Uncertain Facial Expression Recognition via Multi-task Assisted Correction

Yang Liu, Xingming Zhang, Janne Kauttonen, and Guoying Zhao*, *Fellow, IEEE*

***Abstract*—Deep models for facial expression recognition achieve high performance by training on large-scale labeled data. However, publicly available datasets contain uncertain facial expressions caused by ambiguous annotations or confusing emotions, which could severely decline the robustness. Previous studies usually follow the bias elimination method in general tasks without considering the uncertainty problem from the perspective of different corresponding sources. This paper proposes a novel method of multi-task assisted correction in addressing uncertain facial expression recognition called MTAC. Specifically, a confidence estimation block and a weighted regularization module are applied to highlight solid samples and suppress uncertain samples in every batch. In addition, two auxiliary tasks, i.e., action unit detection and valence-arousal measurement, are introduced to learn semantic distributions from a data-driven AU graph and mitigate category imbalance based on latent dependencies between discrete and continuous emotions, respectively. Moreover, a re-labeling strategy guided by feature-level similarity constraint further generates new labels for identified uncertain samples to promote model learning. The proposed method can flexibly combine with existing frameworks in a fully-supervised or weakly-supervised manner. Experiments on five popular benchmarks demonstrate that the MTAC substantially improves over baselines when facing synthetic and real uncertainties and outperforms the state-of-the-art methods.**

***Index Terms*—Facial Expression Recognition, Uncertainty, Action Unit, Valence, Arousal, Multi-task Learning.**

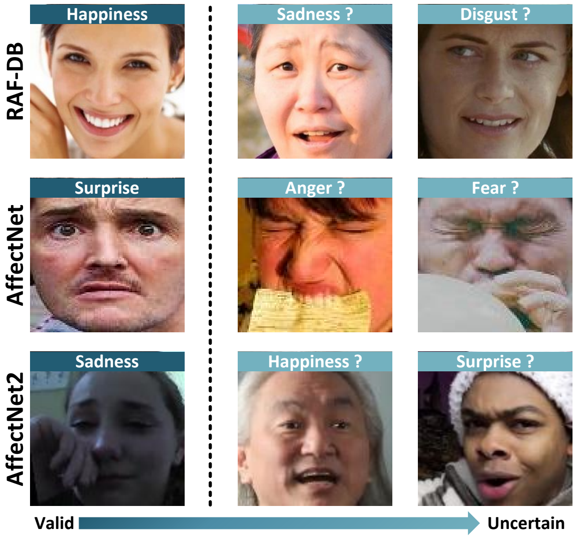


Fig. 1. Uncertainty in RAF-DB, AffectNet, and AffWild2 benchmarks. Top texts indicate their original labels. Uncertainty commonly exists in different facial expression datasets.

I. Introduction

FACIAL expressions carry essential information for perceiving human emotions and attitudes in daily communications. Automatic facial expression recognition (FER) from visual signals of images and videos is a vital technology for realizing human-computer systems such as remote health care, virtual reality, and social robots. Due to sufficient labeled data and high-speed computation resources, deep learning models have achieved excellent performance and dominated the FER research recently [1], [2], [3], [4], [5].

High-quality annotated images are significant when developing a FER method. Early facial expression databases (e.g., CK+ [6] and Oulu-CASIA [7]) usually recruit a small scale of subjects and collect their facial expressions in a lab-constrained environment. Due to the limited number and conditions, experts can annotate the data carefully and precisely. To meet the requirement of massive labeled samples for training a deep FER model, recently released datasets (e.g., RAF-DB [8], AffectNet [9], and AffWild2 [10]) gather images from the Internet. Keeping the labeling consistent in a large-scale manner is hard for those in-the-wild databases. Therefore, a lot of annotations could be uncertain or even wrong, which might cause two undesired effects on model training. Firstly, the considerable number of uncertain facial expressions in the training set will arouse and strengthen the over-fitting. Secondly, the uncertain samples will misguide the direction of model learning and decline the final performance.

According to the source, there are two categories of uncertainty in the FER task. The first one is subjective annotation. Labels for existing facial expression datasets are voted by annotators recruited on crowdsourcing platforms [8]. These annotators usually need to gain expertise and will assign different labels for the same image based on their backgrounds, especially for facial expressions under in-the-wild scenes. Fig. 1 presents a few images in RAF-DB, AffectNet, and AffWild2 to illustrate the prevalence of uncertain samples. Facial expressions in the left column are typical for people to make consistent annotations. However, in the other two columns, non-typical behaviors and occlusions in the wild scenarios could cause different opinions about the same face.

The second source is intrinsic confusion. Previous FER

* Corresponding author

Y. Liu and G. Zhao are with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland, FI-90014 e-mail: Yang.Liu@oulu.fi, Guoying.Zhao@oulu.fi

X. Zhang is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China, 510006 e-mail: cxzxm@scut.edu.cn

J. Kauttonen is with the Haaga-Helia University of Applied Sciences, Helsinki, Finland, FI-00520 e-mail: Janne.Kauttonen@haaga-helia.fi

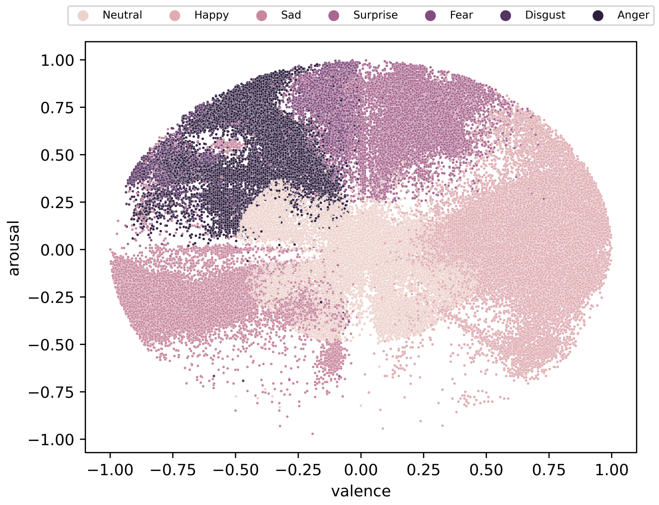


Fig. 2. Distribution of discrete labels and continuous labels in AffectNet's training set. Categories with few samples are more likely to be confused with other classes. Many samples are far from the original category center but close to other clusters, indicating great uncertainty.

methods commonly make coarse category predictions (e.g., six basic emotions). Nevertheless, expression displays in the real world are diverse and spontaneous because of different inducements, postures, and contexts [11], [12]. Some facial expressions are composed of compound or non-typical emotions in unconstrained conditions, which are challenging to be allocated to a discrete category. This phenomenon worsens when encountering imbalance categories in the database, as shown in the distribution visualization of training labels in the AffectNet dataset (see Fig. 2).

To this end, a few studies have proposed solutions to alleviate the uncertainty problem. Most of them focused on migrating the methods of handling data noise in general tasks [13], [14]. Generally, a specific block for uncertainty estimation will be introduced to weight or relabel every sample during the model training [15]. Recently, considering characteristics of the FER task in terms of the variety of annotations and the inter-connectivity of sub- or similar tasks, introducing the interrelation within multiple annotations such as action units (AUs) and valence-arousal (VA) has been explored [16], [17]. However, the uncertainty of facial expressions still plagues these methods due to the following issues: 1) different types of uncertainty are lumped together without targeted treatment; 2) additional multi-label knowledge is just exploited at the label space instead of the feature space; 3) the relabeling solution without semantic preserving is simple, declining the confidence of the proxy annotation.

This paper develops a new framework to perform uncertain FER via **M**ulti-**T**ask **A**ssisted **C**orrection, called MTAC. It contains three parts: a target branch and two auxiliary branches. First, a backbone network is applied to extract facial features for every training batch. Then, a weighted regularization block executes confidence measurement for each sample and motivates the model to prioritize images that have dependable annotations in the target branch. Based on a parameter-sharing backbone, the VA estimation task is introduced to jointly supervise the feature learning with a consideration of category imbalance, while the AU detection task is conducted by adding a graph convolution block and extracting the semantic representation of every sample. When samples are deemed highly uncertain, we assess their semantic representation against memory templates and re-assign their labels based on the criterion of feature-level similarity.

This work is an extension of our preliminary investigation presented on ICPR 2022 [18], where we proposed a new uncertain FER label correction method based on auxiliary AU graphs. In contrast to the previous iteration, we have improved this paper in four areas: 1) we involve a new auxiliary task of VA estimation for collaborative model training, which can address the uncertainty caused by the biases of discrete labels on describing in-the-wild facial expressions; 2) we revise original loss functions and design a new weighted loss for handling data imbalance, which can jointly optimize the feature extractor; 3) we construct a new memory template of weighted semantic centers and improve the relabeling strategy, which can adaptively generate pseudo labels for uncertain samples; 4) To conduct a more extensive evaluation of our approach, we leverage supplementary backbones and benchmarks. In summary, the principal contributions of this paper are:

- MTAC method quantifies the sample confidence and suppresses the effects of uncertain discrete labels during the model training.
- MTAC mitigates the category imbalance and facilitates feature learning on ambiguous facial expressions with continuous labels in the auxiliary VA estimation task.
- MTAC examines the semantic representation derived from the auxiliary AU graph and performs relabeling for uncertain samples while adhering to a feature-level constraint.
- Our MTAC approach is demonstrated to be highly effective in addressing the problem of uncertainty through extensive experiments on five large-scale benchmarks, outperforming state-of-the-art methods with superior performance.

