Learning Transferable Compound Expressions from Masked AutoEncoder Pretraining

Feng Qiu[1], Heming Du[2], Wei Zhang[1], Chen Liu[1,2], Lincheng Li[1,*], Tianchen Guo[2], Xin Yu[2]
[1] Netease Fuxi AI Lab
[2] The University of Queensland
{qiufeng, zhangwei05, lilincheng}@corp.netease.com, chen.liu7@uqconnect.edu.au,
{Heming.du, tianchen.guo, xin.yu}@uq.edu.au

Abstract

*Video-based Compound Expression Recognition (CER) aims to identify compound expressions in everyday interactions per frame. Unlike rapid progress in Facial Expression Recognition (FER) for the basic emotions (*e.g.*, surprised, sad, and fearful), CER with the compound emotions (*e.g.*, fearfully surprised, and sadly fearful) remains underexplored, with an evident gap in the availability of substantial datasets. In this paper, we design a framework to demonstrate the feasibility of predicting compound expressions in-the-wild without relying on domain-specific supervision. To be specific, we first train a model on a large-scale facial dataset using the Masked Autoencoder (MAE) approach to learn comprehensive facial features. Then, to tailor it for facial expression analysis, we fine-tune the ViT encoder on an Action Unit (AU) detection task. To address the issue of insufficient data, we transform the task of recognizing compound emotions into a multi-label recognition task for basic emotions. We train a network by finetuning the pretrained ViT encoder to predict the probability of each basic emotion, and then combine these probabilities to arrive at the final prediction for the compound emotions. Experiments conducted on the C-EXPR-DB dataset demonstrate the effectiveness of our framework in the frame-by-frame prediction of compound expressions in-the-wild. Our framework is recognized as the leading solution in the Compound Expression (CE) Recognition Challenge in the 6th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW). More information for the Competition can be found in: 6th ABAW.*

1. Introduction

In the realm of computer vision, the understanding of human emotions through facial expressions remains a pivotal yet challenging task. Recent progress has greatly advanced the field of Facial Expression Recognition (FER), making it possible to accurately identify basic emotions. These improvements have mainly focused on recognizing the basic emotions, such as surprise, sadness, and fear. Despite these strides, the complexity of human emotional expression, which often encompasses blended or compound states, presents a nuanced challenge that extends beyond the scope of basic emotion recognition.

Despite the critical importance of compound expressions in providing a more comprehensive understanding of human emotions, the field of Compound Expression Recognition (CER) remains notably underdeveloped. This is partly attributable to the limited availability of extensive datasets specifically tailored to compound expressions. Therefore, the adoption of transfer learning techniques, capable of predicting compound expressions without specialized datasets, becomes imperative. To facilitate the development of understanding compound emotions, the 6th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW) hold the Compound Expression (CE) Recogntion based on the C-EXPR-DB [13, 16–24, 64] database.

In this work, we introduce a comprehensive framework specifically designed to address the challenge of predicting compound facial expressions directly from unstructured environments, thereby eliminating the reliance on domain-specific annotated data. Our approach centers on the development of a robust feature extractor tailored for expression analysis, utilizing a novel Facial Expression Masked AutoEncoder (FE-MAE). This AutoEncoder is pretrained on an extensive and diverse dataset comprising both images and videos, which encapsulates a wide spectrum of facial expressions. Subsequently, we retain the ViT encoder and fintune it on the Action Unit (AU) detection task [67] to extract features that are more effective for analyzing effective behaviors. This pretraining phase is crucial, as it equips the ViT encoder with a broad understanding of facial dy-

*Corresponding author

namics and expressions, laying a solid foundation for the subsequent phases of our methodology.

