FedAffect: Few-shot federated learning for facial expression recognition

Debaditya Shome and T. Kar
KIIT University
Odisha, India
1804372@kiit.ac.in, tkarfet@kiit.ac.in

Abstract

Annotation of large-scale facial expression datasets in the real world is a major challenge because of privacy concerns of the individuals due to which traditional supervised learning approaches won't scale. Moreover, training models on large curated datasets often leads to dataset bias which reduces generalizability for real world use. Federated learning is a recent paradigm for training models collaboratively with decentralized private data on user devices. In this paper, we propose a few-shot federated learning framework which utilizes few samples of labeled private facial expression data to train local models in each training round and aggregates all the local model weights in the central server to get a globally optimal model. In addition, as the user devices are a large source of unlabeled data, we design a federated learning based self-supervised method to disjointly update the feature extractor network on unlabeled private facial data in order to learn robust and diverse face representations. Experimental results by testing the globally trained model on benchmark datasets (FER-2013 and FERG) show comparable performance with state of the art centralized approaches. To the best of author's knowledge, this is the first work on few-shot federated learning for facial expression recognition.

1. Introduction

Humans can communicate their emotional states and intentions through facial expressions, which are one of the most powerful, natural and universal signals [9, 28]. Facial expression recognition (FER) plays an important role in a plethora of human-centric computing applications. There has been significant progress in increasing FER performance due to advances in deep learning aided computer vision and large-scale annotated datasets. However, in real-life situations, facial expressions of people are influenced by a variety of personal characteristics such as gender, age, ethnicity, personality, and expressive styles. Due to these reasons, FER models trained on large datasets are likely to suffer from dataset bias and would fail to generalize on facial data from people which the model hasn't encountered before [19]. Moreover, it has been seen that performance of FER hasn't kept on improving even on merging multiple large-scale datasets [36]. These facts prove that it is essential to go beyond traditional supervised learning and utilize facial expression data from varied set of users under different conditions. Due to privacy sensitive nature of facial expression data, it has become a major bottleneck for research in FER. Strict data protection laws are being imposed in order to eliminate the collection of user's private data with consent. Many publicly available face image datasets have been withdrawn from the internet due to such non-consent based facial data collection [3]. Recently, federated learning (FL) has emerged as an effective technique of training of models on decentralized private data. FL for FER hasn't been looked into yet by the research community as real world user devices hold mostly unlabeled data and only few samples of labeled data per device is possi-

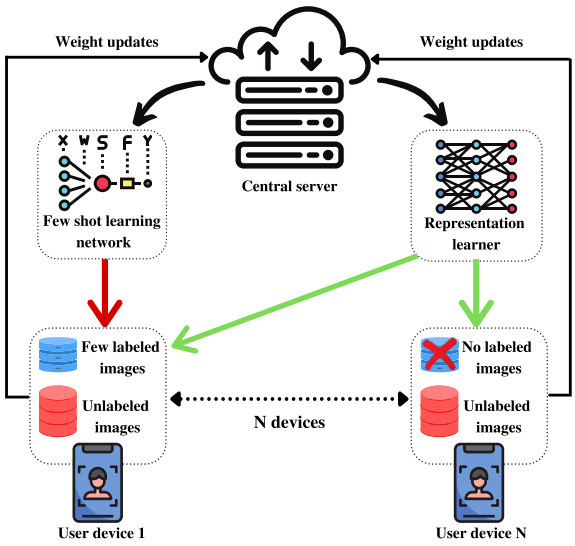


Figure 1: FedAffect framework

ble to acquire on a daily basis. In an attempt to solve the above challenges, in this paper we propose an FL framework called FedAffect which utilizes a few-shot learning network for FER from few labeled samples present on user devices and a self supervised representation learner which utilizes all unlabeled decentralized facial data available to learn effective representations of faces from a varied set of users.