The remainder of this article is structured as follows: Section II provides an overview of various related studies, Section III expounds on the MTAC approach, Section IV presents the experimental results and discussions, and finally, Section V summarizes and concludes this study.

II. Related Work

This section briefly summarizes the current progress of the FER research regarding multi-task facial expression analysis, graph-based affective representation, and deep learning with uncertainty.

A. *Multi-Task Facial Expression Analysis*

Automatically predicting basic emotions is the main task in traditional FER studies [19], [20]. From psychological findings, advanced emotional description models have been utilized to annotate a broader range of facial expressions, such as AUs [21] and VA [22]. Therefore, recent studies have explored combining multiple tasks for a generalized feature extractor of facial images and videos [23]. Zhang *et al.* [24] devised a cohesive adversarial learning scheme that links the predicted emotions and the corresponding distribution of continuous annotations. Similarly, Antoniadis *et al.* [25] captured the dependencies between categorical and dimensional emotions through a graph convolutional network (GCN). Cui *et al.* [16] captured the association between labels at the object and property level, serving as a basis for updating and generating labels for novel datasets. Chen and Joo [26] integrated the *triplet* loss to encode the interdependence between AUs and facial expressions. Besides emotion-related tasks, other close facial tasks like facial landmark detection have provided additional information to facial expression analysis. Chen *et al.* [17] addressed label inconsistency by incorporating landmark detection as a neighboring task and utilizing the cluster distribution. Toisoul *et al.* [27] developed a network that combined facial landmarks with discrete and continuous emotions, utilizing an attention mechanism based on fiducial points. In contrast, our approach utilizes auxiliary AU detection and VA estimation tasks to improve uncertainty correction. Each auxiliary task can be independently integrated during the model training without causing extra burden in the testing stage.

B. Graph-based Affective Representation

Effective facial representations are vital for FER methods [28], [29]. Graph-based methods have emerged as a recent solution, owing to their ability to capture the anatomy and semantic associations among different facial regions concurrently, which are considered crucial clues of human facial perception [30]. Liu *et al.* [31] designed a graph representation of facial expressions that consisted of reasonable facial landmarks and semantic connections, which modeled critical appearance and geometric facial changes. Zhao *et al.* [32] constructed a geometric graph description of facial components more robust to appearance variations like texture noise and light changes. Besides facial landmarks, many studies generate graph representations based on local facial regions. Jin *et al.* [33] cropped 20 local facial areas as graph nodes and linked edges according to a trainable weighted adjacency matrix to exploit intra- and inter-regional relationships. Xie *et al.* [34] correlated a cross-domain graph for global-local feature adaptation to learn invariant representations of facial expressions. Alternatively, graphs constructed from the perspective of AUs are also explored. Luo *et al.* [35] learned a unique graph describing the dependence within each AU pair, comprising its activation level and its correlation with other AUs. Song *et al.* [36] transferred hybrid messages among AUs and inferred possible graph structures to provide complementary information for higher performance. This work focuses on AU graphs where the extracted semantic representation is used to constrain the re-labeling strategy. Compared to existing methods, our AU graph is constructed using a data-driven approach based on golden or automatic annotations.

C. Deep Learning with Label Uncertainty

Label uncertainty is a common and significant problem in FER and plagues deep models for many general tasks [37], [38], [39]. Machine learning researchers usually regard uncertainty as a noisy label issue and rely on modified loss functions to penalize it. Zhong *et al.* [40] propagated the uncertain signals across a confidence graph based on feature similarity and temporal consistency that were used to train a label noise cleaner. Li *et al.* [14] regularized the low-dimensional subspace of embedded images by a consistency loss and a prototypical loss so that alleviated uncertain samples with a neighboring constraint. Analogously, in the FER field, Wang *et al.* [41] introduced a self-cure model that assigned weights to facial images to reduce the effect of suspicious samples by finding and recovering incorrect annotations. She *et al.* [13] utilized a series of mini branches to tackle label and instance space ambiguity with the pairwise distribution. Additionally, Zhang *et al.* [42] developed a weakly-supervised noise estimation method to learn the correlation between feature space and the difference from clean annotations. Gu *et al.* [15] suppressed the label and feature noise by leveraging a multivariate normal distribution and preserving the inter-class correlations. As mentioned before, the FER task suffers from various uncertainties, which need to be fully considered. To this end, we combine multi-task and distribution learning to address subjective annotation and intrinsic confusion problems in this paper. The proposed MTAC can adaptively suppress or correct uncertain samples during the modeling training.

III. Proposed Method

As mentioned above, the uncertainty in large-scale FER datasets comes from two aspects, i.e., subjective annotation and intrinsic confusion. To this end, we need to know which samples are uncertain to reduce their impact on model training and correct them to use existing data fully. Inspired by the previous work [17], [24], we propose leveraging multi-task learning and distribution learning methodologies to address uncertain FER. This work has two bases: 1) features of one sample extracted on similar tasks are correlated, and 2) comparable samples ought to exhibit a shared interdependence between their label space and feature representation. This section provides a pipeline of the MTAC and expounds on its critical components.

A. Overview of MTAC

Fig. 3 provides an outline of the MTAC, which comprises the following components: 1) a target branch that utilizes facial features obtained from a pre-trained backbone network and evaluates the certainty of annotations through a self-attention block. These confidence scores determine the samples' significance during the classification loss computation. 2) one auxiliary branch of the VA estimation task jointly supervises the feature learning accompanying the class-oriented loss to simultaneously deal with the uncertainty of intrinsic confusion and class imbalance in the present batch. 3) the other auxiliary branch of the AU detection task constructs data-driven AU graphs, generates a memory template of semantic centers for every emotion category, and then relabels suspicious samples based on the rank regularization and the similarity preserving constraint. The entire MTAC framework is end-to-end, and the

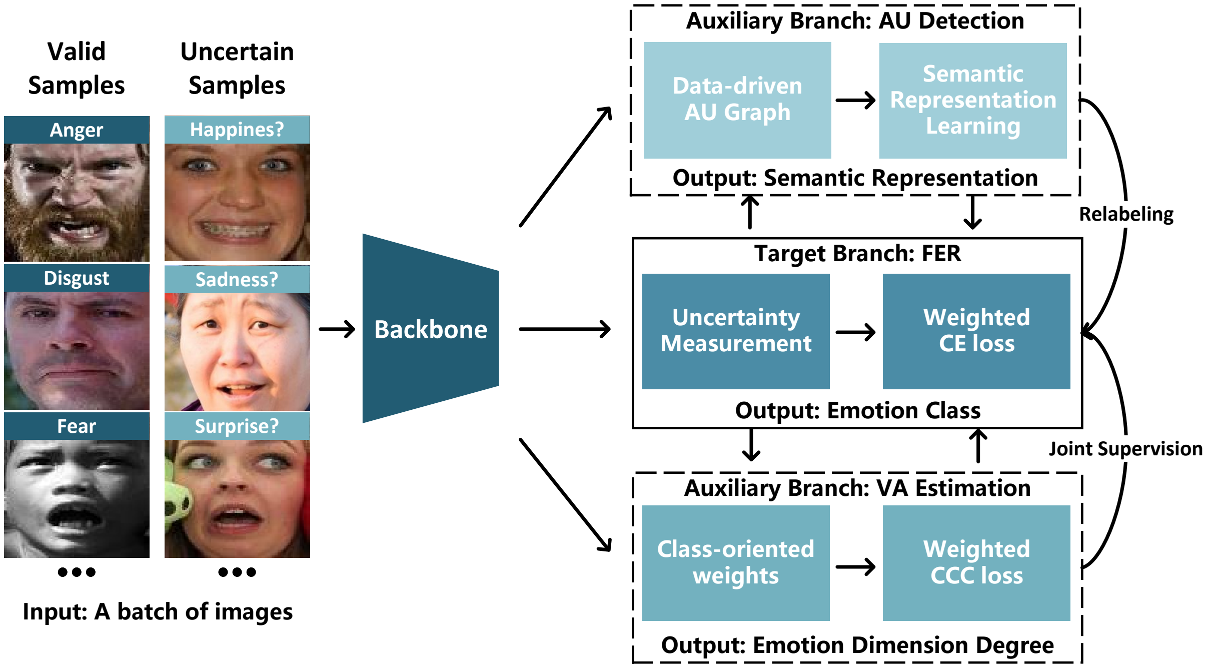


Fig. 3. The outline of MTAC. It has a target branch for FER, one auxiliary branch for VA estimation, and one auxiliary branch for AU detection. The VA branch supports the feature learning of the target branch by using continuous emotion labels and considering category imbalance, while the AU branch relabels extremely uncertain samples based on the semantic similarity constraint. Samples are re-annotated if they appear semantically closer to another category center than the original one in the feature space. Both auxiliary branches are free to work or disabled and will not be involved in the testing phase.

two auxiliary branches can operate separately or jointly and will not be involved in the testing phase.

