Visual-Textual Attribute Learning for Class-Incremental Facial Expression Recognition

Yuanling Lv, Guangyu Huang, Yan Yan, *Senior Member, IEEE,* Jing-Hao Xue, *Senior Member, IEEE,* Si Chen, *Member, IEEE,* and Hanzi Wang, *Senior Member, IEEE*

***Abstract*—In this paper, we study facial expression recognition (FER) in the class-incremental learning (CIL) setting, which defines the classification of well-studied and easily-accessible basic expressions as an initial task while learning new compound expressions gradually. Motivated by the fact that compound expressions are meaningful combinations of basic expressions, we treat basic expressions as attributes (i.e., semantic descriptors), and thus compound expressions are represented in terms of attributes. To this end, we propose a novel visual-textual attribute learning network (VTA-Net), mainly consisting of a textual-guided visual module (TVM) and a textual compositional module (TCM), for class-incremental FER. Specifically, TVM extracts textual-aware visual features and classifies expressions by incorporating the textual information into visual attribute learning. Meanwhile, TCM generates visual-aware textual features and predicts expressions by exploiting the dependency between textual attributes and category names of old and new expressions based on a textual compositional graph. In particular, a visual-textual distillation loss is introduced to calibrate TVM and TCM during incremental learning. Finally, the outputs from TVM and TCM are fused to make a final prediction. On the one hand, at each incremental task, the representations of visual attributes are enhanced since visual attributes are shared across old and new expressions. This increases the stability of our method. On the other hand, the textual modality, which involves rich prior knowledge of the relevance between expressions, facilitates our model to identify subtle visual distinctions between compound expressions, improving the plasticity of our method. Experimental results on both in-the-lab and in-the-wild facial expression databases show the superiority of our method against several state-of-the-art methods for class-incremental FER.**

***Index Terms*—Facial expression recognition, Class-incremental learning, Multi-modality learning, Attribute learning.**

I. Introduction

With the recent advance of deep learning, a large number of facial expression recognition (FER) methods [1]–[5] have been developed and achieved promising performance in unconstrained environments. These methods mainly focus on the classification of basic expressions (including angry, disgusted, fearful, happy, sad, surprised, and neutral) according to Ekman and Friesen's pioneering study [6]. Unfortunately, basic expressions cannot completely characterize the diversity and complexity of human emotions in real-world scenarios. Later, Du *et al.* [7] define compound expressions, which are meaningful combinations of basic expressions (e.g., the happily-surprised expression can be viewed as a combination of the happy and surprised expressions). Compared with basic expressions, compound expressions often involve more subtle visual distinctions.

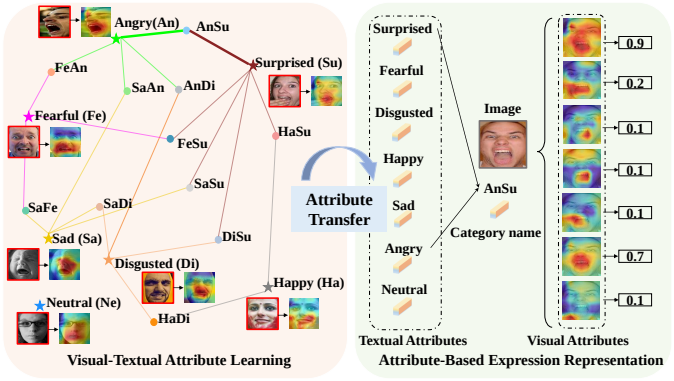

Fig. 1. Illustration of our motivation. We model the intrinsic relationship between basic expressions and compound expressions from the perspective of visual-textual attribute learning. In this way, each old/new expression image and its category name can be represented in terms of visual attributes and textual attributes, respectively.

In many practical applications, new compound expression data usually arrive sequentially since collecting all the expression categories at once is difficult. Conventional FER methods need to retrain the whole model by combining old and new data. However, such a way is time-consuming and sometimes infeasible, especially when we have no access to all the old training data due to memory limitations or data restrictions. Only fine-tuning the model with new data renders it prone to catastrophic forgetting of old classes (which refers to the drastic performance drop on previously learned old classes after learning new classes). This phenomenon can be ascribed to the stability-plasticity dilemma (i.e., the balance between accommodating new classes while retaining old classes) [8].

To address the above challenges, class-incremental learning (CIL), which requires the model to have the ability to acquire new knowledge from new classes while retaining previously learned concepts, is becoming a hot research topic. A number of CIL methods [9]–[11] have been proposed. In this paper,

This work was partly supported by the National Natural Science Foundation of China under Grants 62372388, 62071404, and U21A20514. (*Corresponding author: Yan Yan.*)

Y. Lv, G. Huang, Y. Yan, and H. Wang are with the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, and the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen 361005, China (e-mail: lvyuanling@stu.xmu.edu.cn; 23020211153933@stu.xmu.edu.cn; yanyan@xmu.edu.cn; hanzi.wang@xmu.edu.cn).

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, UK (e-mail: jinghao.xue@ucl.ac.uk).

S. Chen is with the School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China (e-mail: chensi@xmut.edu.cn).

we study the class-incremental FER task, which aims to classify both basic and compound expressions under a CIL paradigm. In particular, we define well-studied and easily-accessible basic expressions as initial classes while learning new compound expressions incrementally.

For our class-incremental FER task, due to subtle visual distinctions between compound expressions, depending only on the visual modality may not guarantee satisfactory performance. As a common sense of human vision, humans can easily identify expressions by using not only the visual modality, but also the textual, voice, and gesture modalities. Hence, we take advantage of both visual and textual modalities (i.e., images and category names) for class-incremental FER to mitigate the stability-plasticity dilemma. In fact, the textual modality involves rich prior knowledge of the relevance between expressions. Thus, we are able to leverage such knowledge as an auxiliary signal to guide the incremental training process, alleviating the forgetting of old classes when learning new classes.

As we mentioned before, compound expressions are meaningful combinations of basic expressions. Inspired by the RGB color model (i.e., each color is represented in terms of RGB components), we consider each basic expression at the initial task as an attribute (i.e., the semantic descriptor which captures some facial characteristics with inherent stability) while compound expressions are described in terms of attributes. Accordingly, we define the latent visual features extracted from basic expression images as visual attributes and the textual word vectors from their category names as textual attributes. In this way, each old/new expression image and its category name can be represented in terms of visual attributes and textual attributes, respectively. Fig. 1 illustrates our motivation for introducing visual-textual attribute learning, which serves as the basis of our proposed method for class-incremental FER.