Our overall contributions are summarized as follows,

- We propose FedAffect, a few-shot meta learning based FL framework which is capable to learn from few labeled FER images present on decentralized user devices.
- We design an FL based self supervised representation learning approach to effectively utilize the large corpus of unlabeled facial images to train a globally optimal feature extraction network.
- We evaluate our FedAffect algorithm on benchmark datasets where it achieves an accuracy of 84.9% on FER-2013, and 97.3% accuracy on FERG.

2. Related works

In this section we give a brief overview of the relevant literature in the domains of FER, FL and few-shot learning for image classification.

Facial expression recognition: FER received a lot of attention in the past few years due to the advancement of deep neural network architectures and availability of large annotated datasets [10, 22, 25]. Research on FER can be grouped into two categories, handcrafted feature based FER and end to end FER. The widely used handcrafted features include HOG [8], LBP [24], etc. End to end FER approaches have been based on large deep learning models. Yang *et al*. [34] proposed a FER model to learn expressions using De-exprssion Residue learning (DeRL) network. They generated de-expressioned neutral images by training a cGAN. Hayale *et al*. [15] exploited deep siamese neural networks with a supervised loss function for building a FER system by minimising intra class feature differences and maximising inter class feature differences. Gera and Balasubramanian [13] presented a method termed CCT for efficiently training a FER system with noisy crowd-sourced annotations. By co-training three networks, CCT was able to counteract the noisy labels. Vo *et al*. [30] utilized a pyramid network based super resolution architecture to solve the problem of varying FER image sizes in the wild which improved classification performance on benchmarks. Aouayeb *et al*. [5] proposed an extension of Vision transformers with a squeeze and excitation based MLP head and trained it for FER.

Few-shot learning: Recently, meta learning and few-shot learning approaches have been proved to be highly efficient when it comes to scarcity of labeled data [32]. Delta-encoder was designed by Schwartz *et al*. [23] which utilized a modified autoencoder like architecture in order to synthesize unseen classes of objects only from a few samples. Douze *et al*. [12] leveraged the idea of large scale similarity graph construction using a diffusion setup to learn from few labeled samples. Chu *et al*. presented Spot and Learn which utilizes reinforcement learning based positive and negative sampling policies to determine the favorable regions in an image and to regularize the learning process in order to boost performance of few-shot learning. In a comparative study on few-shot learning for FER [7], it was seen that the few-shot models achieved comparable performance with supervised approaches, but their performance significantly degrades upon shift in domain of data distribution. Although, shifting from one FER dataset to the other showed a negligible drop in performance, thus proving that few-shot learning for FER can be of significant value when training in decentralized setups such as FL.

Federated learning: A lot of work has been done recently on FL in variety of applications due to it's efficiency in utilizing private user data for training models on multiple devices collaboratively [2, 17]. Li *et al*. [18] proposed MOON, an FL framework which makes use of similarities between representations of models to perform contrastive training at model level. He *et al*. [16] proposed FedGKT, an approach for federated training of large CNNs at the edge and regularly transferring locally learnt knowledge to the global server's CNN in order to reduce computation by upto 17 times as reported in the experimental results. Aggarwal *et al*. [3] presented FedFace, an FL framework for collaboratively learning to recognize faces from user's private data.

3. FedAffect framework

In this section, we provide an overview and implementation details of our FedAffect framework and present it in four modules.

3.1. Local representation learning

Large amounts of unlabeled face data are stored on user devices which are not used in the training of FL algorithms using traditional supervised approaches. Using frozen feature extractors trained on large datasets such as Imagenet [11], the features extracted from varying user's faces won't represent much rich information to deal with the complexity of FER. In an attempt to learn effective face representations in the wild using FL and inspired from SimCLR [6], we design a self-supervised learning approach to utilize decentralized facial data from users to learn robust and scale invariant features. In a user device, at each communication round there would be a set of N images. At each local iteration, a mini-batch of M images is randomly sampled

