblenet: Deep 1d time-channel separable convolutional neural network for voice activity detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6818–6822. IEEE, 2021. 3

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[17] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. *arXiv preprint arXiv:2106.15318*, 2021. 2

[18] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020. 5

[19] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 3

[20] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 2

[21] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019. 2

[22] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 2, 5

[23] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 3

[24] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 2

[25] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016. 2, 5

[26] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019. 7

[27] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 1, 2

[28] Hossein Mobahi, Mehrdad Farajtabar, and Peter L Bartlett. Self-distillation amplifies regularization in hilbert space. *arXiv preprint arXiv:2002.05715*, 2020. 2, 5, 6

[29] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 1

[30] Emily Mower, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Interpreting ambiguous emotional expressions. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8. IEEE, 2009. 1

[31] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019. 2

[32] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019. 2

[33] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–8. IEEE, 2015. 1, 2

[34] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 8

[35] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 2

[36] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017. 1

[37] Zixing Zhang, Florian Eyben, Jun Deng, and Björn Schuller. An agreement and sparseness-based learning instance selection and its application to subjective speech phenomena. In *Proceedings of the 5th International Workshop on Emotion Social Signals, Sentiment & Linked Open Data (ES3LOD 2014), satellite of the 9th Language Resources and Evaluation Conference (LREC 2014)(B. Schuller, P. Buitelaar, L. Devillers, C. Pelachaud, T. Declerck, A. Batliner, P. Rosso, and S. Gaines, eds.), Reykjavik, Iceland*, 2014. 2

[38] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018. 1, 2

[39] Zhilu Zhang and Mert R Sabuncu. Self-distillation as instance-specific label smoothing. *arXiv preprint arXiv:2006.05065*, 2020. 2