Investigating the problem domain of FER in the wild

Shabnam Khodadadi
Queen Mary University of London
London, E1 4NS
s.khodadadi@se21.qmul.ac.uk

1. Introduction

One of the most potent, inherent, and universal signals used by people to communicate their emotional states and intents is facial expression [7]. These expressions are nonverbal channels through which Human Machine Interaction (HMI) systems can recognize human internal emotions [39]. Facial expression recognition (FER) has been a subject of study for decades due to its importance in a variety of industries, including computer vision, digital entertainment, advertising, health care, and robotics [5] [4] [14] [37] [41]. However, recognizing facial expressions in real-life environments is still very challenging due to variations, occlusions, and the ambiguity of human emotion which the cultural and ethnic background of a person can affect his expressive style [7]. Ekman et al. identified six universal facial expressions (anger, disgust, fear, happiness, sadness, and surprise) as basic emotional expressions that are in common among humans [13]. FER is typically conducted utilizing handcrafted features in traditional methods, such as LBP [48], HOG [6], and SIFT [42]. Although these features have a strong intuitive basis and perform admirably on a number of lab-controlled databases, they lack generalizability and adequate learning capability [30]. In recent years, new in-the-wild facial expression databases were created that provide FER research the opportunity of being conducted in more challenging environments. By learning potent high-level features, Deep Convolutional Neural Networks (CNNs) have demonstrated promising recognition performance [9]. In this article, we are investigating the problem domain of FER in the wild which includes the three major stages of Hypothesis, Data preparation, and Data preprocessing. The dataset that we work on is the synthetic dataset consisting of images that are artificially generated from the Aff-Wild2 database [18] [28] [21] [23] [27] [25] [22] [24] [26] [19] using methods [44] [20] [19] [26]. Images in the provided dataset are annotated in terms of Ekman et al [13] basic expressions.

2. Related Work

Facial Expression Recognition (FER) in real-life environments faces significant challenges, the most important one is the Inconsistency of annotated data, there are errors and biases in human annotations exist among different datasets. Different cultures, living conditions, and other experiences affect how well humans comprehend facial expressions. These Inconsistency of existing datasets cause two main issues in FER, 1) Usually FER systems inherit the recognition bias from the training dataset. It means that in the case which one similar image has different labels in various datasets, the models that trained on these datasets will predict inconsistant label for same unlabeld sample. 2)It is not possible to tackle inconsistency problem by simply merging various databases during training, Experiences in [56] shows that the performance of the models which trained on single dataset is better than the one that use accumulation of various datasets for training. [56] proposed a framework named IPA2LT to address the mentioned issues. IPA2LT has three stages, in first stage, two FER models are trained by adopting any classification methods on two different human annotated datasets the output is the model A and B which trained on datsets A and B respectively. Then each machine is used as an annotator for the other dataset. Finally each image has multiple labels which consist of its own human annotation in addition to machine-predicted label. These labels are probably inconsistent. In the last stage IPA2LT trains an end-to-end Latent Truth Net(LTNet) [56] to discover the latent truth considering the inconsistent labels and the input images.This is the first work that addresses the annotation inconsistency in different FER datasets. Three FER in-the-wild datasets (SFEW [31], RAF [8], AffectNet [40]) and four in-the-lab ones(Oulu-CASIA [57], MMI [50], CK+ [35], CFEE [11]) use for experiments. In comparison to the state-of-the-art methods, the IPA2LT(LTNet) achieve the accuracy of 87.23 which is higher than the previous best 84.41 [56].

Existing FER methods are not sufficiently generalized, they don't achieve significant results when applied to unseen data or images that are captured in an unconstraint environment. This problem happened because lighting and pose conditions are strictly regulated in most databases which cause the FER recognition methods to only able to identify exaggerated or limited expressions similar to those in the training database. Moreover, they are limited available databases in this area while neural network methods need a large amount of data to train. [38] proposes a cross-dataset approach to address the two mentioned problems across multiple well-known standard datasets. Cross-database tasks are difficult to solve as each database has its own fingerprint in case of illumination, posture, resolution, etc. This characteristic makes it hard to extract features for cross-dataset tasks. [38] build network-in-network architecture [34] using the Inception layer structure [45] which provides the dense network structure required for efficient computation. Extensive experiments are carried out on seven public facial expression databases, MultiPIE [15], MMI [43], CK+ [35], DISFA [36], GEMEPFERA [2], SFEW [8], and FER2013 [49]. One of the mentioned databases is used for evaluation and rest of them are used for training to achieve the goal of model generalization. The proposed method outperforms the state-of-the-art cross-dataset methods in four of the datasets. Also, since there are no studies on cross-database evaluation of FERA, SFEW, and FER2013, the [38] results can be a baseline for cross-database of these challenging datasets. The network structure is the first architecture that applies the Inception layer to the FER problem across multiple datasets.