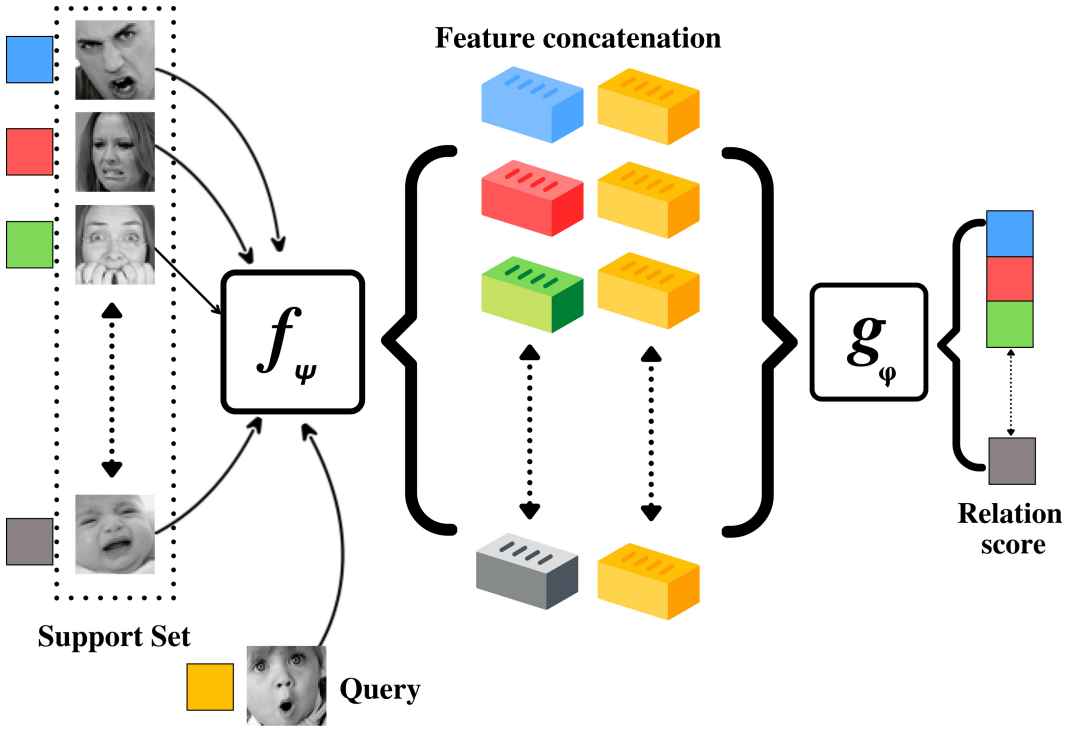


Figure 2: Few-shot learner network

from the N available images. For each local image x in the mini-batch, the representation learner creates two different views, x_i and x_j using a stochastic data augmentation pipeline, thus resulting in $2M$ samples per minibatch. The representation learner consists of an encoder network E and a small projector network P. Vector representations of x_i and x_j are extracted by E and passed on to P which maps them to a latent feature space where the loss function is applied. The contrastive loss for each pair of the projected vectors among the other $2(M-1)$ pairs in the minibatch can be calculated as :

$$l(i,j) = -\log \frac{\exp(\frac{sim(x_i,x_j)}{\tau}))}{\sum_{k=1}^{2M} I_{[k\neq i]} \exp(\frac{sim(x_i,x_j)}{\tau})} \quad (1)$$

where $sim(.)$ is the pairwise cosine similarity, τ refers to the temperature parameter. The net loss over a minibatch is calculated by taking the sum over all augmented pairs of images in that minibatch. This loss has been used in multiple previous works [6, 21, 26, 33]. We use a Resnet-50V2 backbone without pre-trained weights to design the encoder network E.