B. Target Branch with Uncertainty Measurement

Before handling the uncertainty, we want the model to provide confidence for each input while making the prediction. As illustrated in Fig. 4, our target branch follows a broad pipeline with a feature extractor and a classifier for the FER task. Let $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_N] \in \mathbb{R}^{D\times N}$ denote the facial features obtained from a pre-trained backbone network for a batch of N images, where D represents the dimension of each feature. To detect the uncertain samples and estimate their levels of uncertainty, we utilize a self-attention block inspired by [41], [43], which comprises a fully connected (FC) layer and the sigmoid function. Formally, we can compute the confidence score of the i-th sample as follows:

$$\alpha_i = Sigmoid(\boldsymbol{W}_a^\top \boldsymbol{f}_i), \tag{1}$$

where the parameters of the self-attention layer is denoted by $\boldsymbol{W}_a^\top$. Benefiting from the pre-trained backbone network, it is effective to learn an effectual confidence score for each sample in one training batch through this simple block.

During feature learning, it is desirable for images with lower confidence to have a lesser impact, while samples with higher confidence should have greater attention within the current batch. We adopted a weighted Cross-Entropy (CE) loss like [13], [41] to achieve this. The loss of the FER classifier is computed explicitly as follows:

$$L_{wce} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\alpha_i \boldsymbol{W}_{y_i}^\top \boldsymbol{f}_i}}{\sum_{j=1}^{C} e^{\alpha_i \boldsymbol{W}_j^\top \boldsymbol{f}_i}}, \tag{2}$$

where $\boldsymbol{W}_j^\top$ represents the parameters of the j-th classifier, f_i refers to the facial feature, C and y_i are the number of classes and the original discrete label, respectively. According to [44], L_{wce} and α are positively correlated.

C. Auxiliary VA Estimation Branch with Category Balancing

We exploit the VA estimation task as an auxiliary branch to mitigate the uncertainty of intrinsic confusion and complement the biases of discrete emotion labels. As shown in Fig. 5, the VA estimation branch shares the same backbone network as the target branch. However, it replaces the final classifier with a VA regressor for continuous predictions. Specifically, we choose the Concordance Correlation Coefficient (CCC) [45] as our metric here because it reflects both the trend and the error between the dimensional label and the regressed value, which can be computed as:

$$\rho = \frac{2\sigma_{y\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}, \tag{3}$$

where y and $\hat{y}$ represent the continuous label and prediction, respectively, μ and σ denote the mean and variance, respectively, and $\sigma_{y\hat{y}}$ indicates the covariance of y and $\hat{y}$.

In addition, considering the heavy category imbalance in existing FER datasets, the class-oriented weight is presented as follows:

$$\gamma_j = 1 - \frac{N_j}{N}, j \in \{1, 2, ..., C\}, \tag{4}$$

where C denotes the total number of categories, N_j and N represent the count of images in class j and the number of all samples in the current batch, respectively. The network ought

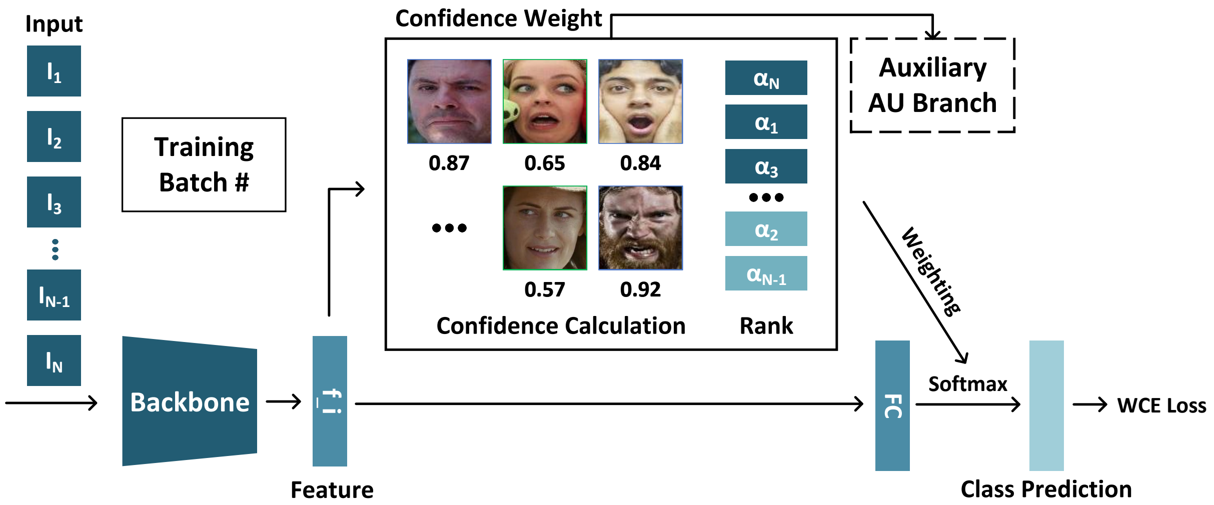


Fig. 4. The pipeline of the target branch. Given a training batch, the confidence score of every sample is calculated by applying a self-attention block and is then used to suppress the uncertainty in the loss function. These confidence scores are further passed to the auxiliary AU branch for semantic representation learning and memory template establishment.

to focus more on categories with fewer samples. The class-oriented weight helps to avoid the situation where the training network converges to major classes faster than minor classes [37]. Therefore, we propose a weighted CCC loss function for the VA estimation task:

$$L_{w3c} = \sum_{j=1}^{C} \gamma_j (1 - \frac{\rho_j^v + \rho_j^a}{2}), \tag{5}$$

where ρ_j^v and ρ_j^a denote the valence CCC and the arousal CCC of the j-th category, respectively. We put this category balancing on the VA branch rather than the target branch for two reasons: 1) feature learning in dimensional emotion estimation is not influenced by imbalanced discrete labels; 2) small categories have higher uncertainties of the subjective annotation and larger intra-class distances as shown in Fig. 2. Alternatively, γ can also be added to Eq. 2 similarly to prevent uncertainty from category imbalance when the VA branch is disabled.

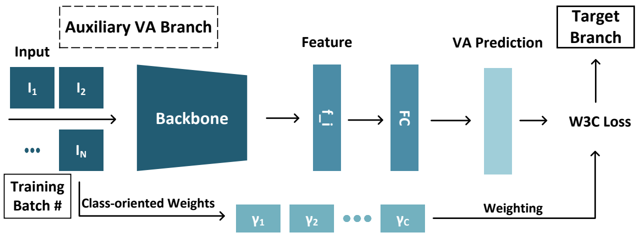


Fig. 5. The pipeline of the auxiliary VA branch. Continuous emotion labels jointly train the parameter-sharing backbone network for better facial feature learning. Furthermore, the class-oriented weight is calculated to address the issue of imbalanced categories in discrete labels.

D. Auxiliary AU Detection Branch with Graph Reasoning

Although the uncertainty is significantly alleviated with the help of the above two branches, low-confidence samples such as those incorrectly labeled will still decrease the recognition performance. Meanwhile, the learned confidence score is a pre-screen step of uncertainty measurement without relevant supervision information, which is not solid enough. Therefore, the AU detection task is further employed as the other auxiliary branch as the Facial Action Coding System is proven to have latent mappings with emotion categories [21], [46], [30]. As illustrated in Fig. 6, we could extract a series of AU features from every sample by utilizing the backbone network, denoted as $\boldsymbol{X}^i = [\boldsymbol{x}_1^i, \boldsymbol{x}_2^i, ..., \boldsymbol{x}_M^i] \in \mathbb{R}^{B \times M}$, where B and M indicate the feature vector dimension and the number of AUs, separately.

Due to the difficulty in ensuring consistent mappings between emotion classes and AUs in large-scale databases [31], [47], we construct a data-driven AU graph that considers independent AU features as graph nodes and co-occurring AU dependencies as graph edges. Specifically, our AU graph is established based on the obtained conditional probability of co-occurring AUs in the training data, which can be computed as follows:

$$\boldsymbol{A}_{p,q} = P(AU_p|AU_q) = \frac{OCC_{p\cap q}}{OCC_q}, \tag{6}$$

where $OCC_{p\cap q}$ represents the count of instances where both AU_p and AU_q are present together, and OCC_q refers to the overall frequency of AU_q. As the co-occurring AU dependency is practically asymmetric, so $P(AU_p|AU_q) \neq P(AU_q|AU_p)$.

Subsequently, a two-layer GCN is employed to learn the semantic representation from the AU graph. Formally, each graph convolution layer can be expressed as follows:

$$\boldsymbol{X}' = g(\boldsymbol{X}, \boldsymbol{A}) = LeakyRELU(\bar{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}_g), \tag{7}$$

where $\bar{\boldsymbol{A}}$ represents the normalized $\boldsymbol{A}$ with rows summing to one, and $\boldsymbol{W}_g$ denotes the weight matrix to be learned in the present layer.