Local patches are image regions such as eyes, nose, and mouth in faces, they play a crucial role in distinguishing various expressions. [54] propose a mechanism to find various local patches and the relationship between them for FER purposes. However, the recognision methods which use this approach face several issues: 1) In images with occlusions or large pose variations, some facial parts are invisible. [54] use Multi-Attention Dropping (MAD) [32] [51] to solve this issue by finding the huge number of local patches and dropping the redundant ones to keep the diverse representations for classifying different emotions. 2) It is difficult to distinguish between similar emotions such as surprise and anger based on just one single region(mouth, exp). To address this issue [54] use Vision Transformer (ViT) [10] to explore the relations among different local patches in a global scope. To conclude, [54] proposes the transform model(Trans-FER) to learn diverse relation-aware local patches for FER, in this method it is generally assumed that mouth, nose, and eyes are the most useful regions to distinguish various emotions. RAF-DB [31], FERPlus [3], AffectNet [40] datasets are used for experience. Trans-FER achieves the accuracy of over 90% on RAF-DB dataset, which is 2.84% better than the best result reported before, Also it outperforms the previous best result by 1.03% in a challenging dataset AffectNet. Finally, it works less than 1% better in FERPlus in comparison to state-of-the-art methods in the same dataset. This is the first work to explore Transformers for FER, to the best of our knowledge [54].

In dimensional models [46], the two latent dimensions of valence (how good or bad a feeling is) and arousal (how likely the individual is to behave as a result of the emotion) are used to characterize emotion. While categorical models categorize expressions as one of the basic emotions defined by [13]. However, there is a strong dependence between the categorical and the dimensional model. [1] Introduced Emotion-GCN, a multi-task learning(MTL) framework that uses a graph convolutional network [53] to distinguish facial expression by utilizing the dependencies between the category and dimensional models. In this framework, a classifier is learned for each facial expression and a regressor for each dimension in VA(valence-arousal) space. Also, a graph convolutional network(GCN) is created [53]. The graph nodes are seven human emotions(the [13] and the natural emotion), In addition to two other nodes which are valence and arousal. Nodes are connected based on their correlation in the graph. The output of the network is vector weights which are applied to image features that are extracted with DenseNet to make a recognition. This work is a novel GCN-based MTL framework that is proposed for in-the-wild FER. AffectNet [40] and Aff-Wild2 [25] are used in this research, these datasets have both categorical and VA annotations(valence-arousal). Emotion-GCN framework vastly outperforms the state-of-the-art methods with an accuracy of 66.46

3. Hypotheses-Restrictions

Facial expression in the wild or real environment refers to the videos that we capture with our own cameras or the streaming videos on television or social networks such as movies or talk shows which contain emotions. In another word, the in-the-wild data capture under uncontrolled conditions [12]. The question is, Can the problem of FER be used in these real-life situations? No, with our current resources and technologies, and pieces of information we are not able to recognize the facial expression in these wild raw data. Therefore, we should apply some restrictions to the problem to make it solvable. We will provide some of the most important Hypotheses in the following paragraphs. Generally, identifying facial expressions under uncontrollable situations is a difficult task because of varying facial positions and view angles, sophisticated light levels, face wrinkles, and partial occlusions. To conduct FER, we need to do some pose alignment or brightness adjustment in the data. Consequently, the problem will change to facial expression recognition in the wild under some conditions. First, we limited the expression recognition subjects to the two main categories(men and women) while they are other creatures in the wild that have emotions, like animals. The real environment images contain expressions of animals but we only detect the human faces to do the FER [52] so the original problem of FER in the wild will convert to a facial expression recognition of humans in the wild. Second, we extract the short video clips from the videos to process. Whilst, they are complex stories in films and movies and lots of metadata that can help us to distinguish emotions but we ignore them due to the lack of resources and high complexity of processing the whole videos. However, we use limited sequential frames for FER so the new problem is FER of some part of wild data. Despite these hypotheses, Facial expression recognition in the wild is still challenging due to the ambiguity of human emotion which is the effect of the cultural and ethnic background of a person. Moreover, patients with Mobius syndrome [7] or with Botulinum toxin (Botox) injections [16] have difficulty in expressing their emotions with an appearance that makes it difficult to recognize by machines. Imbalance datasets are a barrier too, as a result of the rareness of some emotions in comparison to regular ones. With respect to the current situation of science and technology, the FER problem in the wild without mentioned hypotheses will convert into an unsolvable problem. Experts take into consideration these restrictions while creating the in-the-wild FER dataset. [52] [55]

