Fine-Grained Facial Expression Recognition in the Wild

Liqian Liang, *Student Member, IEEE* Congyan Lang, *Member, IEEE* Yidong Li, *Member, IEEE* Songhe Feng, *Member, IEEE* Jian Zhao, *Member, IEEE*

***Abstract*—Over the past decades, researches on facial expression recognition have been restricted within six basic expressions (anger, fear, disgust, happiness, sadness and surprise). However, these six words can not fully describe the richness and diversity of human beings' emotions. To enhance the recognitive capabilities for computers, in this paper, we focus on fine-grained facial expression recognition in the wild and build a brand new benchmark *FG-Emotions* to push the research frontiers on this topic, which extends the original six classes to more elaborate thirty-three classes. Our *FG-Emotions* contains 10,371 images and 1,491 video clips annotated with corresponding fine-grained facial expression categories and landmarks. *FG-Emotions* also provides several features (*e.g.*, LBP features and dense trajectories features) to facilitate related research. Moreover, on top of *FG-Emotions*, we propose a new end-to-end Multi-Scale Action Unit (AU)-based Network (MSAU-Net) for facial expression recognition with image which learns a more powerful facial representation by directly focusing on locating facial action units and utilizing "zoom in" operation to aggregate distinctive local features. As for recognition with video, we further extend the MSAU-Net to a two-stream model (TMSAU-Net) by adding a module with attention mechanism and a temporal stream branch to jointly learn spatial and temporal features. (T)MSAU-Net consistently outperforms existing state-of-the-art solutions on our FG-Emotions and several other datasets, and serves as a strong baseline to drive the future research towards fine-grained facial expression recognition in the wild.**

***Index Terms*—Fine-Grained Facial Expression Recognition, Benchmark Dataset, Action Unit Detection, Attention, Two-Stream Network**

I. Introduction

FACIAL expression plays a vital role in revealing a person's internal thoughts and feelings. Over the past few decades, **f**acial **e**xpression **r**ecognition (FER) has been a hot issue in the field of computer vision and human-computer interaction. In general, FER methods follow two lines of research: categorical methods and continuous methods (also called dimensional methods) [39][37], which map the given resources (images or videos) to discrete classes and describe affective state of resources in dimensional space formulated as valence, arousal and power, respectively. Though continuous methods may represent a wider range of facial expressions, they are less competitive than categorical methods due to the limited ability to show direct and intuitive definitions of facial expressions. Therefore, categorical methods remains as the pioneering perspective in FER. However, most previous categorical models [14][15][16][36][37][38][48][49][50][51][52] are restricted to analyzing six basic emotion classes (*i.e.*, surprise, disgust, fear, happiness, sadness and anger) or seven classes plus an extra neutral emotion according to Ekman's theory [4] widely used in computer vision community, which is disadvantageous for fine-grained FER. Few researches focus on exploring the richness and diversity of human emotional situations.

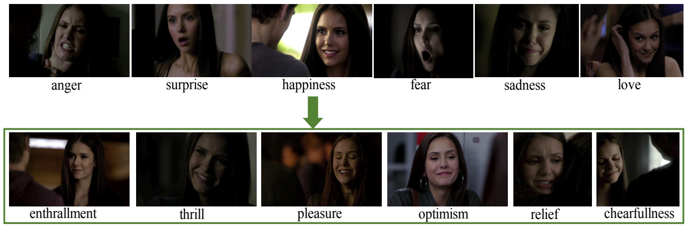


Fig. 1. Illustration of motivation with several samples from our dataset *FG-Emotions*. Previous FER efforts merely focus on the conventional six basic classes, which cannot fully describe the richness and diversity of human beings' emotions, whereas we aim to solve fine-grained FER in the wild, tailored better for real applications.

To alleviate this issue, Du *et al.* [1] first presents a concept named "compound facial expression of emotion", in which "compound" means that the expression here is not an actual expression but a combination of six basic emotion classes, *e.g.*, "happily surprised" and "sadly surprised". However, compound expression recognition can only be seen as an inflexible extension of the six basic classes, which is insufficient for elaborating the subtleties amongst facial expressions effused from different moods. To promote facial expression research towards broader and deeper, we cast insights on psychological fields for assistance. Apart from the most popular Ekman's basic emotion theory, there exist other theories that cover more diverse emotions like the well-known two theories: "Wheel of Emotion" theory proposed by R. Plutchik [61] and hierarchy model of emotions presented by W.Gerrod Parrot [2]. "Wheel of Emotion" represent the emotions in the form of circle, where contrast emotions lie in the opposite sectors of the circle while the adjacent emotions may combine a new emotion. The most related work to ours, F^2ED [1] [53] consults Lee's theory which is derived by "Wheel of Emotion", where 54 emotions are clustered to four none-semantic groups

Liqian Liang, Congyan Lang, Yidong Li and Songhe Feng are with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China. E-mail: lqliang@bjtu.edu.cn; cylang@bjtu.edu.cn; ydli@bjtu.edu.cn; shfeng@bjtu.edu.cn.

Jian Zhao is with Institute of North Electronic Equipment, Beijing 100191, China. E-mail: zhaojian90@u.nus.edu. Homepage: https://zhaoj9014.github.io/.

[1] F^2ED dataset is not available for the present.

based on eye features and each group occupies a quadrant of the circle. Nevertheless, here we consider W.Gerrod Parrot's theory since hierarchy architecture suits more for the fine-grained classification task to model the relationships between basic classes and its subclasses and can be generalized well to Ekman's expression system. Besides, the F^2ED is created in the controlled environment, while we intend to explore the fine-grained FER in the wild.

Fine-grained FER in the wild is essentially a fine-grained classification problem with the nature of tiny differences among the fine-grained classes belonging to the same upper class. It enables better understanding of human beings' emotions under unconstrained settings, which means the samples are collected in real-world scenes with various viewpoints, illumination, poses, scales, occlusion *etc.* Thus it is more challenging than fine-grained FER in the controlled environment. However, the existing benchmark datasets [5][6][7][8][9][47][53] are not suitable for such a new task. Most of them [5][6][7][8][9][47][58] are built for conventional FER based on six, seven or eight basic classes. Few may move forward to finer classes with compound expressions like EmotioNet with 23 classes (including basic and compound emotions) [38] and RAF-DB [51] with 7 basic classes and 12 compound classes. To understand human beings' emotions more comprehensively, we build a brand new benchmark *FG-Emotions*, as shown in Fig.1, where thirty-three facial expression classes are carefully defined to cover the abundance of facial expressions by consulting the psychological studies of Darwin [3] and W.Gerrod Parrot [2]. It in total includes 10,371 images and 1,491 video clips annotated with fine-grained class labels. Specifically, these fine-grained classes are constructed as a four-level tree structure as shown in Fig. 2 (b). Note that there are several differences on the definitions of six basic expressions between Parrot's expression system and Ekman's [4]. To remain consistent with the definitions of Ekman and ensure fair and reasonable experimental comparisons between our framework and state-of-the-arts, we made several modifications that will be described in Section III.

We further propose a new end-to-end model, *i.e.*, **M**ulti-**S**cale **A**ction **U**nit based **Net**work (MSAU-Net) for fine-grained FER with image. As for recognition with video, we extend the MSAU-Net to a **t**wo-stream model (TMSAU-Net). In contrast to general object classification problems, FER usually demands utilizing task-specific AUs defined in Facial Action Coding System [4] to complement feature embedding, though there exist some severe drawbacks regarding to AU analysis. AU annotation is labor-consuming especially for large-scale datasets. Besides, AU only considers the spatial information and statistic features regardless of temporal information and motion features when dealing with videos. Unlike most prior methods which take AU as reference and lack the ability to recognize subtle difference among fine-grained classes which belong to the same basic class, inspired by AU detection solutions [17][18][19][20][21], our MSAU-Net introduces AU detection module to learn more reliable facial representations which focus on acquiring clues of AUs spontaneously. Moreover, we locate the most discriminative facial regions stimulated by AU detection module and "zoom in " these regions to investigate more subtleties through a multi-scale network structure. TMSAU-Net incorporates spatial and temporal information jointly for fine-grained FER with video, where attention mechanism is used for aggregation of spatial features and optical field is cast into temporal stream network to capture temporal information. Actually, TMSAU-Net can be regarded as a unified framework for FER with both image and video, regarding that the attention mechanism and temporal stream network can be removed when recognizing images. Extensive experiments on both existing benchmarks and *FG-Emotions* dataset show that our method achieved superior performance on both our dataset and other benchmarks compared with state-of-the-arts.

Our contributions are summarized as follows.

1. We explore the problem of fine-grained FER in the wild and propose a new dataset *FG-Emotions* which extends the standard of previous efforts to fine-grained classes to advance detailed facial analytics. *FG-Emotions* consists of 10,371 images and 1,491 video clips.

2. We propose a simple yet effective end-to-end deep model *i.e.*, (T)MSAU-Net, tailored for fine-grained FER with image and video, which serves as a strong baseline to inspire more future research efforts on this task.

