Privacy Preserving Personalization for Video Facial Expression Recognition Using Federated Learning

Ali N. Salman
Ali.Salman@UTDallas.edu
University of Texas at Dallas
Dallas, Texas, USA

Carlos Busso
Busso@UTDallas.edu
University of Texas at Dallas
Dallas, Texas, USA

ABSTRACT

The increased ubiquitousness of small smart devices, such as cellphones, tablets, smart watches and laptops, has led to unique user data, which can be locally processed. The sensors (e.g., microphones and webcam) and improved hardware of the new devices have allowed running deep learning models that 20 years ago would have been exclusive to high-end expensive machines. In spite of this progress, state-of-the-art algorithms for *facial expression recognition* (FER) rely on architectures that cannot be implemented on these devices due to computational and memory constraints. Alternatives involving cloud-based solutions impose privacy barriers that prevent their adoption or user acceptance in wide range of applications. This paper proposes a lightweight model that can run in real-time for *image facial expression recognition* (IFER) and *video facial expression recognition* (VFER). The approach relies on a personalization mechanism locally implemented for each subject by fine-tuning a central VFER model with unlabeled videos from a target subject. We train the IFER model to generate pseudo labels and we select the videos with the highest confident predictions to be used for adaptation. The adaptation is performed by implementing a federated learning strategy where the weights of the local model are averaged and used by the central VFER model. We demonstrate that this approach can improve not only the performance on the edge device providing personalized models to the users, but also the central VFER model. We implement a federated learning strategy where the weights of the local models are averaged and used by the central VFER. Within corpus and cross-corpus evaluations on two emotional databases demonstrate that edge models adapted with our personalization strategy achieve up to 13.1% gains in F1-scores. Furthermore, the federated learning implementation improves the mean micro F1-score of the central VFER model by up to 3.4%. The proposed lightweight solution is ideal for interactive user interfaces that preserve the data of the users.

CCS CONCEPTS

• **Computing methodologies** → **Unsupervised learning**; **Computer vision tasks**.

ICMI, Nov 07–11, 2022, Bangalore, India

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3536221.3556614

KEYWORDS

facial expression recognition, adaptive facial expression recognition, federated learning,

ACM Reference Format:
Ali N. Salman and Carlos Busso. 2022. Privacy Preserving Personalization for Video Facial Expression Recognition Using Federated Learning. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3536221.3556614

1 INTRODUCTION

One fascinating fact about emotions is that they play an important role in our decision making process [26]. Emotions guide us to survive and thrive, playing a fundamental role in daily human interaction. There are multiple emotions, and each serves a specific purpose. For instance, the basic emotions that we mainly deal with in every day are happiness, sadness, anger, surprise, disgust, and fear [10]. These emotions help us to shape our daily interactions by dictating the expressed behaviors and how others respond to us. We can augment interactive systems with capabilities to create seamless experiences in *human-computer interaction* (HCI) by understanding emotions and learning how to detect them using automatic methods. These emotion-aware systems can have wide range of applications in areas such as healthcare, security and defense, education, and entertainment.

Emotions are expressed through speech [1, 25], facial expressions [6, 11, 18, 22, 28, 30, 37], lexical content [32], posture [8] and physiological signals [11, 22]. Therefore, it is not surprising the advances in algorithms to recognize emotions using these modalities. Furthermore, recent hardware development has allowed us to deploy emotion recognition systems in real-time on edge devices such as smartphones, tablets and portable computers using embedded sensors of the devices such as microphones and cameras. With the ubiquitousness of cameras, *facial expression recognition* (FER) has become an important modality for HCI. There are important barriers for FER technology that need to be carefully considered before these technologies can be massively deployed. State-of-the-art FER systems rely on complex networks that impose computational and memory resources [38, 39]. These systems are not ideal for edge devices. A straightforward solution is to implement cloud-based solutions where images or videos are uploaded to a server that has the resources to run these models. However, this solution has usability problems where users may not be willing to share private sensitive data [17]. These barriers need to be addressed before multimodal interactive systems can safely and effectively use FER algorithms in edge devices without the need for users to share sensitive information.

This study proposes an effective solution to build lightweight models for FER that preserves the privacy of the users, while improving the performance by personalizing the edge models. The proposed approach starts with a central lightweight model that can be easily implemented on edge devices. The model uses the MobileNetV2 framework [31] as the core architecture, which is significantly smaller than other alternative networks such as the VGG-Face network [24]. Then, the central model is locally personalized at the edge level by adapting the model using unlabeled videos from the target subject. We only use videos predicted with high confidence to create pseudo labels using the prediction results. An important novelty in our study is that the personalized models in the edge devices are also used to improve the central model, generating an effective federated learning strategy. Instead of sharing images or videos from the target users, our approach shares the weights of the local models, which are averaged and used by the central model. This approach increases the accuracy by adapting an edge model to a target subject, without the need for hand labeled data. The adaptation can be locally implemented on the edge device without the need to send sensitive data to a server which would compromise the user's privacy.

We evaluate the performance of our approach by adapting each model to a single subject, and reporting the aggregate results for all subjects in the test set. The within-corpus evaluations show that the performance increases by up to 6% in micro F1-score. Furthermore, we evaluate the proposed adaptation strategy using cross-corpus experiments. We observe gains in macro F1-score up to 13.1%. These results highlight the effectiveness of our unsupervised adaptation approach in personalizing the models to a specific subject. The proposed model only has about 3.1M parameters, from which we only adapt 208K parameters. This configuration allows the proposed algorithm to be efficiently implemented on edge devices. The federated learning results also show that updating the central model leads to improvements over the original central model with up to 3.4% absolute gains in macro F1-score. The results suggest that the proposed solution can handle over-time domain shifts. The contributions of this study are:

- We provide a lightweight *video facial expression recognition* (VFER) method that is specifically designed to be implemented on local devices with minimum computational and memory constrains.
- The local models are adapted with unlabeled local data with a simple approach that does not impose much computational resource on the edge device.
- By using a federated learning strategy, we update the central model while preserving the privacy of the user, providing an effective solution to handle over time drifts in the data distribution in the target domain.

These contributions are crucial to deploy FER systems in interactive multimodal systems.