To this end, we propose a novel visual-textual attribute learning network (VTA-Net) for class-incremental FER. VTA-Net mainly consists of a textual-guided visual module (TVM) and a textual compositional module (TCM). Specifically, TVM extracts textual-aware visual features and performs expression classification by incorporating the textual modality, where a novel textual-guided loss is developed for visual attribute learning. TCM gives expression prediction results based on visual-aware textual features by considering the visual modality. TCM effectively models the dependency between textual attributes and category names of old and new expressions based on a textual compositional graph. TVM and TCM are complementary to each other, where an effective visual-textual distillation loss is introduced to calibrate them. Finally, we combine the outputs from TVM and TCM to obtain the final classification results.

On the one hand, at each incremental task, we can leverage well-trained visual attributes guided by textual information to discover subtle visual distinctions between compound expressions, enabling the model to easily adapt to new classes. On the other hand, since visual attributes are shared across old and new expressions, learning new compound expressions is beneficial to enhance the representations of visual attributes. This in turn can reduce the forgetting of old classes.

The main contributions of this paper are as follows:

- We propose a novel VTA-Net method for class-incremental FER, where we effectively take advantage of visual and textual modalities to relieve the stability-plasticity dilemma.
- We model the intrinsic relationship between basic and compound expressions from the perspective of visual-textual attribute learning. Therefore, we can represent old and new expressions by attributes in a simple and unified way. Based on visual and textual attributes, we develop TVM and TCM to extract textual-aware visual features and visual-aware textual features, respectively, for classifying expressions.
- We perform extensive experiments on both in-the-lab and in-the-wild facial expression databases to show the effectiveness of our method against several state-of-the-art methods. This clearly validates the potential of visual-textual attribute learning in our task.

The remainder of this paper is organized as follows. First, we review the related work in Section II. Then, we elaborately describe our proposed method in Section III. Next, we perform extensive experiments on three facial expression databases in Section IV. Finally, we draw the conclusion in Section V.

II. Related Work

In this section, we first introduce facial expression recognition in Section II-A. Then, we briefly review class-incremental learning in Section II-B.

A. Facial Expression Recognition (FER)

A variety of FER methods [2]–[4], which aim to classify an input facial image into one of the basic expressions, have been developed and shown outstanding performance. However, basic expressions cannot comprehensively describe the diversity of human emotions. Du *et al.* [7] define 22 expression categories (consisting of basic and compound expressions) and reveal that compound expressions can be viewed as combinations of basic expressions. Such combinations are consistent with the subordinate categories involved. For example, the happily-surprised expression involves muscle movements observed in the basic expressions of happiness and surprise. Meanwhile, the differences between compound expressions are sufficient to distinguish. Later, a large-scale dataset EmotioNet [12] and a real-world dataset RAF-DB [13] are collected to involve both basic and compound expressions, paving the way for more practical applications of FER. Compared with basic expressions, the visual differences between compound expressions are more subtle. Thus, it is vital to extract fine-grained features for identifying compounding expressions. Guo *et al.* [14] learn the appearance and geometric representations for compound FER while Zhang *et al.* [15] develop a two-stage recognition strategy (including coarse and fine stages) to perform compound expression recognition. Zou *et al.* [16] perform compound FER under the cross-domain few-shot learning setting.

Recently, some methods [17]–[20] explore the facial expression generation and achieve outstanding performance. Wu *et al.* [17] utilize local focuses to preserve details and suppress overlapping artifacts for realistic performance, while Ma *et al.* [20] leverage parametric 3D facial representations and achieve high-quality facial expression transfer. Otberdout *et al.* [21] remove the constraint of the 4D sequence and address 4D facial expression generation. Although expression generation can enlarge the diversity of compound expression images, the quality of the generated expression images may affect the final recognition performance.

Due to the ever-changing environment, training a model on all the expressions at once is struggling and impractical. To fit for more practical applications, we introduce the CIL paradigm to FER, so that we are able to perform the classification of both basic and compound expressions when expression data are collected continuously. Note that Zhu *et al.* [22] also perform FER in the CIL paradigm and develop a center-expression-distilled loss. Regrettably, they focus on only basic expressions and do not fully exploit the intrinsic relationship between expressions. In this paper, we study FER in a different but more practical CIL setting, which considers both basic and compound expressions. Notably, we leverage basic expressions as initial classes and continually learn new compound expressions. Meanwhile, we propose to make use of the textual modality to improve the performance for classifying both old and new expressions during incremental learning.

Note that Chen *et al.* [23] also combine text and images for FER. They first compute the distances between the word embedding and the image embedding, and then transform these distances to weights. These weights are used as prior knowledge for classifying expressions. Unlike this method, we investigate the textual information based on attribute learning and model the intrinsic relationship between different expressions. In this way, we can establish a simple and unified way to represent old/new expressions at each incremental task.

B. Class-Incremental Learning (CIL)

Existing CIL methods can be roughly divided into three groups: regularization-based, distillation-based, and structure-based. Regularization-based methods [24], [25] often impose constraints on the model when learning new class data. Distillation-based methods [9], [10], [26], [27] leverage knowledge distillation to enforce the outputs of old classes on the current model to be similar to those on the previous model. iCaRL [9] uses a distillation loss to retain the knowledge of old classes, while PODNet [10] proposes a novel spatial knowledge distillation. Structure-based methods [11], [28], [29] introduce new modules to improve the capacity for learning new classes. DER [28] proposes dynamically expandable representations for incremental learning and introduces a channel-level mask-based pruning strategy to reduce the parameters. Later, FOSTER [11] employs gradient boosting to dynamically expand and compress the model while MEMO [29] measures the influence of different layers in the model and expands the specialized block incrementally.

Although the above methods have made impressive progress, they ignore the importance of textual information that can be accessed easily and contains rich information. Recently, Ali *et al.* [30] explore word embeddings of class names as the textual information to facilitate the distillation process while Li *et al.* [31] propose to leverage the textual information by adopting the memory prompt for few-shot image classification tasks. Inspired by these methods, we introduce the textual modality (which contains prior knowledge of old and new expressions) as an auxiliary signal for our class-incremental FER task.