3.2. Few-shot classification

In a realistic scenario, only a few samples of labeled data per user device can be possible to acquire in each FL round. Thus, standard supervised classifiers would take a large number FL rounds in order to converge, and considering the fact that there may also be large time intervals between rounds due to non-availability of labeled data, such an approach becomes highly infeasible. In order to tackle the above mentioned challenges, we employ a meta learning strategy inspired from Relation networks [27] for robust few-shot learning. During each local FL round, the available labeled data is sampled in the form of support sets and query sets in a $KC:1$ ratio where C represents the number of classes and K refers to number of samples per class in support set. As illustrated in Figure 2, the overall architecture of the few-shot learning network consists of the local representation learner f_ψ which is used as the embedding module for feature extraction, along with a relation module g_ϕ. When $K=1$, C samples from the support set and one from the query set are passed as input to the representation learner model which produces C feature maps which are depth-wise concatenated. In the case of $K>1$, feature map of each class in the support set is formed by taking the element-wise sum over outputs of local representation learner f_ψ. The extracted feature maps are passed on through the relation module g_ϕ which produces a scalar termed relation score $Y_{i,j} \in [0,1]$ representing the similarity between the support set and query image. We consider the mean squared error (MSE) loss to train our few-shot learning network by regressing the obtained relation score

$Y_{i,j}$ to the ground truth as in equation 2.

$$L_{relation} = arg\ min \sum_{i=1}^{n_q} \sum_{j=1}^{n_s} (Y_{i,j} - 1(y_i == y_j)) \quad (2)$$

3.3. Global learning

As seen in Figure 1, the central server has two global models. One is the representation learner which gets transmitted to multiple user devices periodically in order to utilize their private unlabeled facial data and perform each local representation learning round. The other model is the few-shot learner which is transmitted to only those devices which have atleast C number of labeled samples in order to perform a one shot learning round. Both the models are trained disjointly, and the global few-shot learner's embedding module f_ψ get's updated each time the representation learner get's updated globally. For global aggregation, we consider the standard FedAvg algorithm [20]. The complete procedure of the proposed FedAffect framework is shown in algorithm 1.

4. Experiments

In this section, we discuss about the implementation details and experiments performed in order to evaluate our proposed FedAffect framework.

4.1. Dataset descriptions

FER-2013: FER-2013 is a large collection of images for FER which were obtained automatically using Google image search API. The dataset is available with images of size 48 × 48 pixels categorized into 7 FER labels. It consists of 28,709 training, 3589 validation and 3589 test images.

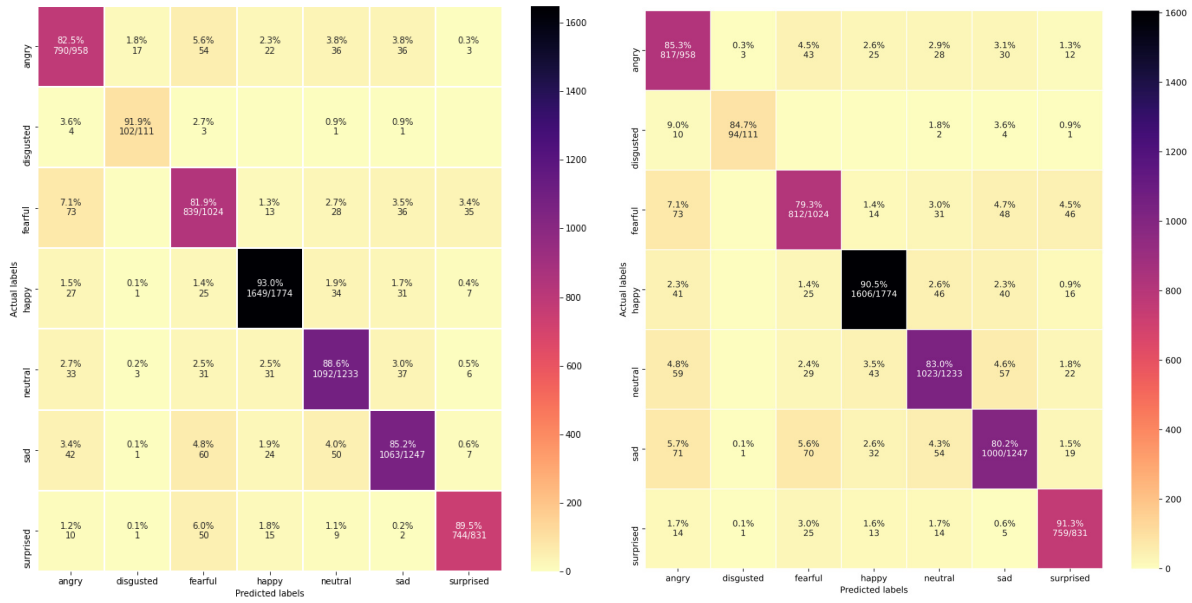
FERG: FERG[4] is a large dataset of stylised characters with face expressions that have been annotated. There are 55,767 annotated face pictures of six stylised characters in the collection. A software called MAYA was used to create the characters. Each character's pictures are divided into seven different sorts of expressions.