2 RELATED WORK

Many studies have reported great advancement in FER by using motion captured data [3], gray scale static images [15], and video sequences [30]. However, not many studies have focused on deploying efficient models that can run on small devices such as smartphones. Song et al. [35] developed a static FER system that works by capturing a photo using the built-in smartphone camera, which is sent to a server that predicts the underlying expression of the image. Song et al. [35] used the approach presented by Ojala et al. [23], consisting of a *support vector machine* (SVM) model trained on local binary patterns extracted from a gray-scale facial image. A limitation of this approach is that sending images to a server requires a fast network connection, which is not always available. It also raises privacy concerns, as sensitive data from the user is shared with the cloud-based server. Another limitation of *image facial expression recognition* (IFER) is that it can only predict posed expressions, since the focus is on static images and not dynamic series of images. Even if we aggregate the predictions of multiple images in a video, the aggregated predictions will be inferior than using models trained on dynamic data. Salman and Busso [29] showed that human emotional perception of static isolated facial images does not provide a good description of the emotional perception observed after watching the entire video. This study showed the importance of modeling the dynamic facial movements to obtain good VFER.

When the focus of *deep neural networks* (DNNs) is maximizing performance, the resulting architectures tend to be quite complex. However, several applications require more compact architectures. Efficient architecture based on *convolutional neural networks* (CNNs) have been proposed for visual tasks. These networks include the MobileNet [14] and EfficientNet [36], and different implementations of these networks containing different depths. These networks often have less than 10M parameters and are purposely made to run on low-powered devices. For comparison, the VGG16 [34] and ResNet50 [13] architectures contain 138M and 25M parameters, respectively. Efficient compact networks are particularly important for FER systems working on edge devices which have limited computational and memory resources [27, 33]

This study relies on federated learning to avoid sending sensitive information from users to update the FER model. Federated learning is an approach proposed by McMahan et al. [20]. The training occurs on the clients, where local data is stored. Instead of sharing the user's data, the approach shares the weights or the gradients of the individually trained models, which are used to update the central model. This process improves the performance of the initial shared central model without the need of the local data from the users, preserving the privacy of the clients. Multiple methods of federated learning have been proposed. McMahan et al. [20] proposed the *federated stochastic gradient descent* (FedSGD) and the *federated averaging* (FedAvg) methods, which are the most popular algorithms. FedSGD sends the gradients of the local models back to the central model. This model requires greater communication between the local models and central model compared to the FedAVG approach. FedAVG averages the weights of some or all the models on the edge after training them with local data.

Chhikara et al. [7] used a federated learning approach to train local and central models using images and audio signals to predict the underlying expression and monitor the mental health status of the users. They use two local models, an audio and a visual model. The audio model extracts features from the audio signal such as MFCCs and Mel-spectrogram, and a CNN model process and predict the emotion. The visual model uses a CNN and a SVM classifier to

detect the emotion from 48x48 gray-scale images. The emotion prediction is made on a time window, which is then equally weighted to give a representation to a sequence for the visual and audio signals. If one modality is not present, the prediction is weighted solely on the present modality. To the best of our knowledge, the closest paper to our methodology was presented by Shome and Kar [33], which trained two models to predict facial expressions in static images using the ResNet50V2 architecture, which contains over 25M parameters. The first model is a representation learner, trained in a self-supervised fashion to extract relevant facial features, while the second model is a few-shot learning model, trained in a self-supervised fashion with a few labeled samples from the client data. Both models are trained on the edge, updating the central models using the FedAVG algorithm. Our study is different from this work since (1) we predict static facial expressions from images and dynamic facial expression from videos, (2) we pre-train the central model on labeled samples, (3) we use a fully unsupervised approach to update the client, and (4) we use less than a fifth of the parameters used on their proposed approach to train both a static and dynamic FER model.

3 PROPOSED APPROACH

We propose an efficient algorithm to personalize a VFER model for each user by relying on an unsupervised adaptation approach. Figure 1 describes our proposed privacy-preserved formulation, which considers a central model and several client models that have available unlabeled local data. The model uses a lightweight image-based network (Sec. 3.1) that is used to build our proposed VFER (Sec. 3.2). The adaptation approach to personalize our VFER model leverages predictions from the subject with high reliability (Sec. 3.3). The weights of the adapted local models are shared to update the central model (Sec. 3.4). This section describes these blocks in our framework.

3.1 Image Facial Expression Recognition

In this study, we propose an end-to-end system that can be adapted to each subject to increase the accuracy of the system. The first building block in our proposed system is the IFER model, which relies on a CNN-based architecture to extract high level facial features. Our system is built with the purpose of being deployed in real-time on edge devices with limited computational resources. We achieve this objective by using a small and efficient model as our CNN backend. While there are several lightweight CNN-based models, we implement our IFER model with the MobileNetV2 architecture [31], which has just over 2 million parameters (CNN only), a size of under 14MB, and a depth of 105 layers. We represent this network with the function $f(x_i)$, where x_i is the input image. MobileNetV2 is suitable for our application since it can run on low-powered devices such as smartphones.

We use a pre-trained version of the MobileNetV2 trained on the ImageNet dataset [9], which allows us to quickly adapt the model for our task. We adapt the MobilNetV2 model using the ImageNet weights for the CNN portion of the model $f(x)$, adding a global average pooling layer followed by a fully connected layer and a softmax layer (Fig. 2(a)). Then, we train the IFER model on the AffectNet dataset (Sec. 4.1), while freezing the weights of the