3. Comprehensive evaluations on our *FG-Emotions* as well as other benchmark datasets verify the superiority of (T)MSAU-Net over the state-of-the-arts.

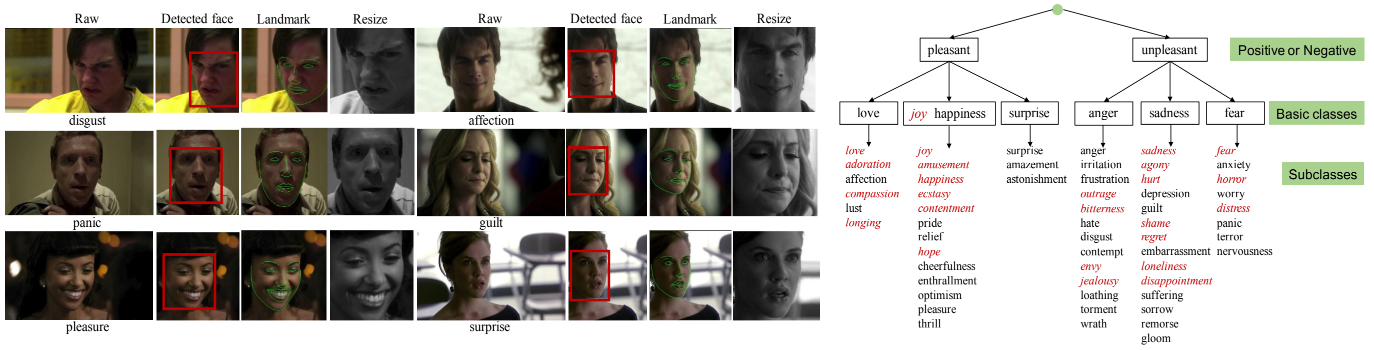
II. Related Work

In this section, we revisit the related works on FER in chronological order from traditional methods to deep learning based methods. The latest developments in FER, which is facial action unit detection, are summarized in a single subsection.

A. Facial Expression Recognition

The journey of FER starts from the definition of Ekman [4], who defines six basic emotions and facial action units. Conventional studies normally made a progress by improving three fundamental modules: feature representation, feature selection and classifiers. For feature representation, static features, *e.g.*, LBP [10] and motion feature, *e.g.*, dense trajectories [11] have been applied for this task. Most researchers put efforts on devising an effective classifier by developing new models or the fusion of different classifiers, such as [12][13][35].

Recent works have shifted their focus on deep networks, *e.g.*, [14][15][16][36][37][38][48][49][50][51][52][56][57] to learn a more distinctive facial representation automatically. Ping Liu *etc.* [14] presented a **B**oosted **D**eep **B**elief **N**etwork (BDBN) for performing three training stages iteratively in a unified loopy network for a joint fine-tune process. Samira Ebahimi Kahou *etc.* [15] explored different methods of combining predictions of modality-specific models including a deep belief net and CNN trained to recognize facial expressions in single frames. Martin Wollmer *etc.* [16] proposed a fully automatic audiovisual recognition approach based on Long-Short-Term-Memory modeling of word-level audio and video features. In [48], a multilayer RBM was used to learn



(a) Examples from FG-Emotions

(b) Hierarchical structure adopted in FG-Emotions

Fig. 2. Examples and hierarchy structure of fine-grained FER definition system from our *FG-Emotions* dataset. Left: Illustration of samples for fine-grained FER. Right: Hierarchy structure which consists of four layers where our fine-grained thirty three classes are defined as leaf nodes.

higher-level features for FER on top of the over-complete representations learned by CNN. GAN-based models have also been studied in resolving the FER task. In [56], a **M**ulti-channel **D**eep **S**patial-**T**emporal feature **F**usion neural **N**etwork (MDSTFN) is presented to perform the deep spatial-temporal feature extraction and fusion from static images, where optical flow and gray-level images are used as temporal and spatial information, respectively. Yang *et al.* [52] proposed a **De**-expression **R**esidue **L**earning (DeRL) procedure based on GAN to filter out the expressive information during the de-expression process yet still embedded in the generator for pose-invariant FER. [53] proposed FaPE-GAN where LightCNN [54] was used as backbone to synthesize face images for solving training data imbalance. Liu *et al.*[57] proposed an **i**dentity-**d**isentangled **f**acial **e**xpression **r**ecognition **m**achine (IDFERM), where the identity is untangled from a query sample by exploiting its difference from its references such as its mined or generated frontal and neutral normalized faces. Other progresses [36][37][38][49][50][51][58] turned to solve a more challenging problem, FER in the wild where background noises severely affect face detection and FER.

These approaches are confined to limited number of class labels and training images while highly dependent on pre-training on large-scale datasets. Besides, the plain CNN network may ignore the subtleties among subclasses within one general class. For example, “astonishment” and “surprise” both belong to the “surprise” class, but astonishment is a feeling of great surprise with more powerful intensity of emotion. Hence we require a model sensitive to minute details among subclasses of the same general class. To satisfy this demand, we derive MSAU-Net and TMSAU-Net under a unified framework setting, both of which perform well not only on our newly constructed dataset but also on existing six classes datasets.

The most related work with regard to the fine-grained FER task is [53], where a database F^2ED is created in the controlled environment with 4 poses and 54 expressions. However, our work aims to explore fine-grained FER in the wild which will promote this research topic from another perspective.

B. *Facial Action Unit Detection*

Recently, many research efforts have been devoted to automated AU detection. The task of this study is to detect active action units and their corresponding intensities. We mainly review CNN-related works below.

In general, AU detection methods are categorized into two types: patch-based methods and structured deep methods. Patch-based methods target to define a subset of facial regions for better action unit detection, *e.g.*, **A**ctive **P**atch **L**earning (APL) [17] , JPML[18] , AUDN [19] , **B**oosted **DBN** (BDBN) [20] and **D**eep **R**egion and **M**ulti-label **L**earning (DRML)[21]. AUDN [19] combined three independent modules sequentially that learn expression-specific representation, and searched subset of the representation that simulates a best AU. BDBN [20] performed three training stages including feature learning, patch selection and classifier construction iteratively in an end-to-end network. DRML [21] was inspired by JPML but this framework naturally fused two tasks into one framework, allowing multi-label learning and region learning to interact more directly. Normally Structured models refer to the methods that learn task-specific constraints and relations between output variables directly from the data. Stefanos *et al.* [22] adopted a latent variable CRF to jointly detect multiple AUs. Robert [23] placed a CRF graph on the fully connected output layer of their network where the CRF was formulated with both unary and binary cliques.

AU detection aims to facilitate the accuracy of FER. In this paper, we leverage the state-of-the-art AU detection methods to advance the distinctive capacity of face representations rather than taking AU as supervisory signals. The proposed framework is presented in detail in Section 4.

Note that fine-grained image classification works are not mentioned here because of the page limit. Moreover, our fine-grained FER problem is not like the other fine-grained image classification problem regarding that expression classes are pre-defined according to psychological works but not on the basis of consensus.

III. Properties of *FG-Emotions*

In this section, we introduce the *FG-Emotions*, a brand new fine-grained FER in the wild dataset. Several appealing

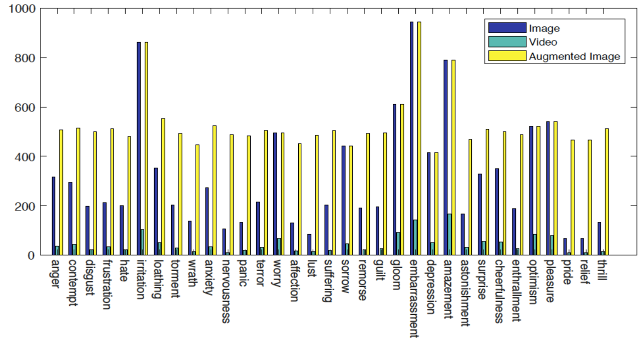


Fig. 3. Data distribution over images, video and augmented image datasets in *FG-Emotions*.

properties are elaborated as follows. Meanwhile, the *FG-Emotions* dataset is expected to serve as a benchmark with our provided models and suitable for deeper understanding of human beings' emotions.