Note that Ali *et al.* [30] utilize word embeddings as semantic information to facilitate the distillation calculation. In contrast, in our paper, the expressions are represented by attributes. Such a way can effectively preserve the relationships between expressions during the incremental learning process. Based on this, we model the dependency between textual attributes and category names, explicitly guiding visual attribute learning. Moreover, the learning strategies between our method and [30] are different. In [30], the authors apply k-means to assign a superclass label to each class, while we utilize a textual compositional graph in TCM and a textual-guided loss in TVM to characterize the semantic correlation explicitly.

III. Methodology

In this section, we introduce our proposed VTA-Net method for class-incremental FER. First, we give the problem formulation in Section III-A. Then, we provide the overview of VTA-Net in Section III-B. Next, we introduce the key components (including the TVM module and the TCM module) of VTA-Net in Sections III-C and III-D, respectively. Subsequently, we give the joint loss in Section III-E. Finally, we summarize the overall training in Section III-F.

A. Problem Formulation

In many real-world applications, expression data often come in the stream format with emerging new compound expressions. To accommodate these applications, we investigate a practical setting for class-incremental FER, which defines the classification of well-studied and easily-accessible basic expressions as an initial task and the learning of compound expressions continuously as incremental tasks.

Assume that there are a sequence of N+1 incremental tasks $\{\mathcal{D}^0, \cdots, \mathcal{D}^N\}$ without overlapping classes, where $\mathcal{D}^n$ denotes the n-th incremental task ($\mathcal{D}^0$ denotes the initial task). Similar to [9], we adopt the rehearsal strategy, which stores a tiny number of exemplars from old classes ever seen as memory and fixes them during the incremental process. At the n-th incremental task, the model is trained with a set of expression samples $\mathcal{B}_n = \{(\mathbf{x}_i^n, y_i^n, \mathbf{t}_i^n)\}_{i=1}^B$, which consist of the exemplars from old classes and all the samples from new classes. Here, B is the number of samples; $\mathbf{x}_i^n \in \{\mathcal{E}^n \cup \mathcal{D}^n\}$, $y_i^n \in \mathcal{Y}^n$, and $\mathbf{t}_i^n$ denote the i-th input image, its ground-truth label, and the textual word vector corresponding to its category name, respectively; $\mathcal{E}^n$ denotes the exemplars of old classes at the n-th incremental task; $\mathcal{Y}^n$ denotes the label set of old classes (i.e., $\mathcal{C}_{old}^n$) and new classes (i.e., $\mathcal{C}_{new}^n$). In this paper, the textual word vector is represented by a word embedding given by the GloVe model [32].

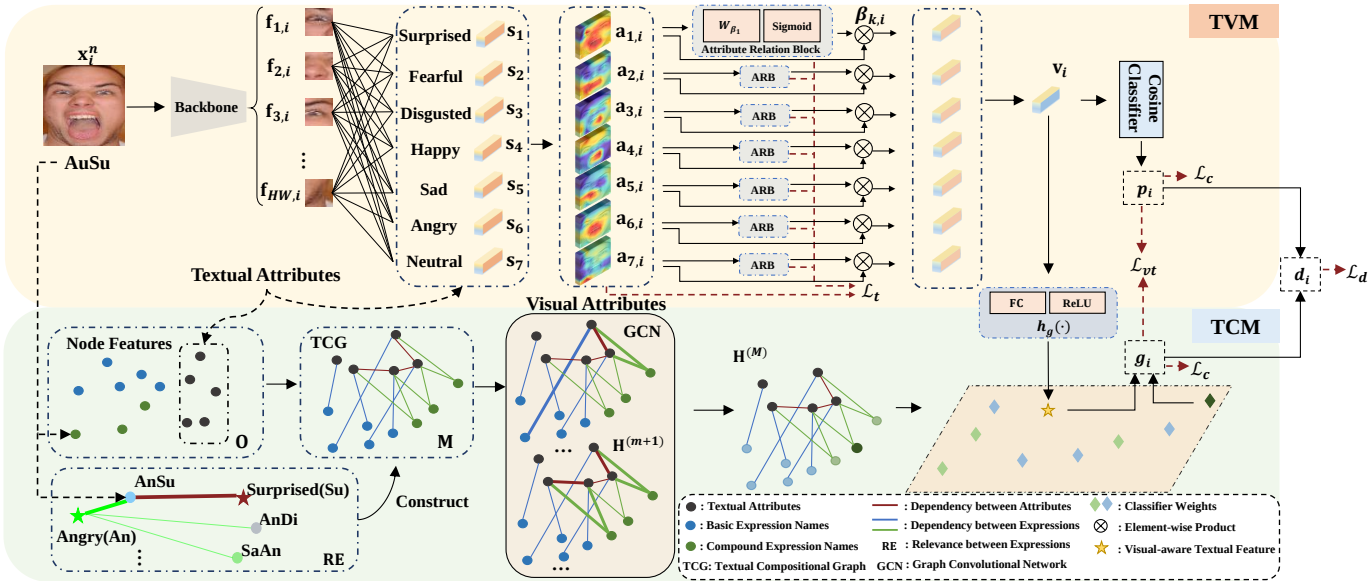

Fig. 2. **Overview of the proposed VTA-Net method**. VTA-Net is mainly composed of a textual-guided visual module (TVM) and a textual compositional module (TCM). TVM extracts textual-aware visual features and predicts expressions by incorporating the textual modality into visual attribute learning, while TCM performs expression recognition based on visual-aware textual features by additionally considering the visual modality. A visual-textual distillation loss is used to calibrate the two modules above. We use ResNet-18 as the backbone. Details of symbols are given in the text.

B. Overview

Considering the intrinsic relationship between basic and compound expressions, we view the latent visual features (which focus on the most related facial regions w.r.t one basic expression) extracted from basic expression images as visual attributes and the textual word vectors from their category names as textual attributes. Based on this, old/new expression images and their category names are respectively described in terms of visual and textual attributes that are common across expressions. In this way, the visual-textual information of expressions is appropriately modeled. Learning visual-textual attributes enables us to represent old and new expressions in a simple and unified manner during incremental learning.

The whole framework of our VTA-Net method is given in Fig. 2. VTA-Net is mainly composed of a textual-guided visual module (TVM) and a textual compositional module (TCM). TVM predicts expressions based on textual-aware visual features by additionally considering the textual modality. TCM performs expression recognition based on visual-aware textual features by additionally considering the visual modality. TVM and TCM are complementary to each other and are beneficial to discover the intrinsic relationship between old and new expressions for class-incremental FER. In particular, a visual-textual distillation loss is introduced to calibrate TVM and TCM. Finally, the outputs from TVM and TCM are simply added as the final prediction results.