The core of our method involves bridging the gap between the recognition of basic and more complex compound emotions. To achieve this, we fine-tune the pretrained ViT encoder using a carefully curated, albeit limited, basic emotion dataset specifically tailored for compound expressions. After that, we are able to obtain the final prediction for compound emotions by combining the probability values of basic emotions. To be specific, we first train on the multi-label classification task for six basic emotions through fine-tuning the ViT encoder. These six emotions are: Surprised, Fear, Disgust, Happiness, Sadness, and Anger. Then, by combining the output probabilities of two of these emotions, we obtain the probability value for the corresponding compound emotion, e.g. Fearfully Surprised, Happily Surprised, Sadly Surprised, Disgustedly Surprised, Angrily Surprised, Sadly Fearful and Sadly Angry. The one with the highest probability is considered as the final prediction result.

These representations, enriched by the preceding phases of pretraining and fine-tuning, contain deep insights into the subtleties of facial expressions, enabling our framework to achieve high accuracy in identifying compound emotions even in the most challenging and uncontrolled environments [28, 29, 38].

The efficacy of our proposed framework is validated through comprehensive experiments conducted on the C-EXPR-DB dataset [14]. We secured the winning position in the Compound Expression (CE) Recognition Challenge at the 6th Workshop and Competition on Affective Behavior Analysis in the Wild (ABAW).

2. Related Work

Facial Expression Recognition (FER) is a cornerstone in the realm of human affective analysis, enabling machines to interpret human emotions through facial cues. This section delves into the evolution and current state of FER, with a particular focus on the progress of expression recognition from the basic 7 universal expression classes (*e.g.*, angry, happy, and natural) to 12 compound facial expression classes (*e.g.*, angrily disgusted, angrily surprised, and happily disgusted).

2.1. Basic Facial Expression Recognition

Ekman *et al*. [6] identify six basic emotions: happy, sad, angry, fear, disgust, surprise, and natural, in 1971 as the start of FER [30, 31, 39, 58–63, 66, 68, 70]. In the domain of traditional facial expression recognition, the methodology predominantly revolves around a tripartite framework: face detection [25, 41, 50, 54], feature extraction [48, 51], and classification [1, 55, 65]. Face detection technologies have reached a level of maturity and now constitute a distinct avenue of research, frequently applied in practical scenarios. Consequently, the crux of enhancing facial expression recognition methods lies in the subsequent stages of feature extraction and classification. Traditional facial feature extraction algorithms bifurcate into two primary categories: geometric feature-based techniques, such as Active Shape Models (ASM) and Active Appearance Models (AAM), and texture feature-based methods, including Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Gabor wavelet transform. Challenges such as occlusion and variable lighting conditions necessitate more refined feature extraction methods for in-field expression detection.

To bolster facial expression recognition technology, initiatives like FER2013 [9] and EmotiW [3] have orchestrated emotion recognition competitions, amassing extensive training datasets from real-world scenarios. This has paved the way for integrating deep learning models that synergize feature extraction and classification processes, markedly enhancing recognition accuracy. Notably, the application of Deep Convolutional Neural Networks (CNN) [4, 8, 12, 26, 46, 47] by researchers like Tang [49] and Kahou *et al*. [11]. has yielded significant advancements, clinching victories in FER2013 and EmotiW competitions, respectively. This marked a paradigm shift, establishing deep learning as the forefront technology in expression recognition. Yang *et al*. [53] propose a weighted hybrid deep neural network, amalgamating a shallow CNN with a VGG16 network pre-trained on ImageNet [42] for facial feature extraction from LBP and grayscale images, and employing weighted fusion for feature enhancement, to epitomize the evolution of this field. Though this approach improved local expression feature extraction, the incremental accuracy gains underscore the ongoing quest for optimizing traditional facial expression recognition methodologies.

2.2. Compound Facial Expression Recognition

The domain of facial expression recognition focuses on identifying basic, universally recognized expressions, primarily guided by Ekman's seminal work which highlighted six fundamental emotions [6]. However, recent advancements underscore the limitations of this basic emotion model in encapsulating the nuanced and complex nature of human emotional displays. In this context, compound facial expressions, which represent combinations of basic expressions, offer a more granular and accurate portrayal of affective states as encountered in daily human interactions [36].