Data collection and definition. In discrete emotion theory, there are many different basic emotion definition systems, of which the most popular one adopted in computer vision community is conducted by Paul Ekman [4], where the six basic emotions are *anger, disgust, fear, happiness, sadness and surprise*. However, Ekman's definition system is not appropriate for the task when involving fine-grained FER. Inspired by hierarchical models of emotions in psychology field, we follow the taxonomy defined by W.Gerrod Parrot *etc.* [2]. Fig. 2 (b) shows the tree structure employed in our paper, where there exist one root node, secondary nodes (pleasant and unpleasant), third-level nodes (anger, fear, love, joy, sadness and surprise) and leaf nodes (*i.e.*, subclasses of basic classes). Among those subclasses, we carefully select thirty three words aiming to adapt to fine-grained classification tasks in computer vision. Therefore, we put forward four principles of modification to satisfy the requirements of consistency with other FER methods and eliminating ambiguous subclasses instead of putting all emphasis on psychological-based study. In fact, under these four principles, the 33 fine-grained subclasses are finally defined from the experience of annotators during data collection process, *e.g.*, videos of certain subclasses are too hard to obtain from Internet, then those subclasses are not adopted in our FG-Emotions. The detailed four principles of modification are as follows:

1. To remain consistent with traditional six basic FER task in computer vision, we replace the basic class "joy" with "happiness";

2. To balance the numbers of pleasant and unpleasant emotions, we take "love" as basic class following W.Gerrod Parrot while classify "disgust" into "anger". In this case, the original basic class set can be seen as the subset of our four-layer class set;

3. To avoid the coexistence of the words with close similarities of meanings, *e.g.*, we keep the "terror" while eliminating "horror";

4. To ensure the number of collected video data, we get rid of those words hardly used in daily life or difficult to be searched such as "agony" and "ecstasy".

TABLE I

DATASET STATISTICS IN *FG-Emotions*. IMAGE, VIDEO, AUGMENTED INDICATE THE IMAGE DATASET, VIDEO DATASET AND AUGMENTED IMAGE DATASET RESPECTIVELY.

Source	Training set	Validation set	Test set	Total
Image	7,826	1,500	1,045	10,371
Video	825	275	391	1,491
Augmented Image	12,826	2,580	2,055	17,461

The whole process of constructing FG-Emotions is comprised of three phases: data collection phase, data definition phase and data inspection phase. During data collection phase, five annotators first vote to define an initial set of 42 subclasses following the four principles. On top of the initial set, we take six basic classes as key words for performing automate Internet search to download corresponding videos. The videos are mainly collected from TV series or movies where sequences contain a complete process starting from a neutral face, reaching the peak emotion in the middle and ending with a neutral face again. Based on these raw videos, we step into data definition phase, which includes two tasks: one is judging whether the videos are rational and classifying them into one of the sub-classes of certain basic emotion; the other one is to pick out the key frames from those defined videos. For each video, given its basic class, an annotator carefully classifies the video clip to corresponding subclass and gives a confidence score of assigning this subclass. If there is no video clip related to a certain subclass, then repeat the data phase until the new videos satisfy the annotator's requirements or annotator find the search so hard that this subclass is excluded. Then another annotator manually rechecks and selects the key frame for each video, which indicates the climax stage in a complete course of facial expression change. During this phase, we also eliminate those subclasses with low confidence (less than 0.5 in our paper) after the annotators finish labeling all the videos. The final thirty three subclasses are confirmed via another round of voting by different five annotators to recheck the four principles and data collection process. The detailed list of the fine-grained expression classes is as below where the words in brackets denote the generic upper classes to which the former words belong: *anger, contempt, disgust, frustration, hate, irritation, loathing, torment, wrath (anger), anxiety, nervousness, panic, terror, worry (fear), cheerfulness, enthrallment, optimism, pleasure, pride, relief, thrill (happiness), affection, lust (love), suffering, sorrow, remorse, guilt, gloom, embarrassment, depression (sadness), amazement, astonishment, surprise (surprise).*

With the definition of 33 fine-grained expression classes, the inspection phase is performed on all the annotated videos and corresponding key frames to verify the correctness. In cases where annotations are erroneous, the information is manually rectified by 5 well-informed analysts. The verifying process follows the rules of voting: in a group of five people, the labels or the key frames are confirmed if greater than or equal to three people come to agree. If not, these five people need to re-annotate the labels of videos or key frames. After the second voting, we keep the videos or key frames without any debates while removing those that remain uncertain. Subsequently, the frames adjacent to the key frames are densely sampled to enlarge the scale of image dataset. Our dataset eventually eliminates the data with ambiguity deviated from the principle of computer vision or human's subjectivity. The whole work took around two months to accomplish by twenty professional data annotators.

The statistics of *FG-Emotions*. The *FG-Emotions* dataset covers both image and video media resources. The resolution of the videos is 720p and the length for each video ranges from 1 second to 4 seconds. After we pick out the key frames from raw videos, we construct the image dataset by automatically sampling the frames near key frames then manually selecting those sampled frames, therefore the labels of images are the same as corresponding video clips. Here the manual process is performed by two experts, one for selection and the other one for inspection.

Thorough data distribution statistics over thirty three classes is illustrated in Fig. 3, from which we can conclude our dataset suffers from class imbalance. Our solution is to utilize data augmentation to improve this situation. Normally data augmentation techniques are divided into two groups: online and offline methods, which focus on enlarging the dataset during training and pre-processing respectively. In our paper, we adopt the combination of online and offline methods to expand data on both size and diversity. The statistics of augmented image dataset, *i.e.*, data split, total and distribution, is described in Tab. I and Fig. 3, and now data distribution after augmented operations is basically balanced over thirty three classes. The details of augmented operations are presented in the following Image pre-processing. In particular, for video dataset, we randomly sample 391 video clips within the constraint of covering all the 33 fine-grained classes to form the test set, while the rest form the training set of 825 video clips with the same constraint and validation set of 275 video clips with no constraint. The ratio among the scale of training, validation and test set is around 3:1:1.5. As for image and augmented image dataset, the training, validation and test set is composed of frames sampled then selected from corresponding video clips.

Image pre-processing. Original images in our dataset are frames directly drawn from the raw video clips. Considering the varying face positions and background noises, a series of pre-processed procedure is conducted including face detection, face alignment and image resize. We adopt state-of-the-art human face detection algorithm libfacedetection and face alignment technique developed by Dlib to preprocess raw images. Finally, we obtain the normalized 224 × 224 images without large deviation angles or much background noises. Thus the features extracted from the images can better represent the facial muscle movement. Fig. 2 (a) exhibits some samples of the fine-grained classes and the pre-processed normalized images in our dataset. Moreover, we undertake data augmentation during image pre-processing by utilizing operations include random perturbations and transforms such as rotation, scaling and noise. Here Gaussian noise [42][43] is employed to enlarge data size.

***FG-Emotions* baselines.** We raise two kinds of baselines for later study: conventional methods and deep learning methods. We employ SVM classifier and LBP features as conventional baseline for FER with image, while dense trajectories [11] are used as features and the classifier remains the same SVM for FER with video. For deep learning methods, InsightFace [24] is fine-tuned on our *FG-Emotions* image dataset , since the network is proposed for face recognition and more easily learned to fit our task. And Neural Aggregation Network (NAN) [25] is fine-tuned on *FG-Emotions* video dataset.

Differences between *FG-Emotions* and F^2ED. The differences between our *FG-Emotions* and F^2ED lie in the following aspects: (1) Data collecting environment. The samples from F^2ED are collected under the lab controlled environment while our *FG-Emotions* concentrates on the resources in the wild. (2) Fine-grained class structure. Our 33 classes are constructed as tree structure following the definition of W.Gerrod Parrot [2], while 54 classes of F^2ED are organized into four groups based on the theory of Lee [55]. Notice that these four groups do not have semantic meanings since their 54 classes are clustered by k-means algorithm utilizing seven eye features. (3) Data type. *FG-Emotions* provides two kinds of data: image and video while F^2ED only provides data in image form. (4) Dataset scale. F^2ED provides 219,719 images of 119 subjects with four poses (half left, half right, front and bird view). *FG-Emotions* includes less samples yet instructive for exploring FER in the wild task.

IV. Framework

In this section, we propose the **m**ulti-**s**cale **AU**-based **Net**work (MSAU-Net) for FER with image and further extend it to a **t**wo-stream model TMSAU-Net for FER with video. MSAU-Net takes images as input, then extracts aggregation of local features via a multi-scale AU detection structure and finally outputs corresponding class label predictions. TMSAU produces label predictions through a two-stream network which incorporates spatial and temporal information of frame sequences, guided by attention module and optical flow respectively. TMSAU-Net can be seen as a unified framework for FER with both image and video since the modules can be flexibly plugged in or out.