C. Textual-Guided Visual Module (TVM)

TVM is designed to identify the most related facial regions for each visual attribute (which corresponds to one basic expression) and extract textual-aware visual features for expression recognition. To effectively exploit the textual modality, we leverage the textual attributes to guide visual attribute learning.

Specifically, assume that we have K (equal to the number of basic expression categories) visual attributes. Given the i-th image $\mathbf{x}_i^n$ at the n-th incremental task, a set of preliminary visual features are first extracted from the backbone. In this paper, we transform the feature map $\mathbf{F}_i \in \mathbb{R}^{H\times W\times C}$ extracted from the last convolutional layer of the backbone into a set of preliminary visual features $\{\mathbf{f}_{1,i}, \mathbf{f}_{2,i}, \cdots, \mathbf{f}_{HW,i}\}$ with $\mathbf{f}_{r,i} \in \mathbb{R}^{C\times 1}$, where H, W, and C denote the height, width, and channel number, respectively. Hence, each preliminary visual feature is constructed by concatenating the same position feature values along the channel dimension. These preliminary visual features corresponding to different facial regions (note that each pixel in the feature map corresponds to a receptive field in the original image) and a set of textual attributes $\{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_K\}$ ($\mathbf{s}_k \in \mathbb{R}^{Z\times 1}$ represents the Z-dimensional textual word vector for the k-th basic expression name) are used as the inputs of TVM.

For the k-th visual attribute, we first calculate a visual-textual compatibility score between the corresponding textual attribute and the r-th preliminary visual feature, which can be defined as

$$\alpha_{k,i}^r = \frac{\exp(\mathbf{s}_k^{\mathrm{T}}\mathbf{W}_\alpha \mathbf{f}_{r,i})}{\sum_{r'=1}^{HW}\exp(\mathbf{s}_k^{\mathrm{T}}\mathbf{W}_\alpha \mathbf{f}_{r',i})}, \tag{1}$$

where $\mathbf{W}_\alpha \in \mathbb{R}^{Z\times C}$ is a learnable matrix; $\alpha_{k,i}^r \in \mathbb{R}^{1\times 1}$ denotes the visual-textual compatibility score. The higher value of $\alpha_{k,i}^r$ indicates that the r-th preliminary visual feature (corresponding to a facial region) is more related to the k-th textual attribute.

Then, the preliminary visual features and their visual-textual compatibility scores are combined to obtain the latent visual

feature, which pays attention to the most related facial regions for the k-th visual attribute (e.g., the mouth region for the visual attribute corresponding to the surprised expression). We denote this feature as the k-th visual attribute, i.e.,

$$\mathbf{a}_{k,i} = \sum_{r=1}^{HW} \alpha_{k,i}^{r} \mathbf{f}_{r,i}, \tag{2}$$

where $\mathbf{a}_{k,i} \in \mathbb{R}^{C\times 1}$ denotes the k-th visual attribute for the i-th image $\mathbf{x}_i^n$.

Attribute Relation Block (ARB). To effectively highlight the relevant visual attributes for each expression, we apply a set of ARBs, where each block consists of a fully-connected (FC) layer and a Sigmoid activation operation. Mathematically, each block is formulated as

$$\beta_{k,i} = \sigma_1(\mathbf{W}_{\beta_k}^{\mathrm{T}} \mathbf{a}_{k,i}), \tag{3}$$

where $\beta_{k,i}$ is the attentive weight for the k-th visual attribute; $\mathbf{W}_{\beta_k} \in \mathbb{R}^{C\times 1}$ and σ_1 denote the linear transformation of the FC layer and the Sigmoid operation, respectively. Ideally, the attentive weight of one relevant visual attribute should be large for a basic expression image while the attentive weights of two relevant visual attributes should be large for a compound expression image.

Next, we combine these features and generate a textual-aware visual feature, defined as $\mathbf{v}_i = \sum_{k=1}^{K} \beta_{k,i} \mathbf{a}_{k,i}$ for the i-th input image. Finally, the textual-aware visual feature $\mathbf{v}_i$ is fed into a cosine classifier [33] to give the predicted results (denoted as $\mathbf{p}_i$).

Textual-Guided Loss. The visual differences between expressions are subtle and fine-grained. To effectively capture fine-grained distinctions, we leverage a co-occurrence loss to explicitly enforce each expression (basic or compound expression) to associate with the relevant visual attributes. Meanwhile, we expect that the different facial images from the same expression category have similar distributions of attentive weights, according to the correlation between the category name and textual attributes (since this correlation is fixed and consistent during incremental learning). Therefore, we also adopt a distribution loss. By combining the co-occurrence loss and the distribution loss, we define the textual-guided loss as

$$\mathcal{L}_t = \underbrace{-\sum_{k=1}^{K} \mathbb{1}_{[k=\tilde{y}_i^n]} \log(p_{k,i})}_{\mathcal{L}_{co}} + \underbrace{\sum_{k=1}^{K} \left(\beta_{k,i} - \tilde{\beta}_{k,i}\right)^2}_{\mathcal{L}_{dis}}, \tag{4}$$

$$p_{k,i} = \frac{\sigma_2(\mathbf{a}_{k,i})^{\mathrm{T}} \sigma_2(\mathbf{l}_k)}{||\sigma_2(\mathbf{a}_{k,i})||_2 ||\sigma_2(\mathbf{l}_k)||_2}, \tag{5}$$

where $\mathcal{L}_{co}$ and $\mathcal{L}_{dis}$ are the co-occurrence loss and the distribution loss, respectively; $\mathbf{l}_k \in \mathbb{R}^{C\times 1}$ is a learnable vector and σ_2 is the ReLU operation; $p_{k,i}$ is the occurrence probability of visual attributes by computing a cosine similarity between $\mathbf{a}_{k,i}$ and $\mathbf{l}_k$; when $k = \tilde{y}_i^n$, the function $\mathbb{1}_{[k=\tilde{y}_i^n]}$ equals to 1, otherwise its value is 0; $\tilde{\beta}_{k,i}$ denotes the correlation learned from the textual modality through a cosine similarity between the category name $\mathbf{t}_i^n$ and textual attributes. For a basic expression image at the initial task, its class label y_i^0 equals to $\tilde{y}_i^0$. For a compound expression image at the n-th incremental task, its class label y_i^n is converted into the labels ($\tilde{y}_p^n$ and $\tilde{y}_q^n$) of its two relevant basic expressions (instead of directly using its class label).