4. Data Preparation

Our dataset contains 43447 image files in JPG format categorized in folders of six simple expressions [13], However the number of samples for each class is not equal, our dataset is imbalance due to class distribution of data. We solve this problem in the data preprocessing stage, after splitting data to train/test, to prevent data leakage. We split the data to training and testing set. Partitioning is done in a class dependent manner with the test ratio of 0.25, our division is not subject(person) independant as we are not able to distinguish between images of same people based on the file names. We choose this split rate for with respect to literatures choice for splitting data to test and train [54]. In details, for each class we shuffle the images then randomly choose the 0.25 of them for training and the rest for testing for that specific class without overlapping. After the splitting, the train set have 32574 images, and test set have 10861 samples in all six classes. The images in our dataset is in RGB format and have the size of 128 128 pixels. For furthers stages we need our data to be in array format so we convert the images to numpy array of (128,128,3), for labels we assign the digit of 0 to 5 to the six categorical class names to form an array of labels.

5. Data Pre-processing

Data pre-processing entails the procedures necessary to make it easier to extract useful features from the data. The usual pre-processing steps in FER are face detection and alignment, image resizing and image normalization. The steps describes below are conducted in this article. The number of samples in different annotated classes in train set is as follows: ANGER:6171, FEAR:2338, SURPRISE:5563, SADNESS:9780, HAPPINESS:6834, DISGUST:1988. These values illustrates that the data is imbalance. Imbalanced class distribution in facial expressions is a common issue, which is a result of the practicality of sample acquirement. For example, collecting and annotating a happy face is simple; however, identifying the suprise, anger and other less common emotions are more effortfull. This problem is occured especially in real-world datasets. One of the solution to solve this problem is using data augmentation along with resampling during the pre-processing step in order to balance the class distribution based on the number of samples for each class. Another alternative is to use a loss function which provides larger weights for minority classes [17]. Since the difference between the number of samples in distinct classes is not high, In detail, the sadness class has 13040 images that are approximately five times more than the disgust class, We applied the data augmentation and resampling process to solve the problem of imbalance dataset [29]. To augment the facial expression data, six types of augmentation techniques are usually used(random horizontal flip, random rotation and changes in brightness, contrast, hue and saturation) [56]. We use the sequence of data augmentation which consists of random rotation in range [0,90] degrees, brightness range [0.2, 1.0], and the horizontal flip [29]. The class distribution after the data augmentation and oversampling process in training data is: ANGER:9776, FEAR:9776, SURPRISE:9773, SADNESS:9776, HAPPINESS:9777, DISGUST:9778 After data augmentation Face detection process conduct with [47] proposed method, which are using the multi-task CNN (MTCNN) face detection without margins. However, because of mistakes in face detection, facial regions have not been detected in some images, or for some of them, more than one face is detected. We ignored images with no or more faces in the training set. In the following step, we used the detected face bounding boxes for image alignment. The purpose of alignment is to achieve a normalized representation of each face. For alignment, we use the approach of [33], which rotates the facial images to align them based on the position of the eyes. location of the eyes is calculated by taking the mean value of the six detected landmarks in each eye [1]. We also applied normalization on the value of pixels, the range of values are 0 to 255 after normalization they convert to the range of 0 to 1. Then images are resized to 96×96×3 pixels for analysis due to the limited resources for processing, Even though most of the deep neural networks accept the 128×128×3. According to [40] reducing this resolution has little effect on accuracy but dramatically boosts network speed.