A. MSAU-Net for Fine-Grained FER with Image

When it refers to face analysis such as face recognition, a common solution is to investigate discriminative facial representation including global feature and local feature. In this paper, we adopt the fusion feature of both global and local feature as facial representation. For global feature extraction, inspired by facial action unit detection methods, we aim to automatically learn the feature more sensitive to AUs but with no AUs annotations as supervision like other FER methods. Here we basically follow the architecture of DRML [21] to perform AU detection. Meanwhile, the MAC (short for **m**aximum **a**ctivation of **c**onvolutions) descriptor of the final convolutional layer (*i.e.*, Conv 7 of DRML as shown in Fig. 4) is used for local feature extraction, since maximum activations of feature maps normally reflect the most few distinguishable regions in the overall images. To better exploit

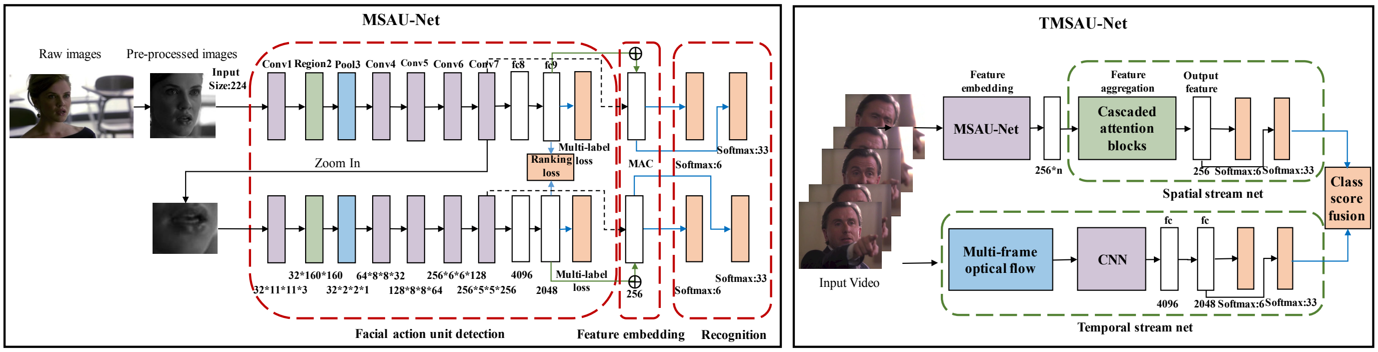


Fig. 4. Illustration of MSAU-Net (left) for FER with image and its extension TMSAU-Net (right) for FER with video. MSAU-Net consists of multi-scale branch networks and each branch includes three modules: AU detection, feature embedding and recognition module, which are depicted in red dotted line. The MAC descriptor is extracted from the final convolutional layer, which is depicted in black dotted line instead of black solid line. TMSAU-Net is composed of a spatial stream net and a temporal stream net to suit for FER with video task, where cascaded attention blocks is used for extracting more distinctive spatial image representations and multi-frame optical flow is adopted as temporal features. The attention mechanism module and temporal stream net are depicted in green dotted line. Besides, boxes of different color represents different functional network layer or module. Specifically, in the illustration of MSAU-Net, convolutional layer, region layer, pooling layer and feature representation (including fully-connected feature and MAC) are depicted as purple, green, blue and white boxes while losses are depicted as orange ones. As for TMSAU-Net, like MSAU-Net, feature representation and losses are presented as white and orange boxes while other functional modules like CNN-based module (including MSAU-Net and CNN), attention blocks and optical flow are presented as purple, green and blue ones.

the information of AU, we derive a recurrent attention CNN framework similar to [26] that is firstly applied for fine-grained image classification by incorporating the information of different scales. However, in MSAU-Net, we simply “zoom in” the most discriminative facial region guided by the pixel which holds the largest activation among all the activations across all the feature maps of the final convolutional layer, *i.e.*, the pixel with the maximum value among elements of the aforementioned MAC descriptor. The most discriminative facial region is essentially the receptive field of this pixel, *i.e.*, a certain area mapped to original image.

Furthermore, as shown in Fig.4 (left), our MSAU-Net is comprised of two recognition branches for incorporating two scales. These two branches are alternatively optimized by softmax losses and pairwise ranking loss. The aim of ranking loss is to enforce the finer-scale branch to generate more confident predictions over a pre-defined margin than the coarser ones. In other words, the finer-scale branch contributes more to the network. Specifically, the pairwise ranking loss is defined as:

$$\mathcal{L}_{\text{rank}}(p_t^{(s)}, p_t^{(s+1)}) = max\{0, p_t^{(s)} - p_t^{(s+1)} + margin\}, \tag{1}$$

where $p_t^{(s)}$ denotes the prediction probability on the correct class labels t and s denotes the scale. In our case, s equals to 1.

Each branch consists of three modules: AU detection module, feature embedding module and recognition module. For AU detection module, like DRML, we employ a multi-label sigmoid cross-entropy loss. Let the number of AUs be C, the number of samples be N, the ground truth $Y \in \{-1, 0, 1\}^{NC}$ where Y_{ij} indicates the (i, j) element of Y, *i.e.*, the j_{th} AU of the i_{th} sample, and the predictions $\widehat{Y} \in \mathbb{R}^{N\times C}$. The loss functions is formulated as follows:

$$\mathcal{L}(\mathrm{Y}, \widehat{\mathrm{Y}}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\{[Y_{nc} > 0]log\widehat{Y}_{nc} + \\ [Y_{nc} < 0]log(1 - \widehat{Y}_{nc})\}, \tag{2}$$

where $[x]$ is an indicator function. Note that the value 0 in ground truth {-1,0,1} is used for data distribution balance during training via disabling multi-label sigmoid cross entropy loss to reduce the number of negative or positive samples. The critical part of DRML network is the introduction of region layer whose weights are shared only within specific facial regions instead of entire feature maps aiming to capture local appearance changes for different facial regions.

Afterwards we conduct feature embedding following the AU detection module. The image representation is constructed as a concatenated feature vector of final fully-connected feature andthe MAC descriptor, an easily-implemented and effective local feature motivated by deep-based content instance image retrieval researches, which is concatenated with the maximum value of the feature maps of the final convolutional layer. In other words, the dimension of MAC equals to the number of channels of the final convolutional layer, *i.e.*, 256. That means the total dimension of feature vectors is the sum of the 2048 and 256. This way of generating features can extract local features from activated AU regions, rather than regarding the action unit as supervisory signals like previous works.

As for the recognition module, we made a simple yet effective change contributing to achieve a promising result on our *FG-Emotions* image dataset. Considering the expression labels are organized as hierarchy, two softmax loss functions are jointly trained where one is for six basic classes and the other one is for thirty three classes. We assume that by reducing the error with reference to six basic expression classification, the accuracy of thirty three expression classification would increase relatively. In this paper, we use the notion $32 \times 11 \times 11 \times 3$ to represent the parameters of a convolutional layer, where 32 denotes the number of kernels and $11 \times 11 \times 3$ indicates the kernel size. Besides, we adopt the training procedure like this:

1. We first fine-tune the AU detection module with multi-label loss on BP4D [46] and DISFA dataset [45] and fix its weight;

2. Then we fine-tune the other two modules with two softmax losses and pairwise ranking loss, the weight of each equals to 0.5, 0.5 and 0.5.

B. TMSAU-Net for Fine-Grained FER with Video

On the basis of MSAU-Net, we extend it by making several modifications: 1) utilizing attention mechanism to concentrate on the key frame of the video clip; 2) taking the temporal and the spatial information into account simultaneously on videos. Observe that the extended version can fit the tasks for fine-grained FER with both image and video. Once MSAU-Net is fine-tuned, the weights of which can be directly shared in TMSAU when testing, considering that the attention mechanism module and temporal stream can be plugged in or out flexibly. Fig. 4 (right) illustrates the TMSAU-Net comprised of temporal stream net and spatial stream net.

TMSAU-Net is jointly trained by the final class score fusion with equal wights 0.5 on two stream networks for the fusion of temporal and spatial features. The architecture of the entire network is related to[28] [34] targeting at analyzing action in videos. The loss functions used in our network are softmax losses, the settings of which are the same as MSAU-Net. Temporal stream net takes multi-frame optical flow of raw videos as input, while spatial stream net attains more efficient feature embedding and aggregation by selecting video frames with better quality over attention block mechanism. Normally an input video may contain frames with varying lighting, resolution, head pose *etc.*, so the aim of attention block mechanism is to prevent poor face images from interfering the recognition. Here we describe the attention block mechanism in brief which consults the pipeline of face verification [25].

We assume a raw input video contains n frames. First features are extracted for each of n frames by MSAU-Net, where the dimension of feature vectors for a single frame is 256. An attention block reads all feature vectors from GoogleNet and generate linear weights for them, where feature vectors are denoted by $\{f_k\}$ and corresponding weights are depicted by by $\{a_k\}$. Therefore for a video, the final feature vector through the aggregation of attention block mechanism is the weighted sum of features from all the video frames, that is $r = \sum_{i=1}^{n} a_k f_k$. The dimension of final features extracted by spatial stream net also equals to 256.

Here we adopt cascaded two attention blocks to perform the aggregation of features $\{f_k\}$. The first attention block filters $\{f_k\}$ with a kernel q via dot product, yielding a set of corresponding significances $\{e_k\}$. They are then cast to a softmax operator to generate positive weights $\{a_k\}$ with $\sum_k a_k = 1$. These two operations are formulated as below.

$$e_k = q^T f_k, \tag{3}$$

$$a_k = \frac{exp(e_k)}{\sum_j exp(e_j)}. \tag{4}$$

Then we compute the weights of the second attention block. Let q_0 be the kernel of the first attention block, and r_0 be the aggregated feature with q_0. We adaptively compute q_1, the kernel of the second attention block, through a transfer layer taking r_0 as the input:

$$q_1 = tanh(Wr^0 + b), \tag{5}$$

where W and b are the weight matrix and bias vector of the neurons respectively, and $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ imposes the hyperbolic tangent nonlinearity. The feature vector r_1 generated by q_1 will be the final aggregation image representations.