Geometry and motion [34], [35] are widely utilized in the FER field, and both of them show promising performance for classifying video-based expressions. In this paper, we mainly consider image-based FER, where the motion information cannot be used. Some existing FER methods [3], [36]–[38] either explore an attention mechanism or divide an image into diverse local patches to adaptively capture the importance of facial regions, learning discriminative features without predefined landmarks or geometry-based features. Motivated by the success of these methods, we expect our visual modality to focus on different regions in TVM. By combining the visual-textual compatibility scores and the preliminary visual features, TVM first learns the important regions of the corresponding basic expressions (as illustrated in Fig. 2). Then, TVM utilizes ARB to highlight and aggregate the relevant visual attributes for each expression. In this way, fine-grained details about the expression can be extracted in TVM.

D. Textual Compositional Module (TCM)

To make full use of textual information at the semantic level, TCM is designed to model the dependency between textual attributes and category names of old and new expressions based on a textual compositional graph.

Compositionality, which has been widely used in zero-shot learning [39], [40] and transfer learning [41], can be viewed as the ability to decompose an observation into its primitives [42]. Inspired by this, at the n-th incremental task, we propose to construct a novel textual compositional graph based on textual attributes and category names (i.e., the compositions of textual attributes). For example, 'Happily-surprised' is a composition of 'Happy' and 'Surprised'. Hence, the graph can model the initial dependency between textual attributes and category names of old and new expressions. Based on it, a graph convolutional network (GCN) is leveraged to learn the dependency structure. In this way, the prior knowledge of the relevance between expressions is well exploited.

Node Features. Both textual attributes and category names are viewed as nodes. Each textual attribute is represented by a textual word vector of one basic expression name, while each compound expression name is represented by averaging textual word vectors of its relevant textual attributes.

Textual Compositional Graph. The textual compositional graph defined in our method is constructed by $Q = K + |\mathcal{Y}^n|$ node features that represent K textual attributes and their $|\mathcal{Y}^n|$ compositions (i.e., category names of old and new expressions) at the n-th incremental task. Technically, the textual compositional graph $\mathbf{M}$, which is a symmetric adjacency matrix, is constructed according to the relevance between expressions. If there is a connection between node i and node j (i.e., two nodes corresponding to overlapped category names (such as 'Happily-Surprised' and 'Happy') or two nodes corresponding to textual attributes), $\mathbf{M}_{ij} = 1$, otherwise $\mathbf{M}_{ij} = 0$.

Training Process. GCN [43] is an efficient type of convolutional neural network (CNN) on graphs and learns new feature

representations of nodes over M layers. We use the textual compositional graph as the adjacency matrix $\mathbf{M} \in \mathbb{R}^{Q\times Q}$. $\mathbf{M}$ and node features $\mathbf{O} \in \mathbb{R}^{Q\times Z}$ (each row is a textual word vector) are taken as the inputs of GCN. Then, GCN is utilized to propagate and aggregate the relations over neighboring nodes. The propagation rule [43] can be formulated as

$$\begin{cases} \mathbf{H}^{(m+1)} = \sigma_2(\mathbf{D}^{-1}\mathbf{M}\mathbf{H}^{(m)}\mathbf{W}^{(m)}), \\ \mathbf{H}^{(0)} = \mathbf{O}, \end{cases} \tag{6}$$

where $\mathbf{W}^{(m)}$denotes a learnable weight matrix; $\mathbf{D}$ represents a diagonal node degree matrix with normalized rows in $\mathbf{M}$; $\mathbf{H}^{(m)}$ is the hidden representations in the m-th layer. The output of GCN is denoted as $\mathbf{H}^{(M)} \in \mathbb{R}^{Q\times Z}$. We choose the last $|\mathcal{Y}^n|$ rows of $\mathbf{H}^{(M)}$ which predict the classifier weights for all the seen expression categories, and denote them as $\mathbf{K} \in \mathbb{R}^{|\mathcal{Y}^n|\times Z}$.

Finally, the output of TCM is given by $\mathbf{g}_i = \mathbf{K}h_g(\mathbf{v}_i)$, where $\mathbf{g}_i \in \mathbb{R}^{|\mathcal{Y}^n|\times 1}$ denotes the prediction of TCM and $h_g(\cdot)$ denotes a nonlinear network (an FC layer followed by a ReLU activation function) to reduce the dimension of $\mathbf{v}_i$ to Z and generate the visual-aware textual feature.

In TCM, the textual compositional graph is initialized in an unweighted way. Based on it, GCN is utilized to propagate and aggregate the relations over neighboring nodes, learning the dependency between nodes flexibly and adaptively.

E. Joint Loss

We obtain the final prediction results by combining the outputs from TVM and TCM as $\mathbf{d}_i = \frac{1}{2}\mathbf{p}_i + \frac{1}{2}\mathbf{g}_i$.

Visual-Textual Distillation Loss. Inspired by [44], a visual-textual distillation loss $\mathcal{L}_{vt}$ is introduced to calibrate TVM and TCM during incremental learning. $\mathcal{L}_{vt}$ is composed of a Jensen-Shannon divergence (JSD) and a squared l_2 distance between the predictions of two modules, which is formulated as

$$\mathcal{L}_{vt} = \frac{1}{2}(D_{KL}(\mathbf{p}_i||\mathbf{g}_i) + D_{KL}(\mathbf{g}_i||\mathbf{p}_i)) + ||\mathbf{p}_i - \mathbf{g}_i||_2^2, \tag{7}$$

where $\mathbf{p}_i$ and $\mathbf{g}_i$ are the probability outputs of all the seen classes from TVM and TCM, respectively, for the i-th input image at the n-th incremental task. $D_{KL}(.||.)$ denotes the Kullback-Leibler (KL) divergence. The first two terms are JSD, while the last term is the squared l_2 distance. JSD is a symmetric and smoothed version of the KL divergence, which measures the similarity between two probability distributions. In our task, JSD can measure how closely the probability distribution of features in one modality matches with the distribution in another modality. By minimizing JSD, the model is encouraged to produce similar distributions for both textual and visual inputs. The l_2 distance measures the difference between two sets of values. Minimizing the l_2 distance can push the model to make the features from the two modalities as close as possible. In a word, the JSD ensures that the distributions of features from the two modalities are aligned, while the l_2 distance ensures the corresponding features from the two modalities are close, providing a more direct measure of similarity. By combining the two losses, our method can give better performance than using either loss alone. This is because $\mathcal{L}_{vt}$ can deal with both the distributional alignment and the feature-wise alignment, leading to a more robust and effective learning process.