4.2. Evaluation

In order to evaluate the local representation learning method, we split the available datasets into training and testing. We use all the images from FERG dataset and train the model on the training set images without utilizing the labels in both centralized learning and federated learning settings using the TensorFlow library [1]. For FL, we use FERG dataset's six characters data separately as 6 decentralized datasets in order to make the simulation close to the real world FL setting where each user would have their own face images in their device. We extract features from the images using the trained representation learner models and implement t-SNE [29] to visualize the projections of the extracted features to interpret our approach. As seen in Figure 4, the same model trained with centralized learning

Algorithm 1 FedAffect framework

Input: number of devices N,
number of communication rounds T,
number of representation learner epochs $E1$,
number of classes C,
learning rate η
Output: Globally trained model weights w_f^t and w_g^t
Server executes:
Initialize w_f^0, w_g^0
Fetch data availability information
for $t = 0, 1, ..., T-1$ **do**
  for $i = 1, 2, ..., N$ in parallel **do**
    if Number of labeled data samples at $i > C$ **then**
      Send w_f^0, w_g^0 to i
      $(w_f^t)_i$, $(w_g^t)_i \longleftarrow$ **LocalFewShot**(i, w_f^t, w_g^t)
    end if
    if i has unlabeled data **then**
      $(w_f^t)_i \longleftarrow$ **LocalReprLearn**(i, w_f^t)
    end if
  end for
  $w_f^{t+1} \longleftarrow \sum_{k=1}^{N} \frac{D_i}{D} (w_f^t)_K$
  $w_g^{t+1} \longleftarrow \sum_{k=1}^{N} \frac{D_i}{D} (w_g^t)_K$
end for
return w_f^t, w_g^t
LocalReprLearn(i, w_f^t):
Initialize projection network p
Initialize encoder network f based on w_f^t
Set batch size B
for sampled minibatch x_k from $k = 1$ to B **do**
  for all $k \in (1, ..., B)$ **do**
    select two augmentation functions $t\ T$, $t\ T'$
    get first projection $z_{2k-1} = p(f(t(x_k)))$
    get second projection $z_{2k} = p(f(t'(x_k)))$
  end for
  $l(i,j) = -\log \frac{\exp(\frac{sim(x_i, x_j)}{\tau}))}{\sum_{k=1}^{2M} I_{[k \neq i]} \exp(\frac{sim(x_i, x_j)}{\tau})}$
  $L = \frac{1}{2B} \sum_{1}^{B} [l(2k-1, 2k) + l(2k, 2k-1)]$
  Update networks f and g to minimize L
end for
return updated weights of encoder, w_f^t
LocalFewShot(i, w_f^t, w_g^t):
Initialize embedding module f based on w_f^t
Initialize relation module f based on w_g^t
Sample support set S and query Q
Train f and g jointly to minimize $L_{relation}$ from equation 2
return $(w_f^t)_i$, $(w_g^t)_i$