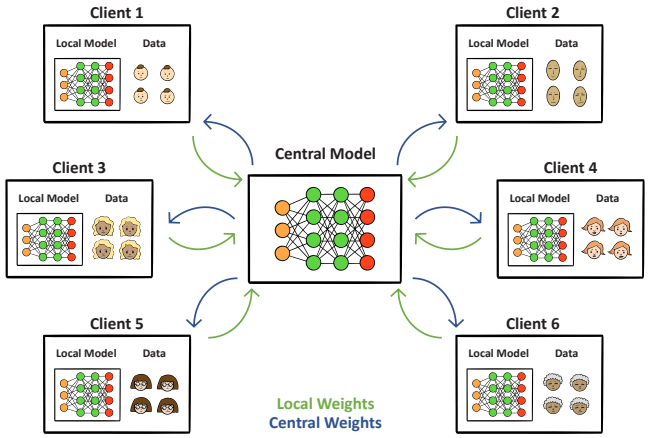
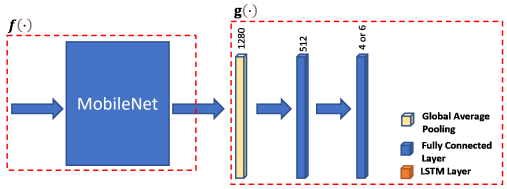
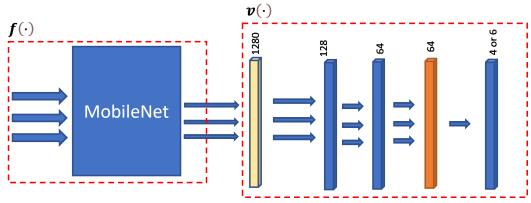


Figure 1: Proposed federated learning formulation for our privacy preserving personalization approach for VFER. The client models update the central model using local unsupervised data. The central model is updated ($v(\cdot)$ network in Fig. 2) without the need of sharing images or videos from the clients.



(a) IFER Model



(b) VFER Model

Figure 2: Diagram of the proposed FER system. The model consists of two parts (a) the image predictor model, which predicts the facial expression for a single face/image, and (b) a video predictor, which aggregates the high-level features extracted from the image predictor to recognize emotions in the video.

CNN blocks. Once the IFER model has been trained, we are able to predict the facial expression present in an image. We denote the IFER block as $IFER(x_i) = g(f(x_i))$, where x_i is the input image of the ith frame and $g(\cdot)$ is the prediction head containing the fully connected layers and the softmax layer.

3.2 Video Facial Expression Recognition

Figure 2(b) shows the architecture of the VFER model. The IFER model is able to extract high-level features from the input images. By using the global average pooling output of MobileNetV2 ($f(\cdot)$), we train a facial expression predictor that takes a sequence of images $(f(x_1), f(x_2), f(x_3), \ldots)$. The video predictor model consists of two fully connected layers implemented with 128 and 64 neurons, respectively. Then, it adds a *long short-term memory* (LSTM) layer, implemented with 64 neurons, and a softmax layer for the predictions. The fully connected layers reduce the dimensionality of the extracted features from 1,280 to 64 per frame. This reduction enables us to use a relatively small LSTM layer to capture temporal information across frames. The LSTM layer additionally reduces the time dimensionality into one time instance of 64 neurons, since we use the last frame as the representative vector of the entire video sequence. The softmax layer is then able to estimate the probabilities for each emotional class. By using this approach, we are able to create a VFER model on top of the IFER model by adding just over 211k parameters, from which 208K parameters are trainable. This configuration only represents a 10% increase in the number of parameters. We denote the VFER block as $VFER(x) = v(f(x_1), f(x_2), f(x_3), \ldots)$, where $x = \{x_1, x_2, \ldots\}$ is the image sequence and $v(\cdot)$ is the prediction head containing a softmax layer to provide prediction on a given image sequence.

3.3 Unsupervised Personalization Strategy

An essential block in our formulation is to adapt the VFER model on the unlabeled local data of each client. We propose an adaptation approach that adapts the VFER model for a single subject by using the output predictions of the IFER model. Because the IFER model is trained on a large number of subjects with more demographic variability, we expect it to be more consistent for samples belonging to subjects not included in the train set than the VFER model, which was trained on a limited number of subject (76 for CREMA-D corpus, and 6 for MSP-IMPROV corpus). First, the central model is shared with each client. The adaptation approach works on a per subject basis and starts by taking into consideration the image level prediction $IFER(x_i) \in \mathbb{R}^{1\times K}$, where K is the number of classes. We process each frame in the video sequence with the IFER model. We define pseudo labels for a video by aggregating the results for the entire sequence. Equation 1 defines the aggregation method for a sequence of length N, which just consists of the average of the probabilities $IFER(\cdot)$ for all the images in the sequence.

$$\widehat{VFER}(x) = \frac{1}{N}\sum_{i=1}^{N} IFER(x_i) \tag{1}$$

The vector $\widehat{VFER}(x) \in \mathbb{R}^{1\times K}$ provides a simple but effective metric to aggregate the temporal predictions across frames into a single vector for a video. For sentences with clear emotional content, we expect that the probability for the correct class will be higher than the probabilities for other classes. Therefore, the values in $\widehat{VFER}(x)$ can be used as a proxy to define the confidence of the model. Our approach assumes that each client has unlabeled local data. We select a subset from the unlabeled data to adapt the model. First, we assign a pseudo label to each unlabeled video by selecting the emotional class with the highest confidence in $\widehat{VFER}(x)$. Then, we limit the number of samples selected per emotional class to at most P samples for each of the K classes, since emotional classes do not necessarily appear in a balanced distribution in the wild. The P samples are the videos with the highest probability per class. Then, we discard videos if the probability of the class with the highest value is lower than a given threshold $p_{\text{threshold}}$. Therefore, the number of videos to adapt the VFER model can be lower than $K \times P$. This restriction aims to only select samples with the highest confidence while still maintaining a balanced distribution across classes.

With the selected unlabeled samples and their pseudo labels, we fine-tune the VFER model by adapting only the $g(\cdot)$ classifier, while freezing the $f(\cdot)$ classifier. Restricting the number of parameters to be adapted is necessary, since we adapt the local VFER model with few unlabeled samples on the edge device.

3.4 Federated Learning Strategy to Update the Central Model

The updated local models can be used to update the central model. This is a particularly important feature of our model from a usability perspective, since it helps the system to deal with a shift in input distributions over time. To preserve the privacy of the users, we do not send the local data. Instead, the local models share their weights with the central model, as described in Figure 1. This is an iterative strategy, corresponding to the FedAVG approach [20]. An iteration starts by deploying the initial central model to the clients. Then, the client models are updated using the strategy described in Section 3.3. The central model is then updated with the shared weights of the local models. We only update the weights of the network $v(\cdot)$ (Fig. 2). The parameters of $f(\cdot)$ are frozen, dramatically reducing the training time, as less than 10% of the parameters have to be updated. This concludes a single round of FedAVG. The central model is then sent to the clients and the training process can be repeated as many times as necessary. However, we only repeat this process one more time for a total of two rounds to keep a low computational overhead on the client.