Compound expressions, constructed from the interplay of two basic emotion categories, embody a richer spectrum of emotional states, extending the descriptive power beyond the basic emotion paradigm. Notable examples include "happily surprised" and "angrily surprised," which individuals can readily identify and distinguish, highlighting

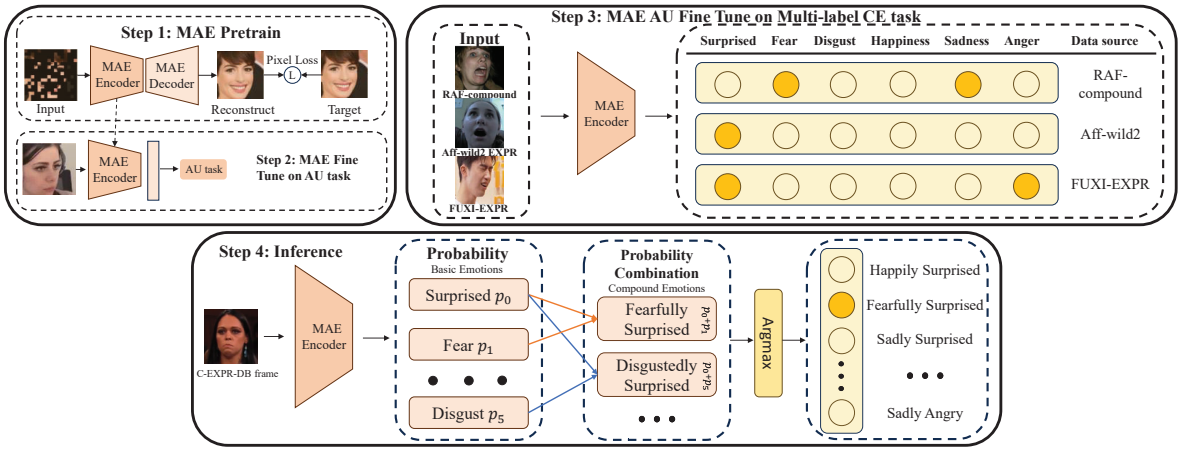

Figure 1. The overview of our proposed framework. Our pipeline includes four steps. Step 1&2: We frist perform MAE pretraining for the facial reconstruction task, and then finetune the ViT encoder on the AU detection task. Step 3: We transform the compound emotion recognition task into a multi-label basic emotion recognition task and further finetune the ViT encoder on this task. Step 4: Finally, when performing inference on compound expression data, we obtain the probability value for the corresponding compound emotion by combining the output probabilities of two basic emotions.

the inherent complexity and diversity in human emotional expression [5]. The leap from basic to compound expression recognition introduces new challenges, chiefly due to the scarcity of comprehensive and well-annotated datasets. Prior efforts rely predominantly on controlled laboratory-generated datasets, which, despite their utility, fall short of capturing the spontaneous and varied nature of real-world expressions. This gap is partially bridged by in-the-wild datasets such as EmotioNet and RAF-DB, which, despite their advancements, still face limitations in terms of size, balance, and modality representation [7, 27].

To address these challenges, Kollias *et al*. [14] introduce the C-EXPR-DB dataset, a substantial in-the-wild audiovisual database meticulously annotated for a wide array of compound expressions. C-EXPR-DB dataset not only expands the available resources for compound expression recognition but also integrates multimodal information, including valence-arousal ratings, action units, and speech cues, thereby enriching the analytical framework for understanding facial expressions in a comprehensive and nuanced manner. Moreover, the advent of multi-task learning frameworks further propel the field, enabling simultaneous learning across related tasks, such as action unit detection and basic expression recognition. These approaches leverage shared representations and inter-task correlations to enhance overall performance, as evidenced by the introduction of C-EXPR-NET by Kollias *et al*. [14], which adeptly combines compound expression recognition with auxiliary tasks to achieve state-of-the-art results. Different from this work, we sequentially conduct pretraining of the ViT encoder first on facial self-reconstruction tasks and then on AU detection tasks to extract more effective facial expression features.