Then, all GCN node features are supplied to a FC layer along with sigmoid functions for every AU prediction. Similar to L_{wce}, we enhance the binary cross-entropy loss with the

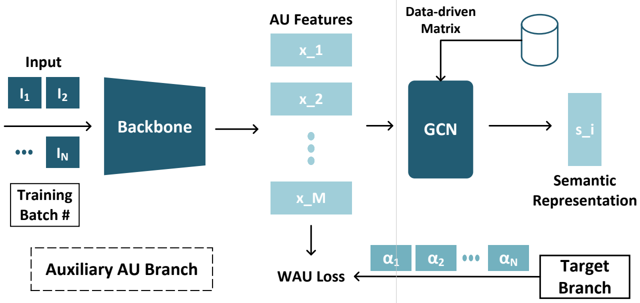


Fig. 6. The pipeline of the auxiliary AU branch. The underlying relationship among AUs is encoded by a data-driven graph from datasets and exploited to generate the semantic representation of each sample.

confidence score to train each AU classifier, and we formulate the overall weighted group loss of the AU branch as:

$$L_{wau} = -\sum_{m=1}^{M} \alpha(z_m \log p_m + (1 - z_m)\log(1 - p_m)), \quad (8)$$

where α denotes the confidence score, z_m and p_m represent the golden/pseudo annotation and the prediction for the m-th AU, separately. Logits $\boldsymbol{s}_i \in \mathbb{R}^{1\times M}$ before the AU classifier serves as the semantic representation of the image.

E. Semantic Similarity Constrained Relabeling

We devise a semantic similarity-constrained relabeling strategy to identify the annotations that require correction and the new categories that need to be assigned (see Fig. 7). For every training batch, a center set for all emotion categories $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_C] \in \mathbb{R}^{M\times C}$ is generated based on the semantic representations and the confidence scores, which can be computed using the following formulas:

$$\boldsymbol{U}_j = \frac{1}{N_j}\sum_{n_j=1}^{N_j} \alpha_{n_j}\boldsymbol{s}_{n_j}, \quad (9)$$

where N_j represents the sample count with the j-th label in the present batch. A memory template $\boldsymbol{T} \in \mathbb{R}^{M\times C}$ is initialized and continuously updated during the entire training procedure as follows:

$$\boldsymbol{T}_j = (1 - e^{-\tau h})\boldsymbol{T}_j + e^{-\tau h}\boldsymbol{U}_j, \tau \in (0, 1], \quad (10)$$

where h denotes the batch index, and τ is a control factor of updating rate. Eventually, the memory template will gradually stabilize as the model converges [48]. After that, the cosine distance between each semantic representation $\boldsymbol{s}_i$ and each of semantic center $\boldsymbol{t}_j$ in $\boldsymbol{T}$ is computed as:

$$dist(\boldsymbol{s}_i, \boldsymbol{t}_j) = 1 - \frac{\boldsymbol{t}_j \cdot \boldsymbol{s}_i}{\|\boldsymbol{t}_j\|\|\boldsymbol{s}_i\|}. \quad (11)$$

Next, for every sample in the current batch, we rank all its semantic distances to the memory template T. Benefiting from the other two branches, the uncertain samples should be suppressed and far from their original category center. Thus, for those samples with extreme uncertainty, we relabel them following the semantic similarity constraint, which could be formulated as follows:

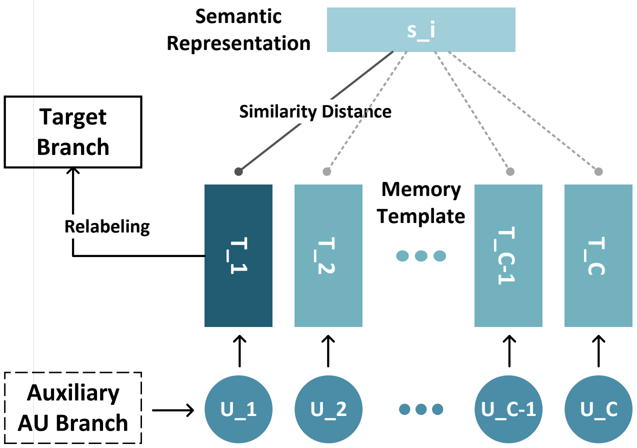


Fig. 7. The pipeline of the relabeling strategy. A memory template is built and updated based on the average category centers and then constrained the relabeling strategy with similarity distance. The new label then participates in network optimization in the target branch.

$$y_i' = \begin{cases} j, & \text{if } dist(\boldsymbol{s}_l, \boldsymbol{t}_{org}) > min(dist(\boldsymbol{s}_l, \boldsymbol{t}_j)) \\ y_i, & \text{otherwise} \end{cases}, \quad (12)$$

where y_i' denotes the pseudo annotation, org indicates the original discrete category, and $j \neq org$. Note that this relabeling strategy will only take effect if the semantic distance to the original category center is not the shortest in the ranking. We assign the sample a new label in such cases by selecting the template class that exhibits the greatest semantic similarity.

F. Network Training

Eventually, the complete loss of the MTAC framework could be expressed as:

$$L_{total} = \lambda_1(L_{wce} + L_{w3c}) + \lambda_2 L_{wau}, \quad (13)$$

where λ_1 and λ_2 are determined using weighted ramp functions that evolve with epoch rounds [49], which can be calculated in the following way:

$$\lambda_1 = \begin{cases} \exp(-(1 - \frac{\beta}{H})^2), & \beta \leq H \\ 1, & \beta > H \end{cases}, \quad (14)$$

$$\lambda_2 = \begin{cases} 1, & \beta \leq H \\ \exp(-(1 - \frac{H}{\beta})^2), & \beta > H \end{cases}, \quad (15)$$

where β denotes the current epoch index, and H is a constant that controls the participation of different branches. Using weighted ramp functions, MTAC can prioritize the AU branch during the initial training phase, where the number of accumulated samples is limited and effective semantic representations and solid memory templates cannot be generated.

As the model undergoes a certain number of training rounds, it progressively focuses on the target and VA branches to extract distinctive features crucial for making accurate predictions. The two auxiliary branches can boost the whole network training with the joint supervision of l_{wce}, l_{w3c}, and

l_{wau} functions. Moreover, our MTAC can work independently with the target branch, while the two auxiliary branches can be flexibly combined into the framework without additional inference burden.

G. Discussion

In this part, we analyze the distinctions between the MTAC approach we propose and other methods for uncertainty-aware FER approaches, such as LDL-ALSG [17], LDLVA [50], SCN [41], DMUE [13], and FENN [15].

1) Difference from LDL-ALSG [17] and LDLVA [50]: LDL-ALSG and LDLVA were two relevant methods that exploit neighborhood knowledge based on k-nearest-neighbor graphs. Nevertheless, the landmark label space used in LDL-ALSG could not keep good accuracy and reflect the actual label distribution, especially on large-scale in-the-wild datasets. Although LDLVA applied pseudo VA annotations and constrained similarities, it led to extra training costs along with the neighbor number of every sample, so that could not perform the efficient uncertain correction. Differently, our MTAC enabled effectively learn semantic feature-level distribution through data-driven AU graphs and adaptively handle uncertain labels with the constructed memory templates without adding too much computation.

2) Difference from SCN [41] and DMUE [13]: Both SCN and DMUE estimated the sample uncertainty. However, the uncertain label cure in SCN was very rough because it relied on the predicted probability without any constraint of extra knowledge, causing inevitable performance degradation. Conversely, DMUE utilized multiple auxiliary branches to alleviate the uncertainty in the label and instance spaces by the latent distribution learning. Nevertheless, it lacked explicit relabeling for extremely uncertain labels, which could lead to weak tolerance when training on large-ratio uncertain data. In contrast, our MTAC initially exploited AU detection and VA estimation tasks separately or collaboratively to assist in uncertainty mitigation. The proposed weighted loss functions, the data-driven semantic dependencies, and the feature-level similarity constraint boosted effective uncertain FER.

3) Difference from FENN [15]: FENN suppressed the uncertainty by exploring the inter-class correlations and extracting compact intra-class descriptions. Nevertheless, it needed to introduce the additional knowledge of neighbor label spaces, which could not achieve label correction during model training. Differently, our MTAC took advantage of multi-task learning, using large-scale training data under both subjective annotation noise and intrinsic confusion noise.

Overall, the proposed MTAC is not a simple combination of existing papers. This work is the first time exploiting the AU graph to realize uncertain label correction based on feature-level similarity preserving. Meanwhile, it is also a pilot study of leveraging continuous emotion labels for uncertain mitigation. The whole framework can be flexibly modified according to the metadata of different datasets. Contributions of all the proposed modules and the improvements against existing methods will be reported in Sec. IV.

IV. Experiments

In this section, we perform extensive experiments to evaluate the effectiveness of MTAC in ablation study, tackling synthetic and real uncertainty and multi-task performance against the state-of-the-art.

A. Datasets

Five challenging FER benchmarks are used for evaluation, i.e., RAF-DB [8], AffectNet [9], AffWild2 [10], FERPlus [51], and SFEW [52]. All five datasets have in-the-wild scenarios and large-scale images containing subjective or intrinsic uncertainty.

RAF-DB comprises $15,339$ in-the-wild images labeled with six basic emotions and neutral. For our studies, we utilized $12,271$ and $3,368$ images for training and testing separately. Since no continuous emotion labels are provided, the VA branch will be disabled in the experiment on RAF-DB.

AffectNet consists of nearly one million facial expression images. We select samples with manual annotations of six basic emotions and neutral for equitable evaluation. Training and test sets comprise $283,901$ and $3,500$ images, separately. Furthermore, we leverage the automatically annotated images in AffectNet as a subset of real noisy data, referred to **AffectNet_Auto**, to evaluate the capacity of MTAC to handle uncertain expressions.