4 EXPERIMENTAL EVALUATION

4.1 Databases

This study uses the AffectNet [21], CREMA-D [5], and MSP-IMPROV [4] datasets. The images of the AffectNet database are used to train the IFER model (Sec. 5.1) and detect pseudo labels (Sec. 3.3). The videos from the CREMA-D and MSP-IMPROV corpora are used for within-corpus and cross-corpus experiments. This section describes these corpora.

AffectNet: The AffectNet corpus [21] contains over 1 million of emotional images collected from different search engines using multiple keywords. Expert evaluators annotated over 440 thousand images with eleven discrete labels (i.e., happiness, sadness, anger, etc), and the emotional attributes for valence (negative versus positive) and arousal (calm versus active). We use a subset of the AffectNet dataset containing either the classes neutral, happiness, sadness, anger (four classes), or these emotions plus fear and disgust (six emotions). The training set of the AffectNet dataset is

down-sampled to 24,882 images per class, thus reducing data imbalance across emotional classes. From this set, we use 80% of samples for each class as our training set and 20% as our development set. Since the labels of the test set provided by the AffectNet dataset have not been released, we use the development set provided by the AffectNet dataset as our test set. This set has 500 images for each class. Therefore, we have a total of 2,000 images for the 4-class problem and 3,000 images for the 6-class problem.

CREMA-D: The CREMA-D corpus [5] is an audio-visual emotional corpus consisting of 91 actors between the age of 20 and 74 (48 actors are male and 43 actors are female). The dataset contains 7,442 video segments with a resolution of 960 × 720 pixels. The videos were manually annotated using a crowdsoucing protocol. There are 12 sentences for each actor for each six primary emotions (neutral, happiness, sadness, anger, fear, and disgust) and four different intensities (low, medium, high, and unspecified). The annotations include three conditions: audio-only without the video, video-only without the audio, and audiovisual. The number of raters per sample ranges from 4 to 12, where over 95% of the samples have eight or more annotations. This study uses the consensus labels obtained by the plurality rule from the audiovisual annotations. We use the CREMA-D dataset to train and test a four class VFER model and a six class VFER model (Sec. 5.2). We use data from 76 subjects for the train set, data from 4 subjects for the development set, and data from 31 subjects for the test set, keeping each set as gender balanced as possible. Additionally, we consider the whole video sequence while training and testing our model using zero pre-padding.

MSP-IMPROV: The MSP-IMPROV corpus [4] is a multimodal corpus of dyadic conversations between 12 subjects (six male and six female). The resolution of the videos is 1,440 × 1,080 pixels. The corpus was designed to elicit 20 target sentences expressing four target emotions: happiness, sadness, anger and neutral state. This goal was achieved by defining hypothetical emotionally dependent scenarios that led one of the actors to utter the target sentence in a specific emotion. The corpus includes not only the target sentences, but also the rest of the recordings in the improvised dialogs. It also includes naturalistic recordings collected during the breaks between the improvisations, and read renditions of the target sentences. Overall, the corpus has 8,428 speaking turns. The segments are annotated using a crowdsourcing protocol described in Burmania et al. [2], which tracks in real-time the quality of the workers, stopping the perceptual evaluation when the quality drops below an acceptable level. Each video segment is annotated by at least five workers, who evaluated the primary emotion, secondary emotion, and three emotional attributes (valence, arousal and dominance). For our experiments, we only consider the plurality rule for the primary emotions. We consider the classes happiness, sadness, anger and neutral state (four class problem). We use data from six actors for the train set, data from 2 actors for the development set, and data from 4 subjects for the test set. These partitions are gender balanced. We also consider the whole video sequence while training and testing our models using zero pre-padding.

4.2 Implementation

We first train the IFER model on AffectNet. We train two models since the classification problems for the MSP-IMPROV corpus (four-class problem) and CREMA-D corpus (six-class problem) are different. For the MSP-IMPROV corpus, we train the IFER model with four emotional classes (happiness, sadness, anger and neutral state). For the CREMA-D corpus, we train the IFER model with six emotional classes (happiness, sadness, anger, fear, disgust and neutral state). We use the MobileNetV2 network up to the global average pooling layer with pre-trained weights using the ImageNet network. We freeze those weights, training the network $g(\cdot)$ for 60 epochs.

Once the IFER has been trained, we freeze the weights of the IFER model. We use the video datasets (CREMA-D or MSP-IMPROV corpora) to train the VFER model using the corresponding IFER model. We train both models for 20 epochs. The model converges fairly quickly, since the VFER model has just over 200k parameters, the $f(\cdot)$ model is frozen, and the video datasets are less diverse compared to the AffectNet database. Once the VFER model is trained, we generate the pseudo labels $\widetilde{VFER}(x)$ for each subject. We set $p_{\text{threshold}} = 0.5$, making sure that the selected class is the dominant class with more than 50% probability. The MSP-IMPROV database contains fewer subjects, but higher number of samples per subject. The CREMA-D database contains more subjects, but fewer samples per subject. Therefore, the value for the parameter P is independently set for each corpora, using $P = 50$ for the MSP-IMPROV database and $P = 15$ for the CREMA-D database.

We adapt the subject specific model using the selected unlabeled data with the corresponding pseudo labels (i.e., at most $P \times K$ samples). Initially, we adapted the model with a learning rate of 0.0001 for 5 to 10 epochs. However, we noticed similar results adapting the model with a higher learning rate of 0.001 for only two epochs. Therefore, we use this setting to reduce the required computations, which is important since this adaptation needs to be implemented on the edge device.