3. Method

3.1. Problem Definition

Given an input video $V_i = \{f_1, f_2, ..., f_N\}$ that consists of N frames, the Compound Expression (CE) recognition task aims to predict one of the pre-defined CE labels for each frame. These pre-defined CE labels consist of Fearfully Surprised, Happily Surprised, Sadly Surprised, Disgustedly Surprised, Angrily Surprised, Sadly Fearful, and Sadly Angry. We first convert this single-class compound emotion classification task to a multi-label basic emotion classification task. Specifically, the label of Fearfully Surprised can be taken as a multi-label of Fear and Surprise. Therefore, the new multi-labels consist of six basic emotions: fear, surprise, happiness, sadness, anger and disgust. One raw CE label can be taken as containing two of these multi-labels.

3.2. MAE pre-training

The pipeline structure of CE recognition is based on Masked Autoencoder [10, 35, 69, 70]. To focus on capturing the facial-related vision features, we employ the Masked Autoencoding (MAE) method to train the model on an image reconstruction task in an unlabeled large-scale face dataset, which includes AffectNet [37], CASIA-WebFace [56], CelebA [32] and IMDB-WIKI [40]. Specifically, we train a model that includes a ViT encoder and decoder on the task of image self-reconstruction. The input to this model is an image with 75% of its patches masked, and the output is the original, unmasked image. MAE learns ef-

ficient and robust representations of the data by reconstructing the original input from partially masked inputs. This encourages the model to understand the underlying structure and features of the data.

3.3. Finetune on AU detection

Then we retain the ViT encoder and finetune it to complete the downstream task of Facial Action Unit (FAU) detection on the Aff-wild2 dataset [13–15, 17, 23, 70]. Facial Action Units serve as the building blocks of facial expressions, providing a detailed and objective way to describe and analyze the complex dynamics of human faces. These movements can reflect subtle changes in facial expression, and the presence or absence of specific AUs can indicate a wide range of emotions or expressions.

3.4. Train on basic emotions

Then, to make fuller use of the available data, we train a multi-label classification network based on six basic emotions. To be specific, we first use RetinaFace to detect faces and crop the corresponding images. Then, we input these face images to train the pretrained MAE encoder in a multi-label classification task. We optimize the network using cross-entropy, which is defined as follows:

$$\mathcal{L} = -\frac{1}{6}\sum_{j=1}^{6} W_j \left[y_j \log \hat{y}_j + (1 - y_j) \log \left(1 - \hat{y}_j\right) \right], \quad (1)$$

where $\hat{y}$ represents the predicted results for the basic emotions, and y denotes the ground truth values for the basic emotion category.

3.5. Test on compound emotions

The goal of Compound Expression Recognition Challenge is to predict the categories of compound emotions, each of which is a combination of two basic emotions. Therefore, based on the probability values of the basic emotions predicted by the network, we combine them to obtain the corresponding results for compound emotions. In final, the compound emotion with the highest combined probability value is taken as the final prediction result.

4. Experiment

4.1. Dataset

4.1.1 C-EXPR-DB

For the Compound Expression Recognition challenge, we need to make the predictions on a part of C-EXPR-DB [14] database. C-EXPR-DB is the largest, diverse, in-the-wild audiovisual database with the annotations for continuous dimensions of valence-arousal, speech detection, facial landmarks and bounding boxes, action units, facial attributes and 12 compound expressions. The C-EXPR-DB dataset comprises 400 videos, with a total duration exceeding 13 hours and an approximate total frame count of 200,000 frames. For this challenge, we only use 56 of the videos as the test set, and we consider only 7 compound expressions, which are Fearfully Surprised, Happily Surprised, Sadly Surprised, Disgustedly Surprised, Angrily Surprised, Sadly Fearful and Sadly Angry.