Algorithm 1 The overall training of our method at the n-th incremental task

Input: The n-th incremental subset $\mathcal{D}^n$; the old class exemplars $\mathcal{E}^n$; the total training epochs T_{max}; a set of textual attributes $\{\mathbf{s}_1, \mathbf{s}_2, \cdots \mathbf{s}_K\}$; the node features $\mathbf{O}$ in $\mathcal{D}^n$;
Output: The updated model;

Initialize the textual compositional graph $\mathbf{M}$ via the node features $\mathbf{O}$;
for each t = 1 to T_{max} **do**
  for each mini-batch in $\{\mathcal{D}^n \cup \mathcal{E}^n\}$ **do**
    for each image in a mini-batch **do**
      Calculate the visual-textual compatibility score $\alpha^r_{k,i}$ via Eq. (1);
      Obtain the k-th visual attribute $\mathbf{a}_{k,i}$ via Eq. (2);
      Calculate the attentive weight $\beta_{k,i}$ for $\mathbf{a}_{k,i}$ via Eq. (3);
      Calculate the textual-guided loss via Eq. (4);
    end for
    Obtain the textual-aware visual feature $\mathbf{v}_i$ and the output $\mathbf{p}_i$ of TVM via the cosine classifier;
    Calculate the updated classifier weights of TCM via GCN via Eq. (6);
    Obtain the visual-aware textual feature and the output $\mathbf{g}_i$ of TCM;
    Obtain the final results via a simple addition between $\mathbf{p}_i$ and $\mathbf{g}_i$;
    Calculate the visual-textual distillation loss via Eq. (7) and the classification loss via Eq. (8);
    Update the model by SGD;
  end for
end for

Classification Loss. We leverage the cross-entropy loss with self-calibration [45] as the classification loss, which can offer non-zero probabilities to old classes, reducing the forgetting of old classes during incremental learning. The classification loss is define as

$$\mathcal{L}_c = -\sum_{l=1}^{|\mathcal{Y}^n|} \mathbb{1}_{[l=y_i^n]} \log(\sigma_3(\mathbf{s}_{l,i})) - \lambda_c \log(\sum_{l'=1}^{|\mathcal{C}^n_{old}|} \sigma_3(\mathbf{s}_{l',i} + \mathbb{1}_{\mathcal{C}^n_{old}}(l'))), \tag{8}$$

where λ_c is the balancing parameter; $\mathbf{s}_i$ denotes the class scores given by the classifier in TVM or TCM; $\mathbf{s}_{l,i}$ is the l-th element of $\mathbf{s}_i$; $\sigma_3(\mathbf{s}_{l,i}) = \exp(\mathbf{s}_{l,i})/\sum_{c=1}^{|\mathcal{Y}^n|}\exp(\mathbf{s}_{c,i})$ denotes the softmax operation; $\sigma_3(\mathbf{s}_i)$ equals to $\mathbf{p}_i$ in TVM and $\mathbf{g}_i$ in TCM; when $l = y_i^n$, the function $\mathbb{1}_{[l=y_i^n]}$ equals to 1, otherwise its value is 0; when $l' \in \mathcal{C}^n_{old}$, the function $\mathbb{1}_{\mathcal{C}^n_{old}}(l')$ equals to 1, otherwise its value is -1;

Based on the above formulations, the joint loss is given as

$$\mathcal{L} = \mathcal{L}_d + \lambda_1\mathcal{L}_c + \lambda_2\mathcal{L}_t + \lambda_3\mathcal{L}_{vt}, \tag{9}$$

where $\mathcal{L}_d$ denotes a simple distillation loss [9]; λ_1, λ_2, and λ_3 are the balancing parameters.

F. Overall Training

We summarize the overall training of our method at the n-th incremental task in Algorithm 1.

IV. Experiments

In this section, we first introduce the databases in Section IV-A. Then, we present the implementation details of our method in Section IV-B. Next, we conduct ablation studies in Section IV-C and give some visualization results in Section IV-D. Finally, we compare our method with several state-of-the-art methods in Section IV-E.

A. Databases

In this paper, we evaluate our method on an in-the-lab database (i.e., CFEE [7]) and two in-the-wild databases (i.e., RAF-DB [13] and EmotioNet [12]).

CFEE first defines the compound expressions. In CFEE, a total of 230 human subjects are recruited from the university area, where different races are included. Meanwhile, CFEE also provides facial action unit (AU) coding system analysis to analyze the differences between basic expressions and compound expressions. Each AU encodes the fundamental actions of individuals or groups of muscles, where the meaningful combinations of AU can give specific facial expressions. In total, CFEE contains 7 basic expressions (with 1,610 images) and 15 compound expressions (with 3,450 images).

RAF-DB is a real-world FER database ranging in age from 0 to 70 years old including 52% female, 43% male, and 5% remain unsure. For the racial distribution, there are 77% Caucasian, 8% African-American, and 15% Asian. Meanwhile, RAF-DB contains basic expressions with a single-modal distribution and compound expressions with a bimodal distribution, consistent with the CFEE observations. In total, RAF-DB contains 7 basic expressions and 11 compound expressions. Specifically, it has 15,339 basic expression images involving 12,271 training images and 3,068 test images. It also has 3,954 compound expression images involving 3,162 training images and 792 test images.

EmotioNet is a large-scale in-the-wild database captured from the internet, where AU and AU intensity are automatically annotated by a trained classifier. Meanwhile, it also divides facial expressions into the basic and compound emotion categories defined in CFEE. We use the second track of the EmotioNet Challenge. It contains 2,478 images with 6 basic expressions and 10 compound expressions.

B. Implementation Details

Each facial image is first aligned and then resized to the size of 224×224. All the results are reported under the same settings based on PyCIL [46] (a Python toolbox for CIL).

Our model is trained using stochastic gradient descent with the initial learning rate of 0.01 at the initial task and 0.001 at the incremental tasks, where we use CosineAnnealingLR [47] as a scheduler. For each incremental task, we train the model for 40 epochs with a batch size of 32. We store 20 exemplars of old classes based on the rehearsal strategy as [9].