AffWild2 is the first audiovisual dataset with labels for various tasks, including FER, VA estimation, and AU detection. It comprises 558 videos with around 2.8 million images of facial expressions. In this work, we use the subset of 'MTL_Challenge' with seven discrete labels, VA labels, and AU labels simultaneously, which consists of $39,614$ and $10,839$ training and testing samples, respectively.

FERPlus is an enhanced version of the FER2013 [53] dataset. It consists of $28,709$ training images and $3,589$ testing images, and all resized to 48×48 grayscale pixels. Each image is labeled from one of eight categories based on the collaboration of ten crowd-sourced annotators. For a fair evaluation, the category with the highest number of votes is selected as the label for each sample, similar to [13], [41], [54]. Like RAF-DB, the VA branch will be turned off in the experiment on FERPlus.

SFEW consists of 958 training images and 436 testing images extracted from video clips of real-world movies, representing seven basic emotions. The dataset is divided into training, validation, and test groups. We provide the overall accuracy achieved on the testing set to ensure a reliable evaluation. Similar to RAF-DB and FERPlus, the VA branch will be disabled in the experiment on SFEW.

Due to the need for trained experts and the time-consuming nature of AU annotation, no dataset except AffWild2 provides AU labels. To address this limitation, we utilized Openface 2.0 [55] to automatically produce pseudo AU annotations, like in [17], [26]. For AffWild2, the original AU labels are used to generate the AU graph. In other words, the AU branch and the relabeling can work either fully or weakly supervised and are compatible with various datasets. Moreover, the AU branch does not participate in the parameter update of the backbone network. Our MTAC utilizes a feature-level semantic similarity constraint to adaptively correct the extremely uncertain sample instead of directly replacing it with the prediction, which can mitigate the adverse effects of inaccurate pseudo AU annotations.

B. Implementation Configuration

The MTAC is developed using the Pytorch platform and is trained on two Nvidia Volta V100 GPUs. We use cropped facial regions as inputs, resized to 224×224 pixels. The ResNet-18 [56] and the DenseNet-121 [57] pre-trained on the MS-Celeb-1M and SwinTransformer-Small [58] pre-trained on the ImageNet-1K are used as backbone networks, as in previous methods [13], [41], [54]. The initial learning rate for the Adam optimizer is set to 0.01 for the target and auxiliary VA branches, which is then reduced to 10^{-3} and 10^{-4} at the 10-th and 20-th epoch, separately. Every GCN layer has 64 channels in the auxiliary AU branch, with control factor τ and decayed learning rate set as 0.9 and 0.005, respectively. To ensure that all templates are effectively updated during training, a batch size 512 is chosen, while the H defaults to 5. The relabeling starts after 10 epochs to ensure a stable memory template of the semantic representation.

C. Ablation Study

A few ablation studies are performed to verify the contribution of every branch in MTAC and the key hyper-parameter proposed in this paper.

1) Components evaluation: MTAC deals with the effects of uncertain samples based on three branches, i.e., target branch, auxiliary VA branch, and auxiliary AU branch. The target branch suppresses suspicious samples and highlights valid inputs through confidence measurement and weighted loss function. The VA branch optimizes the parameter-sharing network with continuous annotations and considers the category imbalance. The AU branch corrects extremely uncertain labels with the data-driven AU graph and semantic memory templates. Various network architectures can incorporate any of the three branches in a versatile manner. In this study, we propose five different configurations to verify the effectiveness. Please note that on RAF-DB, the target branch employs class-oriented weights owing to the absence of continuous labels. When none of the branches is activated, the network performs like a conventional ResNet-18 model.

Tab. I demonstrates that the individual target branch considerably improves the FER accuracy across three datasets. A more remarkable improvement can be achieved using two auxiliary branches, respectively. In particular, the VA branch performs slightly better than the AU branch on AffectNet and AffWild2 because of the additional knowledge from the continuous label space and the manipulation for the huge category imbalance. The best performance comes from the complete MTAC framework, with all three branches considering the uncertainty from subjective annotation and intrinsic confusion.

Additionally, we perform statistical analysis between the backbone and the MTAC. First, a two-sample F-test is employed to assess the quality of variances, demonstrating that the two data have equal variances. Afterward, a paired two-tailed T-test is utilized to examine the performance difference (H_0 : *There is no performance difference between backbone and MTAC*). As reported in Tab. I, the p-value demonstrates that our MTAC achieves statistically significant improvements among three benchmarks.

TABLE I
EVALUATION OF DIFFERENT BRANCHES. '*Target B.(ranch)*' APPLIES THE UNCERTAINTY MEASUREMENT, AUXILIARY *VA B.(ranch)* EXECUTES THE JOINT FEATURE LEARNING AND THE CATEGORY BALANCING, AND AUXILIARY *AU B.(ranch)* EXPLOITS THE DATA-DRIVEN AU GRAPH AND THE SEMANTIC SIMILARITY CONSTRAINED RELABELING. **BOLD** DENOTES THE BEST RESULT, AND *ITALICS* INDICATES THE SECOND BEST RESULT.

Target B.	*VA B.*	*AU B.*	RAF-DB	AffectNet	AffWild2
×	×	×	85.81	57.94	56.41
✓	×	×	*87.74*	61.97	59.09
✓	✓	×	-	*63.71*	*61.47*
✓	×	✓	**89.31**	62.51	61.16
✓	✓	✓	-	**65.09**	**62.78**
$f-value$			$0.8817 < F_{(.05,2,2)}$		
$p-value$			$0.0181\ (< .05)$		

2) Impact of the class-oriented weight: Most large-scale facial expression datasets have severe category imbalances. In MTAC, the proposed class-oriented weight γ is compatible with various loss functions. In this experiment, we design three different settings, i.e., MTAC without γ, γ in the target branch (as our preliminary work in ICPR 2022 [18]), and γ in the auxiliary VA branch. As presented in Tab. II, the category balancing contributes significantly to the model training. It performs better in the VA branch, demonstrating our statement in Sec. III-C.

TABLE II
EVALUATION OF THE CLASS-ORIENTED WEIGHT. **BOLD** DENOTES THE BEST RESULT, AND *ITALICS* INDICATES THE SECOND BEST RESULT.

Method	RAF-DB	AffectNet	AffWild2
w/o γ	88.45	63.46	61.83
Target B. w/ γ	**89.31**	*64.83*	*62.07*
VA B. w/ γ	-	**65.09**	**62.78**

3) Impact of the data-driven graph: Relabeling under semantic similarity constraints is another essential module of MTAC for uncertainty mitigation. Its semantic information of AU co-occurring dependencies is encoded with a data-driven AU graph. To study the established edges, we randomly initialize A with element values from 0 to 1 to shield edge attributes in this experiment. We also design a fully-connected A that every element is fixed as 1. As shown in Tab. III, the random edges introduce additional uncertainty and lead to performance decreases, while the fixed edges can not reflect the AU co-occurrence and approximate the actual distribution. On the contrary, our data-driven AU graph helps the GCN to generate better semantic representations and further boosts the memory templates and the relabeling.

4) Impact of the pseudo AU labels: Since few existing FER datasets provide manual AU annotations, we exploit pseudo AUs extracted by OpenFace. In other words, the auxiliary AU branch of MTAC can work in a fully supervised way with real AU annotations or in a weakly supervised way with pseudo

TABLE III
EVALUATION OF THE DATA-DRIVEN GRAPH. **BOLD** DENOTES THE BEST RESULT, AND *ITALICS* INDICATES THE SECOND BEST RESULT.

Method	RAF-DB	AffectNet	AffWild2
w/ *random edges*	86.10	62.06	60.88
w/ *fixed edges*	*87.14*	*63.94*	*61.63*
w/ *data-driven edges*	**89.31**	**65.09**	**62.78**

AU labels. We experimented with the AffWild2 dataset using pseudo AU labels to study its effect. In addition, due to the AU label also influencing the AU graph construction, we further replace the original data-driven edges with those computed based on RAF-DB and AffectNet, respectively. As presented in Tab. IV, the real AU annotation significantly contributes to the final FER result. Although pseudo AUs lead to a performance decrease, it still outperforms the model without the auxiliary AU branch, as reported in Tab. I. Despite the change of data-driven edges, the AffectNet AU graph shows a competitive performance compared with pseudo AffWild2 AUs because of its large-scale samples, while the RAF-DB AU graph suffers a degradation.

TABLE IV
EVALUATION OF PSEUDO AU LABELS. **BOLD** DENOTES THE BEST RESULT, AND *ITALICS* INDICATES THE SECOND BEST RESULT.

Method	AffWild2
w/ *RAF-DB AU graph*	60.78
w/ *AffectNet AU graph*	61.61
w/ *pseudo AffWild2 AUs*	*61.80*
w/ *real AUs*	**62.78**

D. Performance Evaluation of Handling Uncertainty

Here, we evaluate the proposed MTAC's ability to deal with uncertain samples. Specifically, we performed comparative tests under synthetic and real-world uncertainty scenarios. Two baseline methods, i.e., ResNet-18 and DenseNet-121, along with two state-of-the-art methods considering uncertainty, namely SCN [41] and DMUE [13], are chosen for the competition.