4.1.2 RAF-DB

RAF-DB [27, 45] is a large-scale facial expression database in the wild. It contains 29,672 real-world images labeled with different expressions, age range, gender and posture features. It defines two challenging benchmark experiments: 7-class basic expression classification named RAF-DB-B and 11-class compound expression classification name RAF-DB-C. The categories in RAF-DB-C are composed of combinations of two among the six basic emotions included in RAF-DB.

4.1.3 Aff-wild2

Aff-wild2 [15] contains around 600 videos annotated with AU, base expression category, and VA. For the training of CE track, we mainly utilize the data with the base expression category (anger, disgust, fear, happiness, sadness, surprise). Aff-wild2 includes 548 videos with this kind of annotation. And 318 of them have accessible labels for the EXPR track of the 6th ABAW. We use this part of the data and add them to the multi-label task training for the CE track.

4.1.4 Fuxi-EXPR

In addition to publicly available datasets, we also utilize a private dataset Fuxi-EXPR that contains 71,618 facial images collected from the common video websites. We utilize Fuxi Youling Crowdsourcing platform for data management and annotation. The labeling categories include six basic emotions (Surprised, Fear, Disgust, Happiness, Sadness, Anger) and an additional category for Others. During training, we retain only the data for the six basic emotions, totaling 25,457 images.

4.2. Experimental Setting

We extract frames from the videos in C-EXPR-DB using OpenCV, and then employ RetinaFace [2] for face detection and cropping the faces. All face images are resized to 224×224 before being fed into the network. Our experimental code is implemented using the PyTorch framework, and training is conducted on NVIDIA A30 GPUs. The optimizer used during training is AdamW [34], with an initial learning rate of 0.0001. We set the size of mini-batch to

Val. Set	Suprised	Fear	Disgust	Happiness	Sadness	Anger	Average
fold-1	0.6791	0.6923	0.7587	0.7675	0.5644	0.6991	0.7102
fold-2	0.5899	0.6262	0.7239	0.8641	0.5981	0.6693	0.6786
fold-3	0.5742	0.6262	0.7215	0.8613	0.5870	0.6693	0.6733
fold-4	0.6689	0.6834	0.7547	0.8673	0.6501	0.6867	0.7185
fold-5	0.5833	0.6307	0.7309	0.8669	0.5801	0.6804	0.6787

Table 1. The F1 scores for the classification of 6 basic emotions from models trained on different folds.

Method	MAE Pretraining	Distillation	EMA	Mixup	F1 score
Ours	×	×	×	×	0.6010
Ours	✓	×	×	×	0.6484
Ours	✓	✓	×	×	0.6579
Ours	✓	✓	✓	×	0.6614
Ours	✓	✓	✓	✓	**0.6745**

Table 2. Comparisons of our methods with different settings on 6 basic emotions classification.

Rank	Teams	F1 score
#1	**Netease Fuxi AI Lab**	**0.5526**
#2	HSEmotion [44]	0.2708
#3	USTC-IAT-United [57]	0.2240
#4	SUN_CE [43]	0.2201
#5	USTC-AC [52]	0.1845

Table 3. Final competition results of the Compound Expression Recognition Challenge. The F1 scores are evaluated on the test set of C-EXPR-DB.

64. During the training process, we adjust the learning rate using the Cosine Annealing [33] policy with the minimum learning rate of 1e-8 and the number of restart epochs of 5. Additionally, the duration of training for each model is regulated by an early-stopping mechanism, which is set to trigger after a patience period of 10 epochs. During training, we implement the Exponential Moving Average (EMA) strategy, utilizing a decay rate of 0.999, to bolster the stability of the training process. Furthermore, we employ a variety of image augmentation techniques to improve the model's generalization capabilities, including RandomResizedCrop, RandomHorizontalFlip, RandomRotation, and ColorJitter.