V. EXPERIMENTS

We evaluate the proposed method on both our *FG-Emotions* dataset and public six basic expression datasets. Our models MSAU-Net and TMSAU-Net achieve pretty good performance on *FG-Emotions*, and they are also suited for six expression classification, the accuracies of which exceed the-state-of-the-art works.

A. Expression Datasets and Evaluation Metrics

Our *FG-Emotions* dataset has been elaborated in Section III, we then provide an overview of several publicly available databases labeled with seven facial expressions, *i.e.*, six basic classes plus neutral.

CK+ [5]. The extended **C**ohn**K**ande (CK+) database is the most extensively used laboratory-controlled database for evaluating FER systems. CK+ contains 593 video sequences, each of which includes a complete process shifting from neutral expression to the peak expression. The sequences are from123 subjects and the length varies from 10 to 60 frames. Seeing that CK+ does not provide specified training, validation and test sets, therefore we assume the most common data selection method, which is to extract the last one to three frames with peak formation and the first frame (neutral face) of each sequence. Next the subjects are divided into 10 groups for person-independent 10-fold cross-validation experiments.

MMI [6][7]. The MMI database is also laboratory-controlled and consists of 326 sequences from 32 subjects. In contrast to CK+, sequences in MMI are in the form of onset-apex-offset, *i.e.*, the sequence starts at a neutral face, then reaches peak and finally returns to the neutral face again. For experiments, we choose the first frame (neutral face) and the three peak frames in each frontal sequence to conduct person-independent 10-fold cross-validation.

Oulu-CASIA [8]. The Oulu-CASIA database includes 2,880 sequences collected from 80 subjects. Each of the videos is captured with one of two imaging systems: **n**ear-**i**nf**r**ared (NIR) or **vis**ible light (VIS). The form of the videos is similar to CK+ which starts at a neutral expression and ends with the peak expression. We follow the typical 10-fold cross-validation experiments by collecting last three peak frames and the first frame (neutral face) from the 480 videos via the VIS System.

AFEW [9]. The **A**cted **F**acial **E**xpressions in the **W**ild (AFEW) has served as an evaluation platform for the annual

TABLE II
FER QUANTITATIVE COMPARISON ON FG-EMOTIONS

Source	Method	anger	contempt	disgust	frustration	hate	wrath	anxiety	panic	terror	worry	pride
Images	LBP+SVM	.643	.667	.611	.625	.750	.602	.536	.563	.538	.541	.703
	Plain CNN	.675	.694	.631	.648	.746	.625	.570	.579	.575	.583	.724
	InsightFace	.695	.685	.667	.684	.781	.672	.650	.628	.635	.624	.739
	AU+CNN	.714	.707	.682	.690	.790	.668	.659	.641	.656	.635	.735
	AU+Z-CNN	.722	.716	.687	.707	.801	.683	.673	.659	.672	.648	.747
	MSAU-Net (Ours)	**.735**	**.729**	**.713**	**.728**	**.811**	**.702**	**.685**	**.673**	**.692**	**.663**	**.768**
Videos	DenseTraj	.611	.575	.412	.710	.545	.572	.464	.571	.500	.503	.571
	Plain TSNet	.655	.618	.567	.738	.613	.628	585	.629	.562	.571	.636
	NAN	.663	.631	.559	.751	.611	.624	.609	.638	.557	.573	.646
	TMSAU-Net (Ours)	**.702**	**.669**	**.603**	**.772**	**.645**	**.664**	**.640**	**.673**	**.598**	**.612**	**.691**
Source	Method	relief	thrill	lust	affection	gloom	guilt	sorrow	remorse	surprise	amazement	suffering
Images	LBP+SVM	.682	.714	.751	. 750	.556	.545	.553	.563	.682	.770	.714
	Plain CNN	.693	.737	.764	.765	.610	.579	.585	.587	.698	.799	.738
	InsightFace	.703	.752	.783	.779	.647	.610	.605	.615	.714	.806	.751
	AU+CNN	.732	.767	.801	.794	.690	.639	.626	.642	.733	.825	.766
	AU+Z-CNN	.745	.779	.813	.815	.704	.658	.647	.669	.753	.832	.764
	MSAU-Net (Ours)	**.757**	**.803**	**.821**	**.833**	**.723**	**.682**	**.667**	**.688**	**.774**	**.858**	**.780**
Videos	DenseTraj	.584	.593	.545	.597	.537	.541	.521	.543	.687	.674	.545
	Plain TSNet	.633	.647	.591	.635	.570	.579	.576	568	.733	.719	.581
	NAN	.654	.662	.603	.629	.568	.589	.572	.575	.745	.733	576
	TMSAU-Net (Ours)	**.702**	**.710**	**.646**	**.667**	**.613**	**.631**	**.615**	**.609**	**.783**	**.768**	**.604**
Source	Method	pleasure	optimism	torment	loathing	irritation	depression	astonishment	cheerfulness	enthrallment	embarrassment	nervousness
Images	LBP+SVM	.565	.688	.583	.675	.680	.575	.625	.568	.750	.598	.587
	Plain CNN	.614	.694	.604	.693	.702	.606	.667	.590	.769	.601	.593
	InsightFace	.657	.705	.612	.712	.696	.648	.678	.605	.759	.623	.618
	AU+CNN	.679	.729	.645	.738	.745	.650	.710	.636	.781	.616	.643
	AU+Z-CNN	.726	.736	.668	.753	.750	.675	.723	.682	.813	.635	.656
	MSAU-Net (Ours)	**.738**	**.754**	**.701**	**.779**	**.764**	**.702**	**.746**	**.698**	**.827**	**.659**	**.678**
Videos	DenseTraj	.500	.625	.582	.515	.591	.511	.550	.588	.523	.572	.568
	Plain TSNet	.553	.657	.548	.635	.628	.549	.574	.613	.569	.594	.587
	NAN	.568	.679	.627	.579	.667	.567	.592	.635	.583	.631	.603
	TMSAU-Net (Ours)	**.604**	**.706**	**.648**	**.613**	**.696**	**.601**	**.623**	**.673**	**.624**	**.672**	**.659**

Emotion Recognition in the **W**ild Challenge (EmotiW) since 2013. The latest version AFEW 7.0 is divided into three data partitions in terms of subject and movie/TV sources: Train (773 samples), Validation (383 samples) and Test (653 samples). All of the videos provide various environmental conditions in both audio and video.

SFEW [47]. The **S**tatic **F**acial **E**xpressions in the **W**ild (SFEW) was collected by selecting key frames from the AFEW dataset via facial point clustering. The most commonly used version, SFEW 2.0, has been split into three sets: Train (958 samples), Validation (436 samples) and Test (372 samples). Each of the images is assigned to one of seven expression categories, *i.e.*,basic classes plus neutral. Regarding that the testing set are held back by the challenge organizer, we compare our model with other methods on its validation set.

RAF-DB [51]. The **R**eal-world **A**ffective **F**ace **D**ata**b**ase (RAF-DB) contains 29,672 highly diverse facial images downloaded from the Internet. The samples are annotated by one of 7 basic and 11 compound emotion labels with manually crowd-sourced annotation and reliable estimation. In particular, 15,339 images from the basic emotion set are partitioned into two groups, *i.e.*, 12,271 training images and 3,068 test images.

AffectNet [37]. AffectNet includes more than one million real-world images from the Internet in total by querying different search engines using emotion-related tags. It is by far the largest database that provides facial expressions in two different emotion models (categorical model and dimensional model), of which 450,000 images with manually annotated eight basic expression labels are used for categorical model.

FER2013 [58]. The FER2013 database was first introduced during the ICML 2013 Challenges in Representation Learning. FER2013 is a large-scale and unconstrained database collected automatically by Google image search API, which contains 28,709 training images, 3,589 validation images and 3,589 test images with seven expression labels (*i.e.*, anger, disgust, fear, happiness, sadness, surprise and neutral).

Evaluation metrics. The performance is evaluated on two common metrics: classification accuracy and confusion matrix, which are widely used in FER. For each reimplemented method and baseline, we computed average accuracy over all the fine-grained classes.

B. Implementation Details

Online data augmentation is performed like other FER methods during training to alleviate overfitting: we randomly crop the input images from the four corners and center region, then flipped horizontally. Meanwhile, we adopt such prediction mode for test: only the center patch of the face is used for prediction. In addition, two softmax losses and multi-label loss are calculated at the same time for each batch on images. For FER with video, we train the TMSAU-Net via the fusion of softmax losses given the fixed weights of MSAU-Net on images. All models are initialized with learning rate of 0.0001. A momentum of 0.9 and weight decay of 0.0005 is used. Our method is implemented by extending the Pytorch framework [44]. All networks are trained on four TITAN XP GPUs. Due to the GPU memory limitation, the batch size is set to be 8.