For all the experiments, we use the facial region extracted by using the MediaPipe [19] toolkit, which is a cross-platform framework that can be used in real-time on smartphones and other machines. Then, the facial region is resized to 224x224x3. We train using the ADAM optimizer [16], with a learning rate of 0.001 and a 0.5 multiplier on the development set plateau. Also, we use weighted categorical cross-entropy back-propagation while training on the AffectNet, MSP-IMPROV, and CREMA-D corpora to mitigate data imbalance. Regular cross-entropy (non-weighted) is used during adaptation.

5 EXPERIMENTAL RESULTS

5.1 Image Facial Expression Recognition

We need the IFER model to train the VFER model and generate the pseudo labels. We train the model on the AffectNet corpus. Since we use the IFER model to generate the pseudo labels, the emotional classes for IFER, and VFER must be identical. Therefore, we train the model on a four class problem (MSP-IMPROV corpus) and a six class problem (CREMA-D corpus). Table 1 shows the classification results for both models on our test set generated from the AffectNet corpus (Sec. 4.1). For the four class problem (neutral, happiness, sadness, and anger) the model predicts happiness with the highest probability (85.6%) and neutral state with the lowest probability (59.9%). Sadness (64.7%) and anger (64.4%) have similar

Table 1: Within-corpus performance of the IFER model on the AffectNet corpus on the test set described in Section 4.1. The approach is implemented with the MobileNet2 network trained on a four-class problem (happiness, sadness, anger, and neutral state) and on a six-class problem (happiness, sadness, anger, fear, disgust, and neutral state).

Emotion	Precision [%]	Recall [%]	F1-score [%]	Precision [%]	Recall [%]	F1-score [%]
Model	4 Class			6 Class		
Happiness	83.6	87.6	85.6	76.0	86.8	81.0
Anger	65.4	63.4	64.4	51.3	52.2	51.7
Sadness	67.9	61.8	64.7	53.1	53.2	53.1
Neutral	57.8	62.0	59.9	48.7	58.4	53.1
Fear	–	–	–	71.3	64.6	67.8
Disgust	–	–	–	60.4	44.3	51.1
Average	68.7	68.7	68.6	60.1	60.0	59.7

F1-scores. However, sadness shows a higher precision score, while anger shows a higher recall score, indicating that anger tend to have more false positive predictions, while sadness tends to have more false negative predictions. For the six class problem (adding fear and disgust), the model behaves slightly differently by having anger (51.7%) and disgust (51.1%) with a lower F1-score. Happiness remains the class with the highest F1-score of 81.0%. Overall, the four class model has an average F1-score of 68.6%, while the six class model has an average F1-score of 59.7%. Additionally, the average precision and recall for both models remain almost identical, so on average the models have the same number of type 1 and type 2 errors. The results show that the models are effective in extracting discriminative emotional information from static images.

5.2 Within Corpus Evaluation of Adaptation Approach

This section presents the results of the VFER model with and without the proposed personalization method. Table 2 shows the results for the CREMA-D dataset on the test set, using the macro and micro means for the precision, recall, and F1-score. The macro results equally weigh each class, while and micro results equally weigh each sample. These metrics are important because the test set is unbalanced. The adapted models achieve a higher F1-score for all the emotional classes, with the exception of sadness, where the F1-score drops in 1.8%. The highest improvement is for neutral state that increases its F1-score in 9.1%. Overall, the adapted models have almost 3% improvements in macro F1-score and 5% improvement in micro F1-score compared to the models before the adaptation. Table 3 shows the same experiment on the MSP-IMPROV corpus. The adaptation is particularly beneficial for neutral state (11.5%) and sadness (3.6%). Even though the proposed adaptation does not work as well for anger (drop of 6.2%), the adapted model has an overall 2.3% increase in macro F1-score and 6% increase in micro F1-score across emotions. The results reveals that by using the pseudo labels obtained from the IFER model, we are able to adapt the model to better predict the underlying emotion of each subject.

Figure 3 separately shows the differences in Micro f1-score before and after the adaptation for each subject. Figure 3(a) shows the difference for the subjects in the CREMA-D corpus. Out of the 31 subjects used for the testing set, 23 of them saw an increase of micro F1-score ranging from 1.22% up to 9.8% with a mean increase of 4.6%. Four subjects showed no change in performance, and only 4 subjects saw a decrease in micro F1-score. Figure 3(b) shows the results for the subjects in the MSP-IMPROV corpus. There are only four subjects in the testing set. The proposed adaptation approach improves the F1-score for three of them (15.7%, 7.9%, and 3.1%, respectively). Overall, we see that for both corpora, the performance for the majority of the subjects increased with our unsupervised adaptation method.

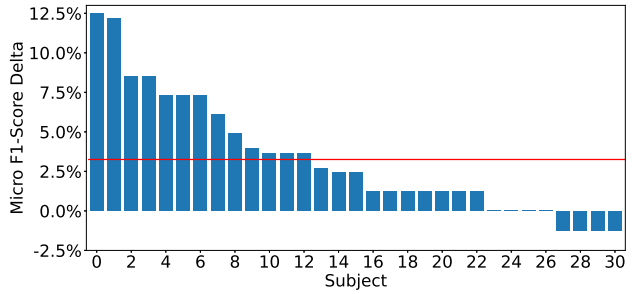
To the best of our knowledge, there is no direct comparison between our approach and others methods, because (1) our approach uses a light-weight model, (2) there is no standard training/development/test split in the emotional corpora used in this study, and (3) most approaches have explored either audio-only models, or audiovisual models. For example, the AuxFormer framework [12] relies on a large transformer model. The evaluation relies on a different data partition than our approach (70% for the train set, 15% for the development, and 15% for the test set). Even with these differences, this model reports a micro F1-score of 53% on the CREMA-D corpus, and a micro F1-score of 72% on the MSP-IMPROV corpus for video-only models. Even though our approach uses a

Table 2: Within-corpus performance of the VFER model on the CREMA-D corpus for the six-class problem. The reported metrics are based on the test set which consists of 31 subjects. The table reports the results before and after the proposed personalization using the local data from the target subject.