References

[1] Panagiotis Antoniadis, Panagiotis P. Filntisis, and Petros Maragos. Exploiting Emotional Dependencies with Graph Convolutional Networks for Facial Expression Recognition. *arXiv e-prints*, page arXiv:2106.03487, June 2021. 2, 3

[2] Tanja Bänziger and Klaus R Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook*, 2010:271–94, 2010. 2

[3] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, 2016. 2

[4] Paris Mavromoustakos Blom, Sander Bakkes, Chek Tien Tan, Shimon Whiteson, Diederik Roijers, Roberto Valenti, and Theo Gevers. Towards personalised gaming via facial expression recognition. In *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2014. 1

[5] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1548–1568, 2016. 1

[6] 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. 1

[7] Charles Darwin. The expression of the emotions in man and animalsli university of chicago press. 1

[8] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2106–2112. IEEE, 2011. 1, 2

[9] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 118–126. IEEE, 2017. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[11] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the national academy of sciences*, 111(15):E1454–E1462, 2014. 1

[12] Damien Dupre, Nicole Andelic, Gawain Morrison, and Gary McKeown. Assessment of automatic facial expressions recognition" in the wild": a time-series analysis using gamm and sizer methods. In *20th International Conference on Human-Computer Interaction*, 2018. 2

[13] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999. 1, 2, 3

[14] Panagiotis Paraskevas Filntisis, Niki Efthymiou, Petros Koutras, Gerasimos Potamianos, and Petros Maragos. Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction. *IEEE Robotics and automation letters*, 4(4):4011–4018, 2019. 1

[15] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 2

[16] David A Havas, Arthur M Glenberg, Karol A Gutowski, Mark J Lucarelli, and Richard J Davidson. Cosmetic use of botulinum toxin-a affects processing of emotional language. *Psychological Science*, 21(7):895–900, 2010. 3

[17] Wassan Hayale, Pooran Negi, and Mohammad Mahoor. Facial expression recognition using deep siamese neural networks with a supervised loss function. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019. 3

[18] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 1

[19] Dimitrios Kollias, Shiyang Cheng, Maja Pantic, and Stefanos Zafeiriou. Photorealistic facial synthesis in the dimensional affect space. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[20] Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1972–1979. IEEE, 2017. 1

[21] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 1

[22] 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. 1

[23] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 1

[24] 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*, pages 1–23, 2019. 1

[25] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1, 2

[26] Dimitrios Kollias and Stefanos Zafeiriou. Va-stargan: Continuous affect generation. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 227–238. Springer, 2020. 1

[27] 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. 1

[28] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 1

[29] Biao Leng, Kai Yu, and QIN Jingyan. Data augmentation for unbalanced face recognition training sets. *Neurocomputing*, 235:10–14, 2017. 3

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

[31] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 1, 2

[32] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018. 2

[33] Guang Liang, Shangfei Wang, and Can Wang. Pose-invariant facial expression recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021. 3

[34] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 2

[35] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010. 1, 2

[36] 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. 2

[37] Daniel McDuff, Rana El Kaliouby, Jeffrey F Cohn, and Rosalind W Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing*, 6(3):223–235, 2014. 1

[38] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016. 2

[39] Ali Mollahosseini, Gabriel Graitzer, Eric Borts, Stephen Conyers, Richard M Voyles, Ronald Cole, and Mohammad H Mahoor. Expressionbot: An emotive lifelike robotic face for face-to-face communication. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 1098–1103. IEEE, 2014. 1

[40] 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, 2, 3

[41] Ghulam Muhammad, Mansour Alsulaiman, Syed Umar Amin, Ahmed Ghoneim, and Mohammed F Alhamid. A facial-expression monitoring system for improved healthcare in smart cities. *IEEE Access*, 5:10871–10881, 2017. 1

[42] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003. 1

[43] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE, 2005. 2

[44] Andreas Psaroudakis and Dimitrios Kollias. Mixaugment & mixup: Augmentation methods for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2375, 2022. 1

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2

[46] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 2

[47] Andrey V Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 119–124. IEEE, 2021. 3

[48] 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. 1

[49] James Tetazoo. Challenges in representation learning: Facial expression recognition challeng. Available at `http://www.kaggle.com/c/challengesin-representation-learning-facial-expression-recognitionchallenge.l`. 2

[50] Pantic M Valstar, M.F. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. *International Conference on Language Resources and Evaluation, Workshop on EMOTION*, pages 65–70, 2010. 1

[51] 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. 2

[52] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20922–20931, 2022. 3

[53] Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*, 2016. 2

[54] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 3601–3610, 2021. 2, 3

[55] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 3

[56] 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, 3

[57] Huang X. Taini M. Li S.Z. Pietik¨ainen M. Zhao, G. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 1