C. *Results on* FG-Emotions

Facial expression recognition with image. We first evaluate MSAU-Net over all the thirty three classes on the augmented image dataset of *FG-Emotions*. And Tab. II shows the experimental comparisons of classification accuracy between our model and baselines, where the first three rows indicate the baseline models, while the last three rows are designed to prove the effectiveness of crucial modules in our framework, *i.e.*, AU detection backbone, “zoom in” operation and modified softmax loss. Among the last three models, AU + CNN means that the “zoom in” operation is removed while AU + Z-CNN includes “zoom in” operation. Meanwhile, both AU + CNN and AU + Z-CNN is directly connected to one softmax loss for thirty three classes without finetuning of softmax loss for six basic classes. MSAU-Net indicates the method with all the three key modules.

From Tab. II we can see that, of three baseline models, InsightFace outperforms the other two models. We argue that InsightFace, the best model for face recognition and verification, can generate a more discriminative facial representations. Furthermore, AU + CNN is slightly better than InsightFace at around 1 to 2 percent for most categories. The assumption is that for FER task, aggregation of local features concentrating on AU regions achieves more promising results than other models in regardless of AU face regions. The accuracies of AU + Z-CNN are higher than those of AU + CNN proves that the “zoom in” operation is reliable, since multi-scale faces can offer finer information of face regions. The results of last row suggest that our framework with a minor change on loss functions is the best option for the task of fine-grained FER. Note that the accuracies are raised by around 3 to 5 percent compared to InsightFace. In conclusion, our proposed method explores one way of aggregating local features and constraining the relationships among labels while the overall performance of our model shows there exists room to be improved, especially for certain unpleasant classes such as “disgust” and “depression”.

Facial expression recognition with video. We compared TMSAU-Net with alternative variants as well as a baseline SVM which takes dense trajectories as features. There are two stream branches in our network, one is used for selecting frames with better quality automatically, while the other one is for extracting temporal information by utilizing optical flow. To prove the efficiency of our model, we design two variants: a plain two stream network that adopts GoogleNet as feature extractor which discards attention mechanism and NAN (*i.e.*, the spatial stream network) without class score fusion of temporal branch. From the last three rows in Tab. II, we observe that our method reaches the highest performance, which is about 4 to 5 percent higher than the other two variants.

TABLE III
QUANTITATIVE COMPARISONS ON OTHER BENCHMARKS FOR FER WITH IMAGE

Method	CK+ [5]	MMI [7]	SFEW [47]	AffectNet [37]	RAF-DB Basic [51]	RAF-DB Compound [51]	FER2013 [58]
Liu *et al.* [48]	.937	.758	-	-	-	-	-
Ding *et al.* [49]	.986	-	.551	-	-	-	-
Liu *et al.* [50]	.971	.785	.542	-	-	-	-
Li *et al.* [51]	-	.784	.510	-	.742	.446	-
Yang *et al.* [52]	.973	.732	-	-	-	-	-
Zhang *et al.* [59]	-	-	-	-	-	-	.751
Pramerdorfer *et al.* [60]	-	-	-	-	-	-	.752
Mollahosseini *et al.* [37]	-	-	-	**.720**	-	-	-
MSAU-Net (Ours)	**.991**	**.865**	**.574**	.712	**.758**	**.502**	**.783**

TABLE IV
QUANTITATIVE COMPARISONS ON OTHER BENCHMARKS FOR FER WITH VIDEO

Method	CK+ [5]	MMI [7]	Oulu-CASIA [8]	AFEW 7.0 [9]
PPDN [30]	.993	-	.846	-
DCPN [31]	**.996**	-	.862	-
ST Net [32]	.975	.915	.863	.408
CNN+LSTM [33]	-	.875	-	.454
TMSAU-Net (Ours)	.995	**.991**	**.874**	**.476**

Meanwhile, these two variants also outperform the conventional method Dense Trajectories + SVM at a large scale. An interesting phenomenon deserved to be considered is that NAN variant performs slightly better than plain two stream network. That raised an assumption: more emphasis on better frames in a video can benefit the distinctiveness of features for the overall framework. Meanwhile, we discover that the classification accuracies over certain fine-grained classes are inconsistent for FER with image and video. For “hate” recognition, our model with image performs pretty well while the model with video is not so satisfying. We assume that the static features may affect more on the accuracy. Therefore, we discuss the impact of the balanced weight in fusion softmax loss later in Subsection F.

D. *Results on Other Datasets*

In this subsection, we proved the robustness of our framework through experiments on six basic facial expression datasets for FER with both image and video. To make our model comparable to state-of-the-arts, we introduce the strategy of generalizing our network for FER on six classes datasets. When assigning a image with certain label during testing, our model takes the corresponding basic class regardless of the fine-grained class. However, this operation is not suitable for “disgust” and we directly calculates its accuracy since “disgust” belongs to the basic classes in standard FER. In particular, if our network predict the test image with basic class“love”, then the test image are treated as negative for all the six basic classes.

By adopting the strategy above, The performance comparison of the proposed MSAU-Net and TMSAU-Net with state-of-the-art methods on the conventional benchmarks are reported in Tab. III and Tab. IV respectively, *i.e.*, MAP for FER with image and video. In Tab. III, RAF-DB is divided into two datasets, RAF-DB Basic and RAF-DB Compound, which include basic class annotated images and compound expression images, respectively. In Tab. IV, PPDN [30] is short for peak-piloted deep network; DCPN [31] is short for deeper cascaded peak-piloted network; STNet [32] is a network ensemble of spatial network and temporal network;

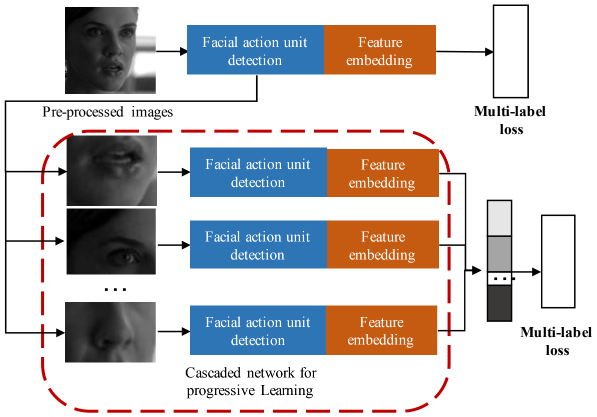

Facial action unit detection
Feature embedding
Pre-processed images
Multi-label loss
Facial action unit detection
Feature embedding
Facial action unit detection
Feature embedding
...
Facial action unit detection
Feature embedding
Multi-label loss
Cascaded network for progressive Learning

Fig. 5. Illustration of the variant for progressive learning based on our original framework.

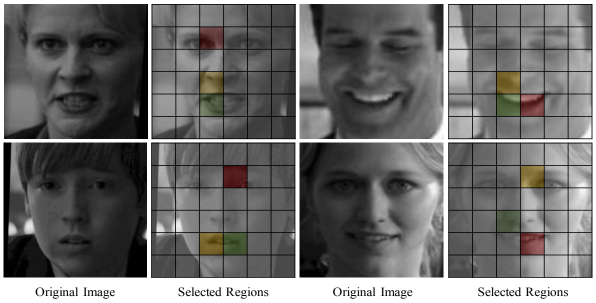


Fig. 6. Visualization of the first three AU-specific receptive fields generated by the progressive order in our framework, which are depicted in red, green and yellow respectively.

and CNN+LSTM [33] takes LSTM as temporal network. From Tab. III and Tab. IV, our model outperforms almost all the other models for FER with image and video on all the datasets except for AffectNet from Tab. III and CK+ from Tab. IV. But the accuracy of our model for video FER is close to that of the best one on AffectNet and CK+. The experimental results for FER with image clearly demonstrate the effectiveness of feature aggregation via automatically detecting AU and multi-scale architecture. Meanwhile, our model shows superior generalization ability to large-scale datasets such as FER2013 and RAF-DB from Tab. III. From the last three rows of the Tab. IV, the accuracies for STNet and CNN+LSTM is far more less than our model ranging from 1 to 10 percent, though these two models adopts different temporal branch network. Again it proves the importance of the introduction of attention mechanism.

E. Deeper Exploration on MSAU-Net

Variants of learning strategy.

In Section IV, we employ facial action unit detection network to help locate the most discriminative facial part from raw images. Nevertheless, there are other less important regions which may contribute to FER. Considering this limitation, inspired by other weakly-supervised tasks such as weakly-supervised semantic segmentation, we adopt a progressive learning strategy for "zoom in" operation. The "progressive" means that the non-overlapping facial regions are utilized during different cascaded stages by the order of discriminativeness. In other words, once a certain facial region is employed, we erase this region and turn to select another most important region and the process of this progressive learning strategy is shown in Fig. 5. The only change to the overall framework is that the final feature of the "zoom in" branch is concatenated by the features aggregated in feature embedding module for each cascaded network. The discriminative power of facial regions are measured in a simple way: the value of MAC. We conduct an ablation study on the number of discriminative regions and the results are depicted in the left chart of Fig. 7, from which we conclude that the performance reach the peak when the number of cascaded networks equals to 4, which means investigating information of local facial parts excessively can not benefit overall performance. Besides, the more networks are cascaded, the overall framework is more time-consuming. Thus we recommend that an appropriate number of cascaded network would be three or four.