Following most existing FER methods [2], [3], [16], all the competing methods here adopt ResNet-18 pre-trained on the MS-Celeb-1M face recognition dataset as the backbone because of its good tradeoff between classification accuracy and model efficiency. Note that vision Transformer (ViT) [48] is not used as the backbone. This is due to the relatively large number of network parameters in ViT, which adversely affects the performance (since the number of training expression images is not sufficient during each incremental task in our experiments). In all experiments, we first train our method on basic expressions as the initial task and learn new compound expressions as the incremental tasks. We run experiments on three different compound class orders (i.e., random seed 1993 as common). We first set the classification of basic expressions as our initial task and then the compound expressions are randomly selected and learned incrementally at each incremental task according to the class orders. The number of incremental classes is set to C=3 or C=5 in incremental tasks. The parameters λ_1, λ_2, λ_3, and λ_c are empirically set to 1.0, 1.0, 0.5, and 0.1, respectively (λ_c=0 at the initial task). The number of basic expression categories K is 7 in RAF-DB and CFEE, while K is 6 in EmotioNet. The results are evaluated on both old and new classes after each incremental task. Then, we average these results as the final result for each class order. Finally, the report average±standard deviations for three different class orders as DER [28].

For RAF-DB, we employ its official training set and test set. For CFEE and EmotioNet, we follow the default evaluation protocols provided in the original papers, where we conduct a 10-fold cross-validation test on CFEE and a 5-fold cross-validation test on EmotioNet. The random state of division is 1993. The evaluation protocols in these databases are designed to evaluate the performance of FER under various conditions (such as different races, genders, poses, illuminations, and identities). By adopting the default evaluation protocols, the results obtained by our method and those in other papers can be fairly compared.

C. Ablation Studies

The results obtained by different variants of VTA-Net on RAF-DB are given in Table I. ResNet-18 with a cosine classifier and a simple distillation strategy is used as our baseline method.

Influence of TVM. We evaluate a variant of our method (denoted as Baseline+TVM), which incorporates TVM into the baseline. Moreover, we evaluate Baseine+TVM without the textual-guided loss (denote as Baseline+TVM (w.o. $\mathcal{L}_t$)). We also evaluate a simple version of our method (denoted as Baseline+Textual), which simply exploits the textual information by projecting the visual features into the textual space, as done in [49].

Compared with the Baseline, Baseline+Textual obtains higher accuracy for both C=3 and C=5, This indicates that making use of the rich textual information is beneficial to improve the performance of class-incremental FER. Baseline+TVM significantly outperforms Baseline+Textual. Notice that Baseline+Textual simply learns a visual-textual mapping,

TABLE I
ABLATION STUDIES FOR SEVERAL VARIANTS OF OUR METHOD WITH THE DIFFERENT NUMBERS OF INCREMENTAL CLASSES C=3 AND C=5 ON RAF-DB. 'AVG±STD' DENOTES THE AVERAGE ACCURACY (%) AND THE STANDARD DEVIATION OVER THE INCREMENTAL TASKS. THE BEST RESULTS ARE MARKED IN BOLD.

Methods	Avg±std	
	C=3	C=5
Baseline	$45.53_{\pm 1.49}$	$48.77_{\pm 1.67}$
Baseline+Textual	$51.17_{\pm 1.69}$	$54.82_{\pm 0.27}$
Baseline+TVM (w.o. $\mathcal{L}_t$)	$60.89_{\pm 1.22}$	$61.85_{\pm 0.42}$
Baseline+TVM ($\mathcal{L}_{co}$)	$62.59_{\pm 1.72}$	$63.96_{\pm 0.27}$
Baseline+TVM ($\mathcal{L}_{dis}$)	$61.59_{\pm 0.55}$	$63.78_{\pm 0.82}$
Baseline+TVM	$66.35_{\pm 0.67}$	$66.56_{\pm 0.44}$
Baseline+TCM	$63.32_{\pm 0.83}$	$63.86_{\pm 0.74}$
Baseline+TVM+TCM (C)	$67.37_{\pm 1.13}$	$67.45_{\pm 0.76}$
Baseline+TVM+TCM (Con)	$67.13_{\pm 1.72}$	$66.76_{\pm 0.80}$
Baseline+TVM+TCM (Att)	$65.10_{\pm 1.13}$	$66.94_{\pm 1.08}$
Baseline+TVM+TCM (w.o.$\mathcal{L}_{vt}$)	$70.51_{\pm 1.08}$	$70.35_{\pm 0.71}$
Baseline+TVM+TCM (JSD)	$70.84_{\pm 0.61}$	$71.67_{\pm 0.34}$
Baseline+TVM+TCM (l_2)	$71.03_{\pm 0.60}$	$71.72_{\pm 0.26}$
Baseline+TVM+TCM	$\mathbf{71.05}_{\pm 1.20}$	$\mathbf{72.17}_{\pm 0.38}$

which cannot effectively describe subtle visual distinctions between expressions. In contrast, Baseline+TVM can extract more discriminative fine-grained features guided by the textual attributes. Compared with Baseline+TVM (w.o. $\mathcal{L}_t$), Baseline+TVM achieves 5.46% and 4.71% improvements in terms of recognition accuracy on C=3 and C=5, respectively. The co-occurrence loss can enforce each expression to associate with the relevant visual attributes, while the distribution loss can make the same expression category give similar weights. Minimizing the two losses can improve the performance from the perspectives of the relationship among expressions and the distribution within one expression. Hence, their combination can make the model obtain better performance. The above experiments show the effectiveness of TVM and the textual-guided loss.

Influence of TCM. We evaluate a variant of our method (denoted as Baseline+TCM), which incorporates TCM into the baseline. For Baseline+TCM, we directly use the features extracted from the backbone network instead of the textual-aware visual features. Baseline+TCM obtains better accuracy than Baseline and Baseline+Textual for both C=3 and C=5. This shows the importance of TCM, which exploits the dependency between textual attributes and category names based on a textual compositional graph.

Influence of Different Fusion Strategies. We investigate three different fusion strategies from the different perspectives of fusion level (including feature level (concatenation and attention) and decision level) to investigate the different combinations of TVM and TCM. Specifically, we denote the method that only uses the outputs (i.e., $\mathbf{g}_i$) of TCM for expression classification as Baseline+TVM+TCM (C), while the method that only uses the outputs (i.e., $\mathbf{p}_i$) of TVM for identifying expressions is denoted as Baseline+TVM. Baseline+TVM+TCM (Con) denotes the method that adopts a concatenate operation between the textual-aware visual feature in TVM and the visual-aware textual feature in TCM, and then passes the concatenated feature through a nonlinear network (an FC layer followed by a ReLU activation function) for expression classification. Baseline+TVM+TCM (Att) denotes the method that applies the attention mechanism [50] between the visual and textual features in TVM and TCM. Both Baseline+TVM+TCM (Con) and Baseline+TVM+TCM (Att) perform fusion at the feature level. Our proposed method is denoted as Baseline+TVM+TCM , while our method without the visual-textual distillation loss is denoted as Baseline+TVM+TCM (w.o. $\mathcal{L}_{vt}$). The comparison results are given in Table I.