4.3. Metrics

For the Compound Expression Recognition Challenge, the performance measure P is the average F1 Score across all 7 categories, calculated as:

$$\begin{cases} F1 = \dfrac{2 \times Precision \times Recall}{Precision + Recall}; \\ Precision = \dfrac{TP}{TP + FP}; \\ Recall = \dfrac{TP}{TP + FN}, \end{cases} \tag{2}$$

$$P_{CE} = \frac{\sum_{c=1}^{7} F1_c}{7} \tag{3}$$

Here, c represents the category ID, TP represents True Positives, FP represents False Positives, and FN represents False Negatives.

4.4. Results

Validation results. As previously mentioned, to utilize more data, we converted the single-label classification problem of multiple emotions into a multi-label classification problem for individual emotions. Training was conducted on a dataset composed of a mixture of RAF-DB-B, Aff-wild2, and FUXI-EXPR. To further enhance the generalization ability and test the models' performance, we use 5-fold cross-validation to train multiple models and then ensemble them.

To further enhance the generalization ability and test the performance of models, we employ a 5-fold cross-validation strategy for training multiple models followed by their ensemble. Specifically, we divide each of the three datasets into five folds and then randomly select one fold from each to form the validation set, with the remaining parts combined to serve as the training set. The results can be found in Table 1. In our experiments, we find that the F1 score for Happiness is the highest. This is because happiness is the only positive emotion among the target emotions, and there is also more data available for it.

Ablation study. As can be seen in Table 2, we also conduct extensive experiments with different settings to further investigate the effectiveness of different components. It is evident that pretraining with large-scale self-reconstruction based on face images is very important. Without this step, the F1 score could only reach 0.6010 in a 6-classification

task. However, adding it can significantly improve the performance of emotion classification, achieving an F1 score of 0.6484. Additionally, we explored various techniques during training. Model distillation and mixup are also effective means to enhance performance. During training, we also employ the Exponential Moving Average (EMA) strategy with a decay rate of 0.999 to enhance the stability of the training process. In final, on the task of classifying the six basic emotions, the F1 score reached 0.6745.

Competition results. In the Compound Expression (CE) Recognition Challenge of ABAW6, we need to recognize the 7 compound expressions for each frame in a part of C-EXPR-DB database, which contains 56 videos and 26145 frames. As can be seen in Table 3, our method achieves a average F1 score of 0.5526 and wins the first place in the CE Recognition challenge. Our score is significantly higher than those of the second place. We attribute this achievement to the transformation of the task from compound emotion classification into multi-label basic emotion classification, allowing us to utilize a large amount of available data for training. Additionally, we annotate a batch of high-quality data based on the Fuxi Youling crowdsourcing platform, which helps us train more generalizable models. Furthermore, our ViT encoder, pretrained through a large-scale self-supervised face reconstruction task, also demonstrates better generalizability. It can extract more effective facial features, thereby enhancing the performance of downstream emotion recognition tasks.

5. Conclusion

In this paper, we have extended the inquiry into the applicability of task-agnostic, large-scale pre-training methodologies to the distinct domain of Compound Expression Recognition (CER). By implementing a framework that leverages a Facial Expression Masked AutoEncoder (FE-MAE) pre-trained on extensive datasets of facial expressions and subsequently fine-tuned for compound emotion detection, we have evidenced the potential for cross-domain adaptation of these advanced pre-training techniques. We have fine-tuned the pre-trained model with limited domain-specific data, demonstrating promising results in transfer capabilities for the nuanced task of CER. The performance metrics, obtained through rigorous testing on the C-EXPR-DB dataset and recognized at the 6th ABAW, underscore the efficacy of our approach, while simultaneously highlighting avenues for future enhancements.

6. Acknowledgments

The experiments and the data management and storage are supported by NetEase Fuxi Youling platform, based on Fuxi Agent-Oriented Programming (AOP) system that is carefully designed to facilitate task modeling. This work is also supported by the National Key R&D Program of China (No. 2022YFF09022303).