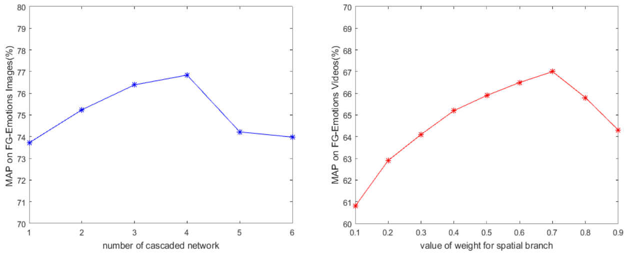


Fig. 7. The left chart illustrates the impact of the number of the cascaded networks for FER with image. The right chart illustrates the impact of the weight in class score fusion for FER with video.

Impact of AU detection. In MSAU-Net, AU detection plays a vital role in locating discriminative facial parts. From Tab. I, we observe that adding facial action unit detection can contribute 1 to 4 percent in comparison with plain CNN. We further explore where the most discriminative facia parts are. We simulate the progressive process of the automatic selection in our image framework and illustrate the first three selected regions depicted in the colored grids from Fig. 6. Noe that the receptive field of the last convolutional layer is roughly divided by 6 × 6 grids, and we relocate the corresponding regions and visualize them in the original images. Normally the important facial regions are around eyes, nose and mouth. Also, we notice that the mouth region are usually the first to be selected when the emotion is optimistic, while the eye or eyebrow regions are normally the first to be selected when the emotion is passive.

F. Deeper Exploration on TMSAU-Net

Variants of temporal-stream network backbone. In Section IV, we introduce a simple two-stream network where the optical flow is treated as temporal flow. However, there are other forms of deep network that can serve as extracting temporal information, such as LSTM [40], C3D [41] *etc.* Therefore, we carried out an ablation study where the optical flow branch is replaced with these networks as backbone.

TABLE V
MAP via different temporal backbones for FER with video. Symbol "*" denotes the re-implementation backbone network in our paper.

Datasets	Temporal Backbones		
	Optical Flow	VGG16-LSTM* [40]	C3D-LSTM* [41]
AFEW 7.0 [9]	.476	.501	.483
FG-Emotions	.659	.694	.655

The experimental results are depicted in Tab. V. From this table, we can observe that though different temporal branches can improve the recognition accuracy, the contributions of temporal branch is still lower than that of spatial branch as can be seen in Tab. I, which proves the larger influence of attention mechanism from the side.

Hyper parameter analysis. Here we analyze the hyper parameter for the class score fusion regarding spatial and temporal branch. Tab. I shows the performance utilizing the model where equal weights 0.5 are adopted for these two branches. The impact of the hyper parameter is shown in Fig. 7, and we only show how the weight of the spatial branch influence the overall FER performance with video, considering that the sum of two weights equals to 1. From the right chart of Fig. 7, at first the performance raises along with the weight increased. When the performance reaches the best, later it decreases along with the weight increased. From the peak value, we conclude that when the value of weight is around 0.7, the framework achieves the most promising results. The study on hyper parameters also proves the the large influence of attention mechanism from the side.

G. Discussion

To investigate more into the fine-grained facial expression problem, we raise the following questions and make a further analysis on these questions, which can sparkle the inspirations for later studies.

How does the network perform across all the fine-grained categories? From Tab. II, we conclude that the accuracies for more positive expressions such as "thrill", "lust" and "affection", are usually higher than those of negative expressions such as "disgust", "terror" and "anxiety". Meanwhile, in comparison to InsightFace model, the accuracies of our framework are increased by more than 5 percent on certain categories for image classification, *e.g.*, "surprise" and "amazement". The assumption for these two phenomenons is that facial muscles of positive expressions normally vary more largely than those of negative expressions, thus the image representations of negative expressions are too subtle to be recognized. The same explanation is suitable for the second phenomenon, certain expressions are more sensitive to AU detection since AU detection offers more discriminative capabilities on the extraction of local features of AU region.

For different basic expressions, how does the network perform on their fine-grained sub-categories? To answer this question, we did another experiments to compute confusion matrixes for a group of fine-grained categories which belong to the same basic expression. In confusion matrix, rows denote true category, columns indicate recognized category, and boldface specifies the best recognized categories. As we can see in the exemplar Tab. VI and Tab. VII for selected classes "sadness" and "happiness" respectively, there are more zeros in Tab. VII than in Tab. VI. That also proves negative facial expressions are more easily confused when compared to other fine-grained categories of the same basic class. As mentioned before, due to minor muscle change of action units, negative facial expressions are hardly distinguished with others. Nevertheless, the sum of non-zero values for each row is usually over 90 percent, which indicates the chance of an expression categorized beyond the range of basic class is small though it can be mistaken with other fine-grained categories of the same basic class. Note that the confusion matrix is generally a symmetric matrix, *e.g.*, "pleasure" is easily confused with "thrill" while "thrill" is easily confused with "pleasure" in turn. Apart from the analysis above, there are still more challenges about fine-grained FER in the wild. We will explore the label relationships and how the expression hierarchy effects each other in the future.

TABLE VI
Confusion Matrix for fine-grained classes of basic class "sadness".

	gloom	guilt	remorse	sorrow	suffering
gloom	**.723**	.037	.040	.052	.049
guilt	.036	**.682**	.078	.037	.044
remorse	.044	.085	**.676**	.052	.072
sorrow	.055	.038	.046	**.712**	.123
suffering	.048	.055	.063	.104	**.665**

TABLE VII
Confusion Matrix for fine-grained classes of basic class "happiness"

	thrill	relief	pride	pleasure	optimism
thrill	**.803**	0	0	.091	.048
relief	.013	**.757**	.028	.064	.059
pride	0	.035	**.768**	.026	.087
pleasure	.082	.070	0	**.782**	.048
optimism	.053	0	.095	.041	**.754**

VI. Conclusion

As far as we know, we are the first to solve the problem of fine-grained facial expression recognition in the wild and introduce a new benchmark dataset named *FG-Emotions* containing both video clips and images with 33 fine-grained labels to facilitate the future research on this topic. Moreover, we further propose MSAU-Net and TMSAU-Net to serve as strong baselines for addressing fine-grained FER. In particular, MSAU-Net focuses on locating the most discriminative facial regions by leveraging AU detection and "zoom in" operation for recognizing with image, whereas TMSAU-Net employs a two-stream structure for recognizing with video where attention mechanism is used to select key frames in a video clip and optical flow is used to capture temporal information. We carried out several ablation studies to verify the efficiency of the variants based on our original model. Extensive experiments on both our FG-Emotions and other benchmark datasets show the superiority of our method over state-of-the-arts. In future, we will continue to take efforts to construct a more comprehensive fine-grained FER benchmark dataset with more data and more detailed category annotations to further push the frontiers of fine-grained FER research.

Acknowledgment

This work was supported by the Beijing Natural Science Foundation (Grants No. 4202057, No. 4202058), the Fundamental Research Funds for the Central Universities (No. 2020JBM403) and National Natural Science Foundation of China (No. 61872032).

References

[1] S. Du, Y. Tao, and A. M. Martinez. *Compound facial expressions of emotion.* In Proceedings of the National Academy of Sciences of USA, 111(15): E1454-E1462, 2014

[2] W.G. Parro. *Emotions in Social Psychology.* In Psychology Press, p.p.102-105, 2001

[3] C. Darwin, and P. Prodger. *The expression of the emotions in man and animals.* In Oxford University Press, 1998

[4] P. Ekman, and W. V. Friese. *Constants across cultures in the face and emotion.* In Journal of personality and social psychology, 17(2):124-129, 1971

[5] 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 CVPRW, 2010

[6] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. *Web-based database for facial expression analysis.* In ICME, 2005.