1) Synthetic uncertain samples: Similar to [41] and [13], we synthesize 10%, 20%, and 30% samples in the training sets of RAF-DB and AffectNet respectively, with a random category other than their original labels. From Tab. V, the proposed MTAC outperforms baselines on two datasets, illustrating that uncertain samples hamper network training. Furthermore, with a growing proportion of uncertainty, the performance decline of MTAC compared to the corresponding baselines is comparatively lower, which serves as additional evidence for the effectiveness of our semantic similarity constraint at the feature level. The DMUE achieves the best results on AffectNet by multi-branch distribution learning when facing 10% and 20% uncertainty. Benefiting from multi-task correction, our MTAC obtains competitive performance in the above two settings and performs the best in the experiment with 30% uncertainty.

2) Real uncertain samples: Besides synthetic uncertainty, we introduce the AffectNet_Auto subset that contains naturally uncertain samples with incorrect labels and confusing emotions to enhance real uncertainty validation, which is barely conducted by existing work. The automatic labeling accuracy reported in the official document is 65% [9]. In Tab. VI, we demonstrate that MTAC outperforms other methods when dealing with actual ambiguous samples, and the increase in performance is more significant than that observed in the synthetic uncertainty experiment. This could be attributed to our approach accounting for the intrinsic confusion in real data. Specifically, MTAC utilizes the class-oriented weight in the auxiliary VA branch to alleviate the uncertainty of imbalanced classes, while the auxiliary AU branch employs a semantic memory template with updated category centers to facilitate promising annotation correction.

TABLE V
EVALUATION OF ENCOUNTERING SYNTHETIC UNCERTAIN SAMPLES. **BOLD** DENOTES THE BEST RESULT, AND *ITALICS* INDICATES THE SECOND BEST RESULT. $*$ MEANS PERFORMING 8-CATEGORY CLASSIFICATION.

Method	Uncertainty	RAF-DB	AffectNet*
ResNet	10%	80.64	57.25
DenseNet		81.03	57.50
SCN [41]		82.18	58.58
DMUE [13]		83.19	**61.21**
MTAC (ResNet)		*83.22*	59.45
MTAC (DenseNet)		**83.64**	*60.20*
ResNet	20%	78.06	56.23
DenseNet		79.48	56.98
SCN [41]		80.10	57.25
DMUE [13]		81.02	**59.06**
MTAC (ResNet)		*81.15*	58.50
MTAC (DenseNet)		**81.92**	*59.05*
ResNet	30%	75.12	52.60
DenseNet		76.51	52.80
SCN [41]		77.46	55.05
DMUE [13]		*79.41*	*56.88*
MTAC (ResNet)		79.01	56.45
MTAC (DenseNet)		**80.86**	**57.33**

TABLE VI
EVALUATION ON THE BENCHMARK WITH REAL UNCERTAIN SAMPLES. **BOLD** REPRESENTS THE BEST RESULT, AND *ITALICS* INDICATES THE SECOND BEST RESULT. † MEANS RE-IMPLEMENTING RESULTS.

Method	AffectNet_Auto
ResNet	53.23
DenseNet	53.83
SCN† [41]	55.43
DMUE† [13]	56.98
MTAC (ResNet)	*57.38*
MTAC (DenseNet)	**57.80**

E. Visualization

To present the specific manipulation effect of MTAC on uncertain samples, we visualize the intermediate results in terms of passive uncertainty suppression with uncertainty measurement and active uncertainty correction with relabeling.

1) Uncertainty measurement: Fig. 8 depicts the visualization of the uncertainty measurement in the target branch on samples in RAF-DB, AffectNet, and AffWild2. Generally, our MTAC effectively figures out the uncertain samples based on the confidence score and adaptively updates the value after the relabeling execution. In the second case of AffectNet, our MTAC accurately identifies the synthetic uncertainty and performs a correction for the original annotation.

2) Relabeling: To exhibit the semantic similarity-constrained relabeling workflow, we demonstrate the prediction distribution in the target branch and the semantic distance in the auxiliary AU branch using samples in RAF-DB, AffectNet, and AffWild2. Additionally, we compare our relabeling technique with subjective annotations obtained from twelve volunteers. As shown in Fig. 9, the memory template generated from the centers of semantic representation can widen the distance between classes. The distribution of predicted emotion categories closely resembles that of manual annotations. These results indicate that our MTAC can handle uncertain samples, facilitating model training and contributing to FER accuracy.

F. Performance Evaluation with the State-of-the-art

Since the proposed MTAC is designed for FER on large-scale datasets and utilizes multiple labels, we verify it with state-of-the-art approaches for single-task and multi-task performance evaluation.

1) Evaluation of single FER task: Tab. VII shows the performance comparison. To summarize, our method performs the best and the top-2 accuracy on RAF-DB and AffectNet, respectively. Despite LDL-ALSG [59] and Face2Exp [60] incorporates supplementary knowledge to facilitate model training, it solely takes into account the distribution at the label level and is unable to rectify uncertain samples, leading to performance degrade. In addition, IPA2LT [59], SCN [41], WSND [42], and FENN explicitly deal with uncertain labels and thus achieve good results. However, intrinsic uncertainty can still limit their feature learning without information in the side space. By leveraging uncertainty estimation, data-driven AU graph, and feature-level constrained relabeling, the proposed MTAC surpasses NMA [24] and delivers comparable performance to DMUE [13], which employ both uncertainty alleviation and auxiliary task concurrently. This demonstrates the contribution of the proposed modules in our study.

Similarly, we execute statistical analysis comparing baselines and MTAC. To assess the performance difference, we employ a series of paired two-tailed T-tests to examine the performance difference (H_0 : *There is no performance difference between baseline # and MTAC*). The p-value presented in Tab. VII indicates that our MTAC exhibits statistically significant improvements against SCN and FENN on three benchmarks and RAN across all four datasets, respectively.

2) Multi-task evaluation: To showcase the ability of MTAC to perform multi-task prediction, we present two advanced techniques, namely Emotion-GCN [25] and EmoFAN [27], which are re-implemented by ourselves for additional validation. As shown in Tab. VIII, our MTAC performs the best in the discrete emotion classification and obtains competitive CCC scores in the continuous emotion regression on both benchmarks, which are more robust than another two multi-task methods. One possible reason is that the uncertainty correction of discrete labels optimizes model parameter updates for more discriminate facial features and improves generalization performance on the VA estimation task.

TABLE VII
EVALUATION WITH THE STATE-OF-THE-ART. $*$ MEANS PERFORMING 7-CATEGORY CLASSIFICATION. † INDICATES THAT UNCERTAINTY HANDLING IS APPLIED. ‡ REPRESENTS ADDITIONAL KNOWLEDGE OF AUXILIARY TASKS INTRODUCED. **BOLD** REFERS TO THE BEST RESULT, AND *ITALICS* DENOTES THE SECOND BEST RESULT.

Method	Year	RAF-DB	AffectNet	FERPlus	SFEW
IPA2LT†[59]	2018	86.77	57.31	-	58.29
SCN† [41]	2020	88.14	60.23	88.01	-
RAN [54]	2020	86.90	59.50	88.55	56.40
LDL-ALSG‡ [17]	2020	85.53	59.35	-	56.50
WSND† [42]	2021	88.89	60.04	-	54.56
NMA†‡ [24]	2021	76.10	46.08	-	-
DMUE†‡ [13]	2021	89.42	**63.11**	89.51	*58.34*
IDFL [2]	2022	86.96	59.20	-	-
Face2Exp‡ [60]	2022	88.54	64.23*	-	-
FENN† [15]	2023	88.91	60.83	89.53	-
SPWFA-SE [61]	2023	86.31	59.23	-	-
MTAC (ResNet)†‡	Ours	89.31	61.58	88.74	56.65
MTAC (DenseNet)†‡	Ours	*89.83*	61.90	*89.58*	57.11
MTAC (SwinTrans)†‡	Ours	**90.52**	*62.28*	**90.42**	**58.94**
$p - value$	MTAC vs.	SCN	RAN	DMUE	FENN
	(< .05)	0.0025	0.0050	0.3838	0.0264

TABLE VIII
MULTI-TASK PERFORMANCE COMPARISON. * DENOTES USING CCC METRIC.† MEANS RE-IMPLEMENTING RESULTS. **BOLD** DENOTES THE BEST RESULT, AND *ITALICS* INDICATES THE SECOND BEST RESULT.

Dataset	Method	Year	Category	Valence*	Arousal*
AffectNet	ResNet	-	57.94	0.672	0.608
	DenseNet	-	59.11	0.701	0.624
	Emotion-GCN† [25]	2021	65.43	**0.762**	0.646
	EmoFAN† [27]	2021	62.37	0.732	**0.651**
	MTAC (ResNet)	Ours	65.09	0.753	0.635
	MTAC (DenseNet)	Ours	**65.80**	*0.758*	*0.649*
	MTAC (SwinTrans)	Ours	65.71	**0.762**	0.647
AffWild2	ResNet	-	56.41	0.365	0.327
	DenseNet	-	58.70	0.406	0.352
	Emotion-GCN† [25]	2021	62.68	*0.451*	**0.510**
	EmoFAN† [27]	2021	61.70	0.429	0.496
	MTAC (ResNet)	Ours	62.78	0.446	0.487
	MTAC (DenseNet)	Ours	*63.51*	0.449	0.503
	MTAC (SwinTrans)	Ours	**64.13**	**0.453**	*0.506*

G. Computation Complexity

Since the proposed MTAC introduces two auxiliary branches during the training process, we explore its computation increases compared with baseline methods. Here, two common metrics, i.e., multiply-accumulate computations (MACs) and accuracy, are used for evaluation. As shown in Fig. 10, compared with the corresponding backbone networks, w.r.t, ResNet-18, DenseNet-121, and SwinTransformer-Small, our MTAC achieves significant improvements with a slight computation burden, further illustrating the flexibility of the proposed framework.