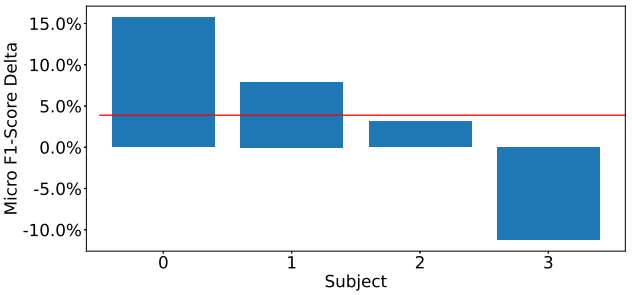
Emotion	Precision [%]		Recall [%]		F1-score [%]	
	Before	After	Before	After	Before	After
Happiness	77.5	82.0	93.7	93.5	84.8	87.4
Anger	45.5	59.5	60.4	48.9	51.9	53.7
Sadness	34.3	38.1	29.6	24.8	31.8	30.0
Neutral	67.0	62.4	51.6	73.4	58.3	67.4
Fear	47.0	49.1	46.4	48.4	46.7	48.7
Disgust	63.3	70.1	61.9	62.2	62.6	65.9
Macro mean	55.8	60.2	57.3	58.5	56.0	58.8
Micro mean	61.2	66.3	59.0	63.6	59.5	64.5

Table 3: Within-corpus performance of the VFER model on the MSP-IMPROV corpus for the four-class problem. The reported metrics are based on the test set which consists of 4 subjects. The table reports the results before and after the proposed personalization using the local data from the target subject.

Emotion	Precision [%]		Recall [%]		F1-score [%]	
	Before	After	Before	After	Before	After
Happiness	77.6	79.5	83.8	81.6	80.6	80.5
Anger	15.2	11.8	23.0	12.4	18.3	12.1
Sadness	20.5	25.6	60.2	51.9	30.6	34.2
Neutral	80.6	69.5	34.0	51.9	47.9	59.4
Macro Mean	48.5	46.6	50.2	49.4	44.3	46.6
Micro Mean	60.2	59.9	54.4	59.8	52.6	58.6



(a) CREMA-D



(b) MSP-IMPROV

Figure 3: Differences in the micro F1-score per subject in the testing set before and after implementing the proposed personalization approach.The horizontal red line depicts the mean difference.

light-weight model without the need for a large transformer-based architecture, our approach outperforms the AuxFormer framework by 11.5% on the CREMA-D corpus (64.5% versus 53%), but it is outperformed by 13.4% on the MSP-IMPROV corpus (58.6% versus 72%).

5.3 Cross Corpus Evaluation of Adaptation Approach

The results presented in Section 5.2 correspond to models trained and evaluated on the same corpus. However, a model deployed in real-world applications will face data that is different from the data used to train the models. To simulate this scenario, we conduct a cross-corpus evaluation where the models are trained on one corpus and tested on the other. Because the number of classes is different for the datasets, we only consider the shared classes (i.e., happiness, anger, sadness, neutral). For this evaluation, we train a separate model for the CREAM-D corpus with four emotional classes.

Table 4 shows the performance of the proposed approach trained on the MSP-IMPROV corpus and tested on the CREMA-D corpus. For consistency, we only report the results on the test set of the CREMA-D (data from 31 speakers). The table displays the results before and after adapting the model on the CREMA-D corpus using our unsupervised approach. Before adapting the model, we achieve an average macro F1-score of 42.8% and an average micro F1-score of 49.3%. After adapting the model, the macro and micro F1-scores increase by 13.1% and 9.3%, respectively. Also, we observe an increase in the F1-score for all the emotional classes. Neutral state presents the highest improvement with almost a 20% improvement.

Table 4: Cross-corpus evaluation, training the models on the MSP-IMPROV corpus, and evaluating the results on the CREMA-D corpus. The adaptations is conducted on the test set of the CREMA-D corpus (31 subjects).

Emotion	Precision [%]		Recall [%]		F1-score [%]	
	Before	After	Before	After	Before	After
Happiness	62.6	73.9	91.2	92.7	74.2	82.3
Anger	28.0	40.4	55.8	52.9	37.3	45.8
Sadness	25.7	34.9	27.6	53.3	26.6	42.2
Neutral	77.5	76.6	21.1	41.1	33.2	53.5
Macro mean	48.4	56.5	48.9	60.0	42.8	55.9
Micro mean	60.9	63.3	46.4	58.3	49.3	58.6

Table 5: Cross-corpus evaluation, training the models on the CREMA-D corpus, and evaluating the results on the MSP-IMPROV corpus. The adaptations is conducted on the test set of the MSP-IMPROV corpus (4 subjects).

Emotion	Precision [%]		Recall [%]		F1-score [%]	
	Before	After	Before	After	Before	After
Happiness	87.9	89.0	66.6	62.9	75.8	73.7
Anger	14.6	10.0	56.5	24.4	23.2	14.2
Sadness	30.7	30.4	17.4	32.4	22.2	31.3
Neutral	67.5	66.1	51.4	63.9	58.4	65.0
Macro mean	50.2	48.9	48.0	45.9	44.9	46.1
Micro mean	55.5	52.5	54.3	57.3	50.1	53.6

Table 5 shows the performance of our approach trained on the CREAM-D corpus, and tested on the MSP-IMPROV corpus (four subjects). Overall, the model increases its macro and micro F1-scores by 1.2% and 3.5%, respectively. The patterns vary across emotions, observing gains for sadness and neutral state and drop in performance for happiness and anger.

5.4 Federated learning VFER

The last step in our formulation is to adapt the central model with information provided by the local models. The objective of this step is not to replace the local models, which are already personalized to the particular subjects. Instead, the objective is to compensate for data shifts in the target domain over time, a common problem when deploying systems on real-world applications. The personalized method results on separate fine-tuned models (one for each subject), which are used to update the central model using the federated learning method described in Section 3.4.

Table 6 shows the experimental results after finishing two rounds of FedAVG for both corpora. For comparison, the results of the original central models are shown in Table 2 for the CREMA-D corpus and Table 3 for the MSP-IMPROV corpus (column “Before”

Table 6: Federated learning evaluation where we update the central model using the FedAVG approach. The table lists the results on the test sets of the MSP-IMPROV and CREMA-D databases.