(a) Centralized learning on FER-2013

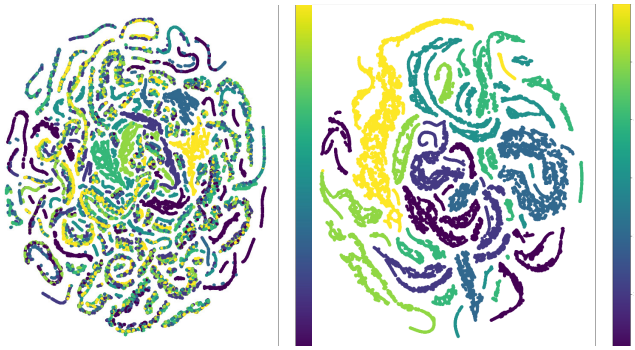
(b) Federated learning on FER-2013

(c) Centralized learning on FERG

(d) Federated learning on FERG

Figure 3: Confusion matrices upon evaluation on FER-2013 and FERG datasets

clearly overfits the data points while the one trained with FL shows separable and distinct features between classes which would be extremely beneficial for boosting the performance of FER. In order to evaluate the few-shot learning network, we divide the FERG dataset in the same way as done during t-SNE visualization discussed above. We consider only 5 of the 6 users for training as we keep aside an unseen user for testing the global model. In every communication round of the FL simulation, for each of the five users a set of $KC + 1$ images are randomly sampled from the number of available labeled images in their local datasets. For simplicity, we consider only one-shot learning per episode ($K = 1$) and in case of FERG dataset, $C = 7$. Hence, we sample 8 images in every training step and use an image from the class which has more than one support sample as the query image. Our representation learner pre-trained on unlabeled



(a) Centralized learning (b) Federated learning

Figure 4: t-SNE visualization of extracted features from FERG dataset

samples acts as the embedding module f_ψ. The features extracted from the 7 support samples are concatenated with the features extracted from the query sample and passed on to the relation module to get the relation score g_ϕ. After training the models by simulating algorithm 1, we use the unseen user's dataset for testing the global model. In table 1, we compare FedAffect's performance with previous state of the art approaches on FERG dataset as well as with our model trained in centralized manner as a baseline. It can be observed that FedAffect achieves a better performance than all of the other approaches including our centralized baseline. This is the result of the efficient personalized feature extraction by the representation learner in the FL setup as demonstrated in the t-SNE plot earlier in Figure 4.

Method	Overall accuracy
Multi-feature ensemble [37]	97%
DeepExpr [4]	89.02%
Centralized (ours)	89.7%
FedAffect (proposed)	**97.3%**

Table 1: Performance comparison on FERG dataset

For cross-dataset evaluation, we take the FER-2013 dataset and split it into five decentralized datasets and train the model using the same FL setup as for the FERG dataset. In table 2, we compare FedAffect with it's centralized counterpart as well as previous state of the art approaches on FER-2013. The globally trained model outperforms and all previous benchmarks except [31] and our centralized baseline. This is due to the in-the-wild nature of face images present in the dataset which has occlusions and even some non-facial images. Moreover, as person-wise images were not available in FER-2013, each decentralized dataset created had random set of faces rather than a particular person's face which would generally be seen in a real life user

Method	Overall accuracy
CNN [35]	65.97%
Ensemble ResMaskingNet [14]	76.8%
RAN-VGG16 [31]	89.16%
Centralized (ours)	87.51%
FedAffect (proposed)	**84.9%**

Table 2: Performance comparison on FER-2013 dataset

device and thus this lack of personalization leads to less utility of our FL framework. The confusion matrices in Figure 3 validate the generalization capability of FedAffect on unseen test samples in order to accurately recognize facial emotions with a negligible amount of misclassification among different classes.