V. CONCLUSION

In this paper, we proposed the MTAC framework to alleviate the uncertainty in facial expression images. The target FER branch measured uncertainty to calculate the confidence score and strengthen valid samples during model training. The auxiliary VA branch executed category balancing and joint feature learning with the support of continuous emotion labels. The auxiliary AU branch constructed the data-driven AU graph

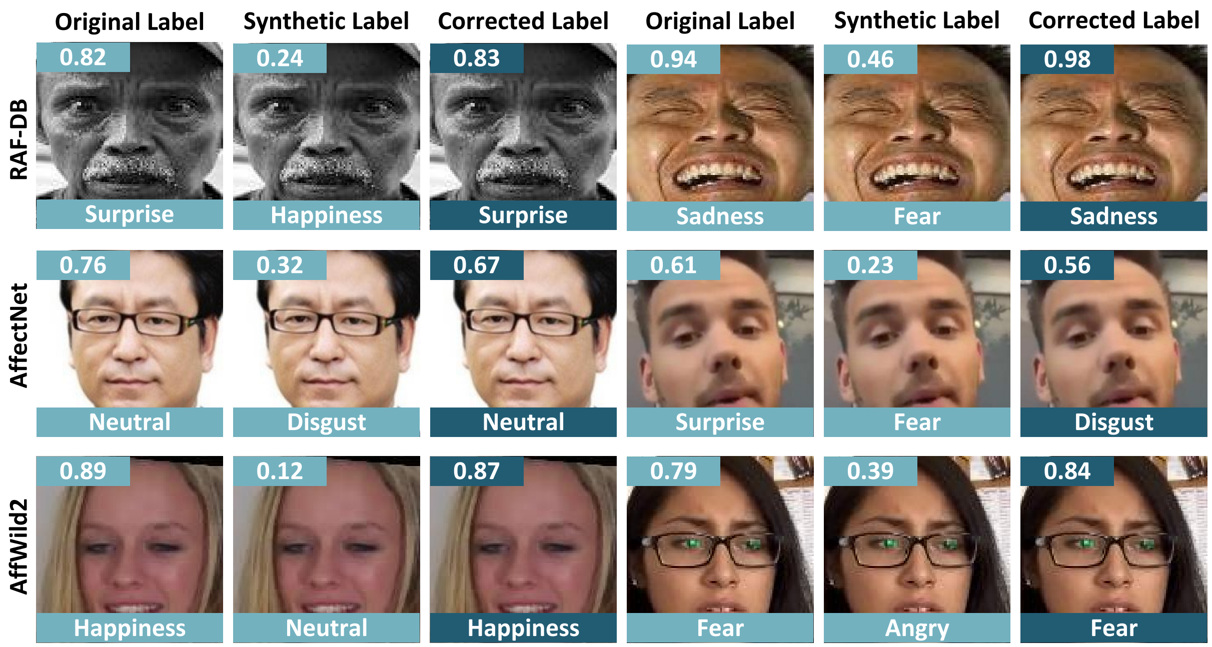


Fig. 8. Visualization of joint feature learning. Two examples of each dataset in RAF-DB, AffectNet, and AffWild2 are shown. The top left block denotes the confidence score α, and the bottom block presents the label of the current sample. From left to right of every three columns are the original, synthetic, and corrected samples, respectively.

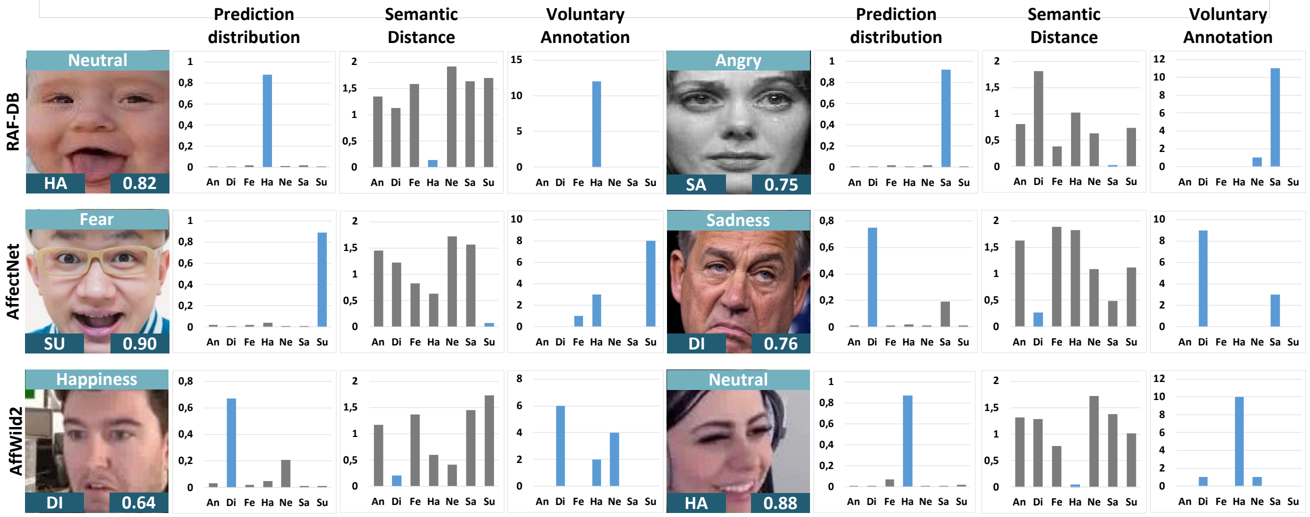


Fig. 9. Visualization of relabeling. Two examples of each dataset in RAF-DB, AffectNet, and AffWild2 are shown. The light color block at the top denotes the synthetic uncertain label, the dark color block at the bottom left indicates the new label after relabeling, and the dark block at the bottom right presents the confidence score α after correction. From left to right of every four columns are the original sample, prediction distribution, semantic similarity distance, and voluntary annotation statistic, respectively. DI, HA, SA, and SU are *disgust*, *happiness*, *sadness*, and *surprise*, respectively.

to generate semantic representations. The relabeling strategy corrected extremely uncertain samples under the feature-level similarity constraint based on the updated memory templates. Our MTAC's modular design allows for adding and removing branches based on what is needed during training and inference. Extensive experiments on five large-scale datasets showed that MTAC was robust to uncertain samples and achieved superior results in the FER task.

Although the MTAC performs competitive FER with full or weak supervision, the requirement of neighboring VA and AU annotations might limit the practical deployment. Alternative auxiliary tasks like face recognition and landmark detection could be introduced. Conversely, MTAC can be expanded to produce labels for incremental learning, pre-train universal encoders of facial expressions, and address the uncertain problem in other data modalities.

Acknowledgment

This work was supported by the Academy of Finland for Academy Professor project EmotionAI (grants 336116, 345122), ICT 2023 project TrustFace (grant 345948), and the University of Oulu & The Academy of Finland Profi 7 (grant

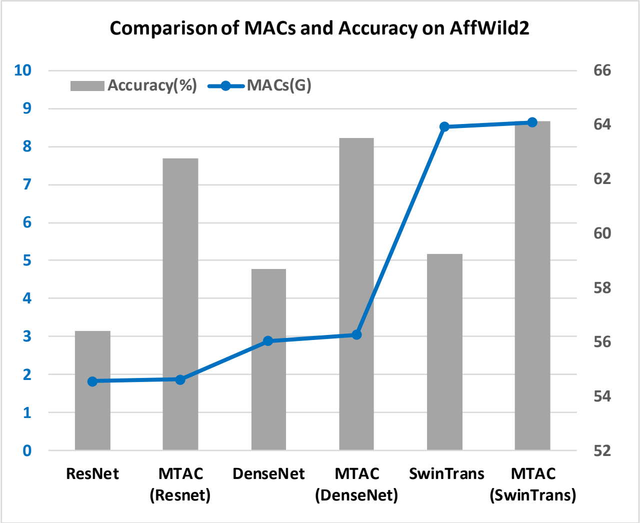


Fig. 10. Computation comparison of MACs and accuracy between baseline methods and MTAC. The grey bar and the blue spot indicates the accuracy and the MACs of each model, respectively.

352788), and the Ministry of Education and Culture of Finland AI Forum project. The work of Dr. Yang Liu was supported by the Finnish Cultural Foundation for North Ostrobothnia Regional Fund 'Towards Crowdsensing Facial Affect Encoder for Trustworthy Mental Wellbeing' (grant 60231712). Dr. Yang Liu also appreciated the support of the China Scholarship Council (grant 202006150091) in the preliminary version of this work. The authors wish to acknowledge CSC-IT Center for Science, Finland, for generous computational resources.

REFERENCES

[1] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep learning for micro-expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2028–2046, 2022.

[2] Y. Li, Y. Lu, B. Chen, Z. Zhang, J. Li, G. Lu, and D. Zhang, "Learning informative and discriminative features for facial expression recognition in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3178–3189, 2022.

[3] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, and Y.-J. Liu, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5826–5846, 2021.