Emotion	Precision [%]	Recall [%]	F1-score [%]	Precision [%]	Recall [%]	F1-score [%]
Dataset	MSP-IMPROV			CREMA-D		
Happiness	80.2	79.7	79.9	79.9	92.5	85.7
Anger	12.2	12.9	12.6	59.5	50.5	54.7
Sadness	22.7	58.5	32.7	33.2	31.1	32.1
Neutral	73.2	48.0	58.0	63.1	62.2	62.7
Fear	–	–	–	45.2	47.9	46.5
Disgust	–	–	–	62.3	60.2	61.2
Macro mean	47.1	49.8	45.8	57.2	57.4	57.1
Micro mean	59.4	58.0	56.0	61.7	60.5	60.9

in both tables). For the CREMA-D corpus, the model after two rounds of FedAVG increases the overall macro F1-score by 1.1% and the overall micro F1-score by 1.4%. The local models that are personalized to the target users still obtain better performance than the updated central model, as expected (column "After" in Table 2). The updated central model using federated learning leads to improvements in F1-score for happiness, anger, sadness and neutral state. The performance slightly decreases for fear (0.2%) and disgust (1.4%). For the MSP-IMPROV corpus, Table 6 shows similar trends. The performance increases the overall micro F1-score by 3.4%, where neutral state presents the highest gain (approximately 10%). We observe a slight drop in performance for the other classes.

In summary, the results from our federated learning approach show better performance than the original central model, but worse performance than the local personalized models. Since we use the same learning rate and parameters (P and $p_{\text{threshold}}$) for the FedAVG rounds as previous experiments, further improvement can be obtained by fine-tuning these parameters.

5.5 Model Size Analysis

Our model contains three networks implemented with 2,223,872 ($f(\cdot)$), 662,534 ($g(\cdot)$), and 208,582 ($v(\cdot)$) trainable parameters (Fig. 2). The central model contains slightly over 3.1M parameters. The whole model is initially shared with the clients. We only share the parameters of the model $v(\cdot)$ from the client to the server during the federated learning stage. This strategy has two advantages. First, it reduces the bandwidth from sharing the 3.1M parameters between the local and central models to only 208K parameters. Second, it reduces the complexity of locally adapting the model. Once the local models have been adapted, the $g(\cdot)$ model can be discarded, as our goal is VFER. Therefore, it further reduces the run time of our model.

6 CONCLUSIONS

This study proposed a novel and effective deep learning method to predict facial expression from images and videos using a lightweight model that can be implemented on edge devices. The proposed model leverages the MobileNetV2 architecture, adding only 10% extra parameters, providing a model for both IFER and VFER. The approach relies on personalizing a central VFER model using unlabeled local data, where data predicted with high confidence is used to generate pseudo labels. The adaptation of the model only affect a reduced number of parameters over two epochs, which can be implemented without imposing computational barriers on local devices. Using a federated learning paradigm, our formulation adapts the central model using the weights shared by the local models. This approach is attractive since it preserves user information by not sharing images or videos from the clients. It is also a practical solution, allowing the central model to deal with over time drifts in the data distribution of the target domain. Within and cross corpus experiments demonstrated that our personalization approach on the local models leads to absolute improvements as high as 13.1% in F1-scores. Evaluations with the federated learning approach demonstrated improvements in the central model over the original model without transferring local images or videos to the server.

In the future, we would like to explore other methods of creating pseudo labels to increase the model's accuracy, leveraging alternative unsupervised strategies. We would also explore longitudinal evaluations where we experience drifts in the data distribution in the target domain to fully assess the need to update the central model over time using federated learning. We will also study in more detail the tradeoff between reducing the number of parameters to be updated and achieving good performance.

ACKNOWLEDGMENTS

This study was funded by the National Science Foundation (NSF) award IIS-1718944.

REFERENCES

[1] M. Abdelwahab and C. Busso. 2018. Study Of Dense Network Approaches For Speech Emotion Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*. IEEE, Calgary, AB, Canada, 5084–5088. https://doi.org/10.1109/ICASSP.2018.8461866

[2] A. Burmania, S. Parthasarathy, and C. Busso. 2016. Increasing the Reliability of Crowdsourcing Evaluations Using Online Quality Assessment. *IEEE Transactions on Affective Computing* 7, 4 (October-December 2016), 374–388. https://doi.org/10.1109/TAFFC.2015.2493525

[3] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. 2004. Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. In *Sixth International Conference on Multimodal Interfaces ICMI 2004*. ACM Press, State College, PA, 205–211. https://doi.org/10.1145/1027933.1027968

[4] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost. 2017. MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *IEEE Transactions on Affective Computing* 8, 1 (January-March 2017), 67–80. https://doi.org/10.1109/TAFFC.2016.2515617

[5] H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova, and R. Verma. 2014. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing* 5, 4 (October-December 2014), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

[6] J. Chen, Z. Chen, Z. Chi, and H. Fu. 2018. Facial Expression Recognition in Video with Multiple Feature Fusion. *IEEE Transactions on Affective Computing* 9, 1 (January-March 2018), 38–50. https://doi.org/10.1109/TAFFC.2016.2593719

[7] P. Chhikara, P. Singh, R. Tekchandani, N. Kumar, and M. Guizani. 2021. Federated Learning Meets Human Emotions: A Decentralized Framework for Human-Computer Interaction for IoT Applications. *IEEE Internet of Things Journal* 8, 8 (April 2021), 6949–6962. https://doi.org/10.1109/JIOT.2020.3037207

[8] M. Coulson. 2004. Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence. *Journal of Nonverbal Behavior* 28, 2 (June 2004), 117–139.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, USA, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[10] P. Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6, 3-4 (1992), 169–200. https://doi.org/10.1080/02699939208411068

[11] V. Gay, P. Leijdekkers, J. Agcanas, F. Wong, and Q. Wu. 2013. CaptureMyEmotion: Helping Autistic Children Understand their Emotions Using Facial Expression Recognition and Mobile Technologies. In *BLED 2013 Proceedings*. Bled, Slovenia, 409–420.