5. Conclusion and future work

In this paper, we tackle the problem of training facial expression recognition directly from decentralized privacy-sensitive data available on user devices. We propose FedAffect, a novel federated learning framework which collaboratively trains two disjoint neural networks, one for self-supervised representation learning from large scale unlabeled facial images and another for utilizing the representation learner as feature extractor and predicting probability scores in a few-shot learning setting on the extracted features for robust facial expression recognition. We evaluate our approach on two standard benchmark datasets namely FER-2013 and FERG. The proposed framework is able to outperform several state of the art centralized learning models without any facial image leaving the user devices. In the future, we aim to extend this work to a Non-IID FL setup and also to be able to extract faces from in-the-wild images automatically in an end to end manner.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 4

[2] Sawsan AbdulRahman, Hanine Tout, Hakima Ould-Slimane, Azzam Mourad, Chamseddine Talhi, and Mohsen Guizani. A survey on federated learning: The journey from central-

ized to distributed on-site learning and beyond. *IEEE Internet of Things Journal*, 8(7):5476–5497, 2021. 2

[3] Divyansh Aggarwal, Jiayu Zhou, and Anil K Jain. Fedface: Collaborative learning of face recognition model. *arXiv preprint arXiv:2104.03008*, 2021. 1, 2

[4] Deepali Aneja, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. Modeling stylized character expressions via deep learning. In *Asian conference on computer vision*, pages 136–153. Springer, 2016. 4, 6

[5] Mouath Aouayeb, Wassim Hamidouche, Catherine Soladie, Kidiyo Kpalma, and Renaud Seguier. Learning vision transformer with squeeze and excitation for facial expression recognition. *arXiv preprint arXiv:2107.03107*, 2021. 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3

[7] Anca-Nicoleta Ciubotaru, Arnout Devos, Behzad Bozorgtabar, Jean-Philippe Thiran, and Maria Gabrani. Revisiting few-shot learning for facial expression recognition. *arXiv preprint arXiv:1912.02751*, 2019. 2

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 2

[9] Charles Darwin. *The expression of the emotions in man and animals*. University of Chicago press, 2015. 1

[10] S. Datta, G. Sharma, and C. Jawahar. Unsupervised learning of face representations. *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2018. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[12] Matthijs Douze, Arthur Szlam, Bharath Hariharan, and Hervé Jégou. Low-shot learning with large-scale diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3349–3358, 2018. 2

[13] Darshan Gera and S Balasubramanian. Affect expression behaviour analysis in the wild using consensual collaborative training. *arXiv preprint arXiv:2107.05736*, 2021. 2

[14] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013. 6

[15] Wassan Hayale, Pooran Negi, and Mohammad Mahoor. Facial expression recognition using deep siamese neural networks with a supervised loss function. pages 1–7, 05 2019. 2

[16] Chaoyang He and Murali Annavaram. Group knowledge transfer: Federated learning of large cnns at the edge. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, (33), 2020. 2

[17] Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pages 794–797. IEEE, 2020. 2

[18] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722, 2021. 2

[19] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 2020. 1

[20] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 4

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[22] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[23] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Rogerio Feris, Abhishek Kumar, Raja Giryes, and Alex M Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. *arXiv preprint arXiv:1806.04734*, 2018. 2

[24] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009. 2

[25] V. Sharma, M. Tapaswi, M. S. Sarfraz, and R. Stiefelhagen. Self supervised learning of face representations for video face clustering. *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2019. 2

[26] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 3

[27] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 3

[28] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001. 1

[29] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 4

[30] Thanh-Hung Vo, Guee-Sang Lee, Hyung-Jeong Yang, and Soo-Hyung Kim. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, 8:131988–132001, 2020. 2

[31] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion

robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 6

[32] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020. 2

[33] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3

[34] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018. 2

[35] Lutfiah Zahara, Purnawarman Musa, Eri Prasetyo Wibowo, Irwan Karim, and Saiful Bahri Musa. The facial emotion recognition (fer-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (cnn) algorithm based raspberry pi. In *2020 Fifth International Conference on Informatics and Computing (ICIC)*, pages 1–9. IEEE, 2020. 6

[36] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018. 1

[37] Hang Zhao, Qing Liu, and Yun Yang. Transfer learning with ensemble of multiple feature representations. In *2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 54–61. IEEE, 2018. 6