References

[1] Brian Cheung. Convolutional neural networks applied to human face classification. In *2012 11th International Conference on Machine Learning and Applications*, pages 580–583. IEEE, 2012. 2

[2] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 4

[3] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. From individual to group-level emotion recognition: Emotiw 5.0. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 524–528, 2017. 2

[4] Heming Du, Xin Yu, and Liang Zheng. Learning object relation graph and tentative policy for visual navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 19–34. Springer, 2020. 2

[5] 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. 3

[6] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 2

[7] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016. 3

[8] Xiaoyu Feng, Heming Du, Hehe Fan, Yueqi Duan, and Yongpan Liu. Seformer: Structure embedding transformer for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 632–640, 2023. 2

[9] 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 *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013. 2

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3

[11] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings*

of the 15th ACM on International conference on multimodal interaction, pages 543–550, 2013. 2

[12] MD Wahiduzzaman Khan, Hongwei Sheng, Hu Zhang, Heming Du, Sen Wang, Minas Coroneo, Farshid Hajati, Sahar Shariflou, Michael Kalloniatis, Jack Phu, et al. Rvd: a handheld device-based fundus video dataset for retinal vessel segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[13] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. *arXiv preprint arXiv:2207.01138*, 2022. 1, 4

[14] Dimitrios Kollias. Multi-label compound expression recognition: C-expr database & network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2023. 2, 3, 4

[15] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 4

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

[17] 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. 4

[18] 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.

[19] 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.

[20] 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.

[21] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019.

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

[23] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. *arXiv preprint arXiv:2303.01498*, 2023. 4

[24] Dimitrios Kollias, Panagiotis Tzirakis, Alan Cowen, Stefanos Zafeiriou, Chunchang Shao, and Guanyu Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. *arXiv preprint arXiv:2402.19344*, 2024. 1

[25] Ashu Kumar, Amandeep Kaur, and Munish Kumar. Face detection techniques: a review. *Artificial Intelligence Review*, 52:927–948, 2019. 2

[26] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995. 2

[27] 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. 3, 4

[28] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: bootstrapping audio-visual segmentation by integrating foundation knowledge. *arXiv preprint arXiv:2308.10175*, 2023. 2

[29] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7590–7598, 2023. 2

[30] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. A neural micro-expression recognizer. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–4. IEEE, 2019. 2

[31] Yuchi Liu, Zhongdao Wang, Tom Gedeon, and Liang Zheng. How to synthesize a large-scale and trainable micro-expression dataset? In *European Conference on Computer Vision*, pages 38–55. Springer, 2022. 2

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 3

[33] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[35] Bowen Ma, Rudong An, Wei Zhang, Yu Ding, Zeng Zhao, Rongsheng Zhang, Tangjie Lv, Changjie Fan, and Zhipeng Hu. Facial action unit detection and intensity estimation from self-supervised representation. *arXiv preprint arXiv:2210.15878*, 2022. 3

[36] Brais Martinez and Michel F Valstar. Advances, challenges, and opportunities in automatic facial expression recognition. *Advances in face detection and facial image analysis*, pages 63–100, 2016. 2

[37] 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. 3

[38] Xingqun Qi, Chen Liu, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *arXiv preprint arXiv:2305.18891*, 2023. 2

[39] Feng Qiu, Bowen Ma, Wei Zhang, and Yu Ding. Multimodal emotion reaction intensity estimation with temporal augmentation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 5776–5783, 2023. 2

[40] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018. 3

[41] Henry A Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 20(1):23–38, 1998. 2

[42] 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*, 115:211–252, 2015. 2

[43] Elena Ryumina, Maxim Markitantov, Dmitry Ryumin, Heysem Kaya, and Alexey Karpov. Audio-visual compound expression recognition method based on late modality fusion and rule-based decision. *arXiv preprint arXiv:2403.12687*, 2024. 5