[7] M. Valstar, and M. Pantic. *Induced disgust, happiness and surprise: an addition to the mmi facial expression database.* In ECCV, 2012

[8] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikaInen. *Facial expression recognition from near-infrared videos.* In IVC, 29(9):607-619, 2011

[9] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon. *From individual to group-level emotion recognition: Emotiw 5.0.* In ACM MM, 2017

[10] C. Shan, S. Gong, and P. W. McOwan. *Facial expression recognition based on local binary patterns: A comprehensive study.* In IVC, 27(6):803-816, 2009

[11] H. Wang, A. Klaser, C. Schmid, and C-L. Liu. *Action Recognition by Dense Trajectories.* In ICCV, 2011

[12] S.W. Chew, S. Lucey, P. Lucey, S. Sridharan, and J.F. Cohn. *Improved Facial Expression Recognition via Uni-Hyperplane Classification.* In CVPR, 2012

[13] B. Sun, L. Li, T. Zuo, Y. Chen, G. Zhou, and X. Wu. *Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild.* In ICML, 2014

[14] P. Liu, S. Han, Z. Meng, and Yan Tong. *Facial Expression Recognition via a Boosted Deep Belief Network.* In CVPR, 2014

[15] S. E. Kahou. *EmoNets: Multimodal deep learning approaches for emotion recognition in video.* In Journal on Multimodal User Interfaces, 10(2):99-111, 2016

[16] M. Wllmer. *LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework.* In IVC, 31(2):153-163, 2015

[17] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas. *Learning multiscale active facial patches for expression analysis.* In TCybernetics, 45(8):1499-1510, 2014

[18] K. Zhao,W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. *Joint patch and multi-label learning for facial action unit detection.* In CVPR, 2015

[19] M. Liu, S. Li, S. Shan, and X. Chen. *AU-aware deep networks for facial expression recognition.* In AFGR, 2013

[20] P. Liu, S. Han, Z. Meng, and Y. Tong. *Facial expression recognition via a boosted DBN.* In CVPR, 2014

[21] K. Zhao, W-S. Chu, and H. Zhang. *Deep Region and Multi-label Learning for Facial Action Unit Detection.* In CVPR, 2016

[22] S. Eleftheriadis, O. Rudovic, and M. Pantic. *Multiconditional latent variable model for joint facial action unit detection.* In ICCV, 2015

[23] R. Walecki, and O. Rudovic. *Deep Structured Learning for Facial Action Unit Intensity Estimation.* In ICCV, 2017

[24] J. Deng, J. Guo, and S. Zafeiriou. *ArcFace: Additive Angular Margin Loss for Deep Face Recognition.* In arXiv:1801.07698v1, 2018

[25] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and Gang Hua. *Neural Aggregation Network for Video Face Recognition.* In CVPR, 2017

[26] J. Fu, H. Zheng, and Tao Mei. *Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition.* In CVPR, 2017

[27] G. Tolias, R. Sicre, and H. Jegou. *Particular Object Retrieval With Integral Max-Pooling of CNN Activations.* In ICLR, 2016

[28] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Gool. *Temporal Segment Networks: Towards Good Practices for Deep Action Recognition.* In ECCV, 2016

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed,D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. *Going deeper with convolutions.* In CVPR, 2015

[30] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan. *Peak-piloted deep network for facial expression recognition.* In ICCV, 2016

[31] Z. Yu, Q. Liu, and G. Liu. *Deeper cascaded peak-piloted network for weak expression recognition.* In Visual Computer, 2017

[32] K. Zhang, Y. Huang, Y. Du, and L.Wang. *Facial expression recognition based on deep evolutional spatial-temporal networks.* In TIP, 26(9):4193-4203, 2017

[33] D. H. Kim, W. Baddar, J. Jang, and Y. M. Ro. *Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition.* In TAC, 2017

[34] K. Simonyan, and A. Zisserman. *Two-stream convolutional networks for action recognition in videos.* In NIPS, 2014

[35] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. *Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark.* In ICCVW, 2011

[36] S. Li and W. Deng. *Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition.* In TIP, 28(1):356-370, 2019

[37] A. Mollahosseini, B. Hasani, and M. H. Mahoor. *Affectnet: A database for facial expression, valence, and arousal computing in the wild.* In TAC, 10(1): 18-31, 2019

[38] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. *Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild.* In CVPR, 2016

[39] H. Gunes, and B. Schuller. *Categorical and dimensional affect analysis in continuous input: Current trends and future directions.* In IVC, 31(2):120-136, 2013

[40] X. Ouyang, S. Kawaai, E. G. H. Goh, S. Shen, W. Ding, H. Ming, and D.-Y. Huang. *Audio-visual emotion recognition using deep transfer learning and multiple temporal models.* In ICMI, 2017

[41] V. Vielzeuf, S. Pateux, and F. Jurie. *Temporal multimodal fusion for video emotion classification in the wild.* In ACM MM, 2017

[42] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos. *Facial expression recognition with convolutional neural networks: coping with few data and the training sample order.* In PR, 61: 610-628, 2017

[43] M. V. Zavarez, R. F. Berriel, and T. Oliveira-Santos. *Cross-database facial expression recognition based on fine-tuned deep convolutional network.* In SIBGRAPI, 2017

[44] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. *Automatic differentiation in pytorch.* 2017

[45] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. *Disfa: A spontaneous facial action intensity database.* In TAFFC, 4(2):151-160, 2013.

[46] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. *A high-resolution spontaneous 3d dynamic facial expression database.* In FGR, 2013.

[47] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. *Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark.* In ICCVW, 2011

[48] M. Liu, S. Li, S. Shan, and X. Chen. *Au-inspired deep networks for facial expression feature learning.* In Neurocomputing, 159:126-136, 2015

[49] H. Ding, S. K. Zhou, and R. Chellappa. *Facenet2expnet: Regularizing a deep face recognition net for expression recognition.* In FGR, 2017

[50] X. Liu, B. Kumar, J. You, and P. Jia. *Adaptive deep metric learning for identity-aware facial expression recognition.* In CVPRW, 2017

[51] S. Li, W. Deng, and J. Du. *Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild.* In CVPR, 2017

[52] H. Yang, U. Ciftci, and L. Yin. *Facial expression recognition by deexpression residue learning.* In CVPR, 2018

[53] W. Wang, Q. Sun, T. Chen, C. Cao, Z. Zheng, G. Xu, H. Qiu, and Y. Fu. *A Fine-Grained Facial Expression Database for End-to-End Multi-Pose Facial Expression Recognition.* In arXiv:1907.10838v1, 2019

[54] X. Wu, R. He, Z. Sun, and T. Tan. *A light cnn for deep face representation with noisy labels.* In TIFS, 13(11):2884-2896, 2018

[55] D. H. Lee and A. K. Anderson. *Reading what the mind thinks from how the eye sees.* In Psychological Science, 28(4):494-503, 2017

[56] N. Sun, Q. Li, R. Huan, J. Liu and G. Han. *Deep spatial-temporal feature fusion for facial expression recognition in static images.* In PR Letters, 119:49-61, 2017

[57] X. Liu, B. V. K. Vijaya Kumar, P. Jia and J. You. *Hard negative generation for identity-disentangled facial expression recognition.* In PR, 88:1-12, 2019

[58] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler and D.-H. Lee. *Challenges in representation learning: A report on three machine learning contests.* In NIPS, 2013

[59] Z. Zhang, P. Luo, C.-C. Loy and X. Tang. *Learning social relation traits from face images.* In ICCV, 2015

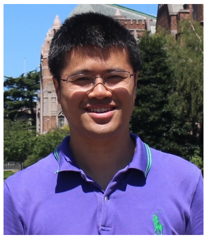
[60] C. Pramerdorfer and M. Kampel. *Facial expression recognition using convolutional neural networks: State of the art.* In arXiv:1612.02903, 2016

[61] R. Plutchik and H. Kellerman. *Emotion: Theory, research, and experience.* In Theories of emotion, 1980.

Liqian Liang received the BSc degree in Computer cience from the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China, in 2014. Currently, she is working toward the PhD degree in the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. She has been a visiting scholar in the School of Computer Science, The University of Adelaide, Australia, from 2016 to 2017. Her research interests include computer vision and machine learning. She is a student member of the IEEE.

Congyan Lang received the Ph.D. degree from the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China, in 2006. She was a Visiting Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, from 2010 to 2011. From 2014 to 2015, she visited the Department of Computer Science, University of Rochester, Rochester, NY, USA, as a Visiting Researcher. She is currently a Professor with the School of Computer and Information Technology, Beijing Jiaotong University. Her current research interests include multimedia information retrieval and analysis, machine learning, and computer vision.

Yidong Li is the Vice-Dean and a professor in the School of Computer and Information Technology at Beijing Jiaotong University. Dr. Li received his B.Eng. degree in electrical and electronic engineering from Beijing Jiaotong University in 2003, and M.Sci. and Ph.D. degrees in computer science from the University of Adelaide, in 2006 and 2010, respectively. Dr. Li's research interests include big data analysis, privacy preserving and information security, data mining, social computing and intelligent transportation. Dr. Li has published more than 80 research papers in various journals and refereed conferences. He has also co-authored/co-edited 5 books (including proceedings) and contributed several book chapters.



Songhe Feng received the Ph.D. degree in the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, P.R. China, in 2009. He is currently a Professor in the School of Computer and Information Technology, Beijing Jiaotong University. He has been a visiting scholar in the Department of Computer Science and Engineering, Michigan State University, USA, from 2013 to 2014. His research interests include computer vision and machine learning.

Jian Zhao received the Bachelor's degree from Beihang University in 2012, the Master's degree from National University of Defense Technology in 2014, and the Ph.D. degree from National University of Singapore in 2019. He is currently an Assistant Professor with Institute of North Electronic Equipment, Beijing, China. His main research interests include Deep Learning, Pattern Recognition, Computer Vision and Multimedia Analysis.