[4] X. Ben, C. Gong, T. Huang, C. Li, R. Yan, and Y. Li, "Tackling micro-expression data shortage via dataset alignment and active learning," *IEEE Transactions on Multimedia*, 2022.

[5] L. Lo, H. Xie, H.-H. Shuai, and W.-H. Cheng, "Facial chirality: From visual self-reflection to robust facial feature learning," *IEEE Transactions on Multimedia*, vol. 24, pp. 4275–4284, 2022.

[6] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proceedings of the 2010 IEEE Computer Society Sonference on Computer Vision and Pattern Recognition-workshops (CVPR)*. IEEE, 2010, pp. 94–101.

[7] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[8] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[9] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[10] D. Kollias, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2328–2336.

[11] Y. Liu, F. Yang, C. Zhong, Y. Tao, B. Dai, and M. Yin, "Visual tracking via salient feature extraction and sparse collaborative model," *AEU-International Journal of Electronics and Communications*, vol. 87, pp. 134–143, 2018.

[12] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 907–929, 2019.

[13] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6248–6257.

[14] J. Li, C. Xiong, and S. C. Hoi, "Learning from noisy data with robust representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9485–9494.

[15] Y. Gu, H. Yan, X. Zhang, Y. Wang, J. Huang, Y. Ji, and F. Ren, "Towards facial expression recognition in the wild via noise-tolerant network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 5, pp. 2033–2047, 2023.

[16] Z. Cui, Y. Zhang, and Q. Ji, "Label error correction and generation through label relationships," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 04, 2020, pp. 3693–3700.

[17] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 13 984–13 993.

[18] Y. Liu, X. Zhang, J. Kauttonen, and G. Zhao, "Uncertain label correction via auxiliary action unit graphs for facial expression recognition," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 777–783.

[19] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 211–220, 2018.

[20] W. Huang, S. Zhang, P. Zhang, Y. Zha, Y. Fang, and Y. Zhang, "Identity-aware facial expression recognition via deep metric learning based on synthesized images," *IEEE Transactions on Multimedia*, vol. 24, pp. 3327–3339, 2021.

[21] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.

[22] J. A. Russell, "Evidence of convergent validity on the dimensions of affect." *Journal of personality and social psychology*, vol. 36, no. 10, p. 1152, 1978.

[23] B. Yoo, Y. Kwak, Y. Kim, C. Choi, and J. Kim, "Deep facial age estimation using conditional multitask learning with weak label expansion," *IEEE Signal Processing Letters*, vol. 25, no. 6, pp. 808–812, 2018.

[24] S. Zhang, Z. Huang, D. P. Paudel, and L. Van Gool, "Facial emotion recognition with noisy multi-task annotations," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 21–31.

[25] P. Antoniadis, P. P. Filntisis, and P. Maragos, "Exploiting emotional dependencies with graph convolutional networks for facial expression recognition," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 2021, pp. 1–8.

[26] Y. Chen and J. Joo, "Understanding and mitigating annotation bias in facial expression recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 980–14 991.

[27] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 42–50, 2021.

[28] J. Zhang, Z. Su, and L. Liu, "Median pixel difference convolutional network for robust face recognition," *arXiv preprint arXiv:2205.15867*, 2022.

[29] Y. Li, Z. Zhang, B. Chen, G. Lu, and D. Zhang, "Deep margin-sensitive representation learning for cross-domain facial expression recognition," *IEEE Transactions on Multimedia*, 2022.

[30] Y. Liu, X. Zhang, Y. Li, J. Zhou, X. Li, and G. Zhao, "Graph-based facial affect analysis: A review," *IEEE Transactions on Affective Computing*, 2022. [Online]. Available: https://doi.org/10.1109/TAFFC.2022.3215918

[31] Y. Liu, X. Zhang, J. Zhou, and L. Fu, "Sg-dsn: A semantic graph-based dual-stream network for facial expression recognition," *Neurocomputing*, vol. 462, pp. 320–330, 2021.

[32] R. Zhao, T. Liu, Z. Huang, D. P.-K. Lun, and K. K. Lam, "Geometry-aware facial expression recognition via attentive graph convolutional networks," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1159–1174, 2023.

[33] X. Jin, Z. Lai, and Z. Jin, "Learning dynamic relationships for facial expression recognition based on graph convolutional network," *IEEE Transactions on Image Processing*, vol. 30, pp. 7143–7155, 2021.

[34] Y. Xie, T. Chen, T. Pu, H. Wu, and L. Lin, "Adversarial graph representation adaptation for cross-domain facial expression recognition," in *Proceedings of the 28th ACM international conference on Multimedia*, 2020, pp. 1255–1264.

[35] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes, "Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 2022, pp. 1239–1246.

[36] T. Song, Z. Cui, W. Zheng, and Q. Ji, "Hybrid message passing with performance-driven structures for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6267–6276.

[37] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[38] Y. Chen, M. Liu, X. Wang, F. Wang, A.-A. Liu, and Y. Wang, "Refining noisy labels with label reliability perception for person re-identification," *IEEE Transactions on Multimedia*, 2023.

[39] Z. Sun, H. Liu, Q. Wang, T. Zhou, Q. Wu, and Z. Tang, "Co-ldl: A co-training-based label distribution learning method for tackling label noise," *IEEE Transactions on Multimedia*, vol. 24, pp. 1093–1104, 2021.

[40] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1237–1246.

[41] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 6897–6906.

[42] F. Zhang, M. Xu, and C. Xu, "Weakly-supervised facial expression recognition in the wild with noisy data," *IEEE Transactions on Multimedia*, vol. 24, pp. 1800–1814, 2021.

[43] W. Hu, Y. Huang, F. Zhang, and R. Li, "Noise-tolerant paradigm for training face recognition cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 887–11 896.

[44] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 212–220.

[45] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[46] Y. Liu, X. Zhang, Y. Lin, and H. Wang, "Facial expression recognition via deep action units graph network based on psychological mechanism," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 311–322, 2019.

[47] L. Lei, T. Chen, S. Li, and J. Li, "Micro-expression recognition based on facial graph representation learning and facial action unit fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1571–1580.

[48] Y. Cheng, Y. Sun, H. Fan, T. Zhuo, J.-H. Lim, and M. Kankanhalli, "Entropy guided attention network for weakly-supervised action localization," *Pattern Recognition*, vol. 129, p. 108718, 2022.

[49] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[50] N. Le, K. Nguyen, Q. Tran, E. Tjiputra, B. Le, and A. Nguyen, "Uncertainty-aware label distribution learning for facial expression recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6088–6097.

[51] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM international conference on multimodal interaction*, 2016, pp. 279–283.

[52] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in *Proceedings of the 16th international conference on multimodal interaction*, 2014, pp. 461–466.

[53] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. 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*. Springer, 2013, pp. 117–124.

[54] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[55] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2018, pp. 59–66.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

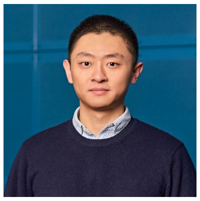
[57] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[58] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[59] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 222–237.

[60] D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang, "Face2exp: Combating data biases for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 291–20 300.

[61] Y. Li, G. Lu, J. Li, Z. Zhang, and D. Zhang, "Facial expression recognition in the wild using multi-level features and attention mechanisms," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 451–462, 2023.



Yang Liu received his D.Sc. degree in computer science and technology from the South China University of Technology, in 2021. He is currently a Post-doctoral researcher at the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. His current research interests include facial expression recognition, affective computing, and trustworthy deep learning.

Xingming Zhang is currently a Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He is a member of the Standing Committee of the Education Specialized Committee of China Computer Federation and the Standing director of the University Computer Education Research Association of China. His research focuses on video processing, big data, video surveillance, and face recognition.

Janne Kauttonen received his Ph.D degrees in physics from the University of Jyväskylä, Finland, in 2012. His post-doctoral research in Aalto University, Finland, and Carnegie Mellon University, USA, included neuroimaging experiments, development of computational methods, data and statistical analyses. Since 2019 he has worked as a researcher at Haaga-Helia University of Applied Sciences, Finland, in fields of data science, applied machine learning, cognitive sciences and human-computer interaction.

Guoying Zhao (IEEE Fellow 2022) received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. She is currently an Academy Professor and full Professor (tenured in 2017) with University of Oulu. She is also a visiting professor with Aalto University. She is a member of Academia Europaea, a member of Finnish Academy of Sciences and Letters, IEEE Fellow, IAPR Fellow and AAIA Fellow. She has authored or co-authored more than 300 papers in journals and conferences with 24300+ citations in Google Scholar and h-index 73. She is panel chair for FG 2023, publicity chair of 22nd Scandinavian Conference on Image Analysis (SCIA 2023), was co-program chair for ACM International Conference on Multimodal Interaction (ICMI 2021), co-publicity chair for FG2018, and has served as area chairs for several conferences and was/is associate editor for IEEE Trans. on Multimedia, Pattern Recognition, IEEE Trans. on Circuits and Systems for Video Technology, Image and Vision Computing and Frontiers in Psychology Journals. Her current research interests include image and video descriptors, facial-expression and micro-expression recognition, emotional gesture analysis, affective computing, and biometrics. Her research has been reported by Finnish TV programs, newspapers and MIT Technology Review.