[44] Andrey V Savchenko. Hsemotion team at the 6th abaw competition: Facial expressions, valence-arousal and emotion intensity prediction. *arXiv preprint arXiv:2403.11590*, 2024. 5

[45] Li Shan and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2018. 4

[46] Xin Shen, Shaozu Yuan, Hongwei Sheng, Heming Du, and Xin Yu. Auslan-daily: Australian sign language translation for daily communication and news. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[47] Hongwei Sheng, Xin Yu, Feiyu Wang, MD Wahiduzzaman Khan, Hexuan Weng, Sahar Shariflou, and S Mojtaba Golzan. Autonomous stabilization of retinal videos for streamlining assessment of spontaneous venous pulsations. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023. 2

[48] V Betcy Thanga Shoba and I Shatheesh Sam. A hybrid features extraction on face for efficient face recognition. *Multimedia tools and applications*, 79(31):22595–22616, 2020. 2

[49] Yichuan Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013. 2

[50] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57:137–154, 2004. 2

[51] Hongjun Wang, Jiani Hu, and Weihong Deng. Face feature extraction: a complete review. *IEEE Access*, 6:6001–6039, 2017. 2

[52] Jiahe Wang, Jiale Huang, Bingzhao Cai, Yifan Cao, Xin Yun, and Shangfei Wang. Zero-shot compound expression recognition with visual language model at the 6th abaw challenge. *arXiv preprint arXiv:2403.11450*, 2024. 5

[53] Biao Yang, Jinmeng Cao, Rongrong Ni, and Yuyu Zhang. Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. *IEEE access*, 6:4630–4640, 2017. 2

[54] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 2

[55] Peng Yao, Huaqiang Wu, Bin Gao, Sukru Burc Eryilmaz, Xueyao Huang, Wenqiang Zhang, Qingtian Zhang, Ning Deng, Luping Shi, H-S Philip Wong, et al. Face classification using electronic synapses. *Nature communications*, 8 (1):15199, 2017. 2

[56] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 3

[57] Jun Yu, Jichao Zhu, and Wangyuan Zhu. Compound expression recognition via multi model ensemble. *arXiv preprint arXiv:2403.12572*, 2024. 5

[58] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *European conference on computer vision*, pages 318–333. Springer, 2016. 2

[59] Xin Yu and Fatih Porikli. Face hallucination with tiny unaligned images by transformative discriminative neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2017.

[60] Xin Yu and Fatih Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3760–3768, 2017.

[61] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European conference on computer vision (ECCV)*, pages 217–233, 2018.

[62] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 908–917, 2018.

[63] Xin Yu, Fatemeh Shiri, Bernard Ghanem, and Fatih Porikli. Can we see more? joint frontalization and hallucination of unaligned tiny faces. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2148–2164, 2019. 2

[64] 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. 1

[65] Lei Zhang, Meng Yang, Xiangchu Feng, Yi Ma, and David Zhang. Collaborative representation based classification for face recognition. *arXiv preprint arXiv:1204.2358*, 2012. 2

[66] Wei Zhang, Feng Qiu, Suzhen Wang, Hao Zeng, Zhimeng Zhang, Rudong An, Bowen Ma, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2428–2437, 2022. 2

[67] Wei Zhang, Lincheng Li, Yu Ding, Wei Chen, Zhigang Deng, and Xin Yu. Detecting facial action units from global-local fine-grained expressions. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1

[68] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Facial affective analysis based on mae and multi-modal information for 5th abaw competition. *arXiv preprint arXiv:2303.10849*, 2023. 2

[69] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Multi-modal facial affective analysis based on masked autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2023. 3

[70] Wei Zhang, Feng Qiu, Chen Liu, Lincheng Li, Heming Du, Tiancheng Guo, and Xin Yu. Affective behaviour analysis via integrating multi-modal knowledge. *arXiv preprint arXiv:2403.10825*, 2024. 2, 3, 4