[12] L. Goncalves and C. Busso. 2022. AuxFormer: Robust Approach to Audiovisual Emotion Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*. Singapore, 7357–7361. https://doi.org/10.1109/ICASSP43922.2022.9747157

[13] L. He, D. Jiang, L. Yang, E. Pei, P Wu, and H. Sahli. 2015. Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. In *International Workshop on Audio/Visual Emotion Challenge (AVEC 2015)*. Brisbane, Australia, 73–80. https://doi.org/10.1145/2808196.2811641

[14] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *ArXiv e-prints (arXiv:1704.04861)* (April 2017), 1–9. arXiv:1704.04861 [cs.CV]

[15] Y. Khaireddin and Z. Chen. 2021. Facial emotion recognition: State of the art performance on FER2013. *ArXiv e-prints (arXiv:2105.03588)* (May 2021), 1–9. https://doi.org/10.48550/arXiv.2105.03588 arXiv:2105.03588 [cs.CV]

[16] D.P. Kingma and J. Ba. 2014. Adam: A Method for Stochastic Optimization. *ArXiv e-prints (arXiv:1412.6980)* (December 2014). arXiv:1412.6980 [cs.LG]

[17] C.-C. Lee, K. Sridhar, J.-L. Li, W.-C.Lin, B.-H. Su, and C. Busso. 2021. Deep Representation Learning for Affective Speech Signal Analysis and Processing: Preventing unwanted signal disparities. *IEEE Signal Processing Magazine* 38, 6 (November 2021), 22–38. https://doi.org/10.1109/MSP.2021.3105939

[18] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang. 2020. TEA: Temporal Excitation and Aggregation for Action Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*. Seattle, WA, USA, 906–915. https://doi.org/10.1109/CVPR42600.2020.00099

[19] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. 2019. MediaPipe: A Framework for Perceiving and Processing Reality. In *Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR 2019)*. Long Beach, CA, USA, 1–4.

[20] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of Machine Learning Research (PMLR 2017)*, A. Singh and J. Zhu (Eds.). Vol. 54. PMLR, Fort Lauderdale, FL, USA, 1273–1282.

[21] A. Mollahosseini, B. Hasani, and M. H. Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing* 10, 1 (January-March 2019), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

[22] D. Nikolova, P. Petkova, A. Manolova, and P. Georgieva. 2018. ECG-based Emotion Recognition: Overview of Methods and Applications. In *Advances in Neural Networks and Applications (ANNA 2018)*. St. Konstantin and Elena Resort, Bulgaria, 1–5.

[23] T. Ojala, M. Pietikainen, and T. Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 7 (July 2002), 971–987. https://doi.org/10.1109/TPAMI.2002.1017623

[24] O.M. Parkhi, A. Vedaldi, and A. Zisserman. 2015. Deep Face Recognition. In *British Machine Vision Conference (BMVC 2015)*. Swansea,UK, 1–12. https://doi.org/10.5244/c.29.41

[25] S. Parthasarathy and C. Busso. 2020. Semi-Supervised Speech Emotion Recognition with Ladder Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (September 2020), 2697–2709. https://doi.org/10.1109/TASLP.2020.3023632

[26] R. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA, USA.

[27] A. Radoi, A. Birhala, N. C. Ristea, and L.-C. Dutu. 2021. An End-To-End Emotion Recognition Framework Based on Temporal Aggregation of Multimodal Information. *IEEE Access* 9 (September 2021), 135559–135570. https://doi.org/10.1109/ACCESS.2021.3116530

[28] A. Raheel, M. Majid, and S. M. Anwar. 2019. Facial Expression Recognition based on Electroencephalography. In *International Conference on Computing, Mathematics and Engineering Technologies (iCoMET 2019)*. Sukkur, Pakistan, 1–5. https://doi.org/10.1109/ICOMET.2019.8673408

[29] A.N. Salman and C. Busso. 2020. Dynamic versus Static Facial Expressions in the Presence of Speech. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. Buenos Aires, Argentina, 436–443. https://doi.org/10.1109/FG47880.2020.00119

[30] A. N. Salman and C. Busso. 2020. Style Extractor for Facial Expression Recognition in the Presence of Speech. In *IEEE International Conference on Image Processing (ICIP 2020)*. Abu Dhabi, United Arab Emirates (UAE), 1806–1810. https://doi.org/10.1109/ICIP40778.2020.9191330

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*. Salt Lake City, UT, USA, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

[32] A. Seyeditabari, N. Tabari, and W. Zadrozny. 2018. Emotion detection in text: a review. *ArXiv e-prints (arXiv:1806.00674)* (June 2018), 1–14. https://doi.org/10.48550/arXiv.1806.00674 arXiv:1806.00674 [cs.CL]

[33] D. Shome and T. Kar. 2021. FedAffect: Few-shot federated learning for facial expression recognition. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW 2021)*. Montreal, BC, Canada, 4151–4158. https://doi.org/10.1109/ICCVW54120.2021.00463

[34] K. Simonyan and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR 2015)*. San Juan, Puerto Rico, 1–10.

[35] I. Song, H.-J. Kim, and P. B. Jeon. 2014. Deep learning for real-time robust facial expression recognition on a smartphone. In *IEEE International Conference on Consumer Electronics (ICCE 2014)*. Las Vegas, NV, USA, 564–567. https://doi.org/10.1109/ICCE.2014.6776135

[36] M. Tan and Q. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of Machine Learning Research (PMLR 2019)*, K. Chaudhuri and R. Salakhutdinov (Eds.). Vol. 97. PMLR, Long Beach, CA, USA, 6105–6114.

[37] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. 2011. The first facial expression recognition and analysis challenge. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2011)*. Santa Barbara, CA, USA, 921–926. https://doi.org/10.1109/FG.2011.5771374

[38] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao. 2020. Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition. *IEEE Transactions on Image Processing* 29 (January 2020), 4057–4069. https://doi.org/10.1109/TIP.2019.2956143

[39] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan. 2016. Peak-Piloted Deep Network for Facial Expression Recognition. In *European Conference on Computer Vision (ECCV 2016)*, B. Leibe, J. Matas, N. Sebe, and M. Welling (Eds.). Lecture Notes in Computer Science, Vol. 9905. Springer Berlin Heidelberg, Amsterdam, the Netherlands, 425–442. https://doi.org/10.1007/978-3-319-46475-6_27