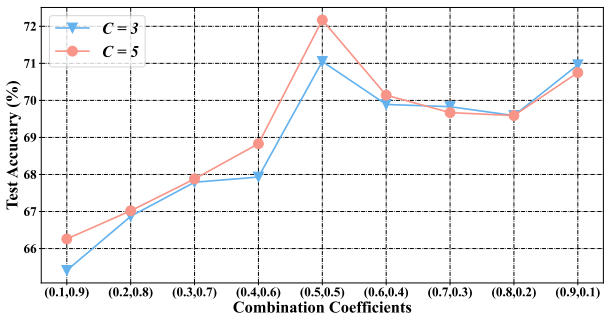

Fig. 3. **Influence of the combination coefficients** (c_1, c_2) of the outputs from TVM and TCM on RAF-DB.

Compared with Baseline+TVM, Baseline+TVM+TCM (C) and Baseline+TVM+TCM (w.o. $\mathcal{L}_{vt}$) gives 1.02% and 4.16% accuracy improvements, respectively, on C=3. More impressively, Baseline+TVM+TCM (w.o. $\mathcal{L}_{vt}$) outperforms Baseline+TVM+TCM (C) by 3.14% and 2.90% improvements in terms of accuracy for C=3 and C=5, respectively. Therefore, the predicted results fused from the two modules can give better performance. Baseline+TVM+TCM (Con) and Baseline+TVM+TCM (Att) obtain worse results than Baseline+TVM+TCM. This shows that these two types of fusion strategies are not optimal choices for our class-incremental FER task. The decision-level fusion gives better results than the feature-level fusion. Baseline+TVM+TCM achieves higher accuracy than Baseline+TVM+TCM (w.o. $\mathcal{L}_{vt}$). JSD can help the feature distributions of the two modalities become more similar since we expect that the two modalities can complement and calibrate from the perspective of distribution. Meanwhile, we also expect that the l_2 distance can make the features from the two modalities more similar. From Table I, we can see the combination of them can achieve better performance than each loss. This shows the superiority of $\mathcal{L}_{vt}$ since these two modules can learn collaboratively to calibrate each other.

Influence of the Combination Coefficients. We study the influence of the combination coefficients (c_1, c_2) of the outputs from TVM and TCM (i.e., the final output is predicted by $c_1\mathbf{p}_i$+$c_2\mathbf{g}_i$) on RAF-DB. The results are given in Fig. 3. From Fig. 3, our method obtains the best results when the combination coefficients are set to (0.5, 0.5). Therefore, both TVM and TCM are equally important to ensure the good accuracy of our method.

Influence of the Number of Incremental Classes. We also

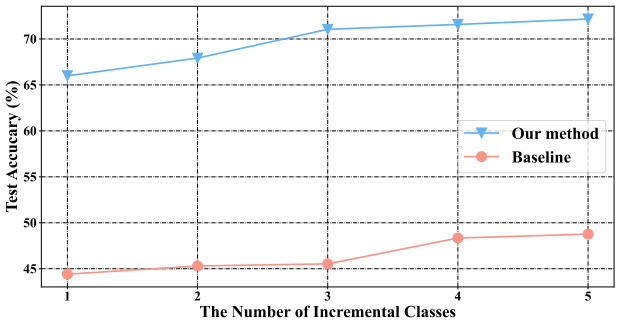

Fig. 4. **Influence of the number** of incremental classes on RAF-DB.

TABLE II
PERFORMANCE (THE AVERAGE ACCURACY (%)) COMPARISONS BETWEEN THE PROPOSED METHOD AND SEVERAL STATE-OF-THE-ART FER METHODS ON TWO POPULAR DATABASES (MMI AND OULU-CASIA.

Methods	MMI	Oulu-CASIA
DDL [53]	83.67	88.26
FDRL [16]	85.23	88.26
ADDL [54]	**86.13**	**89.44**
CBLSTM [55]	83.67	-
Baseline	75.58	85.49
VTA-Net (Ours)	85.79	88.54

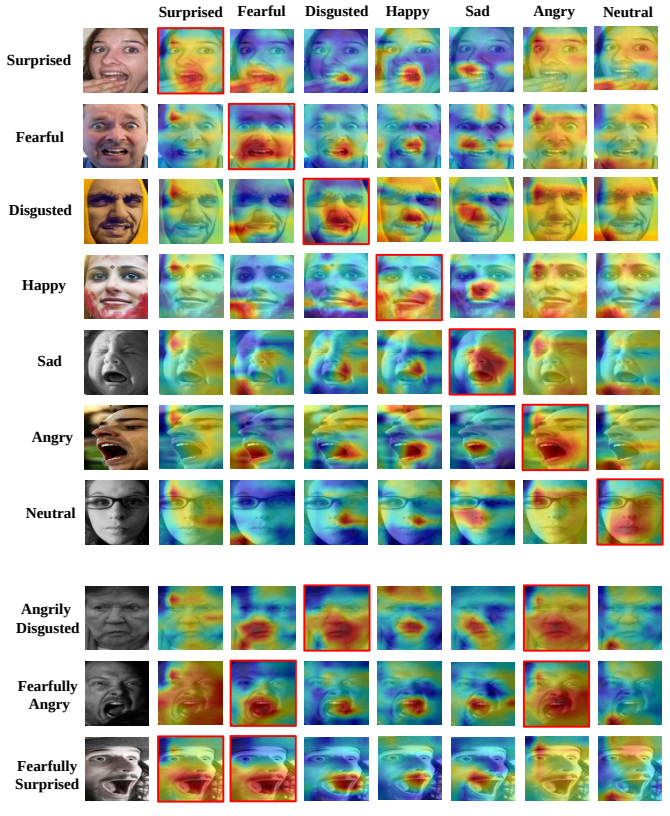

Fig. 5. **Visualization of the attention maps** of 7 visual attributes corresponding to 7 basic expression and compound expression images on RAF-DB. The image in the red box in each row denotes the attention map of the relevant visual attribute associated with one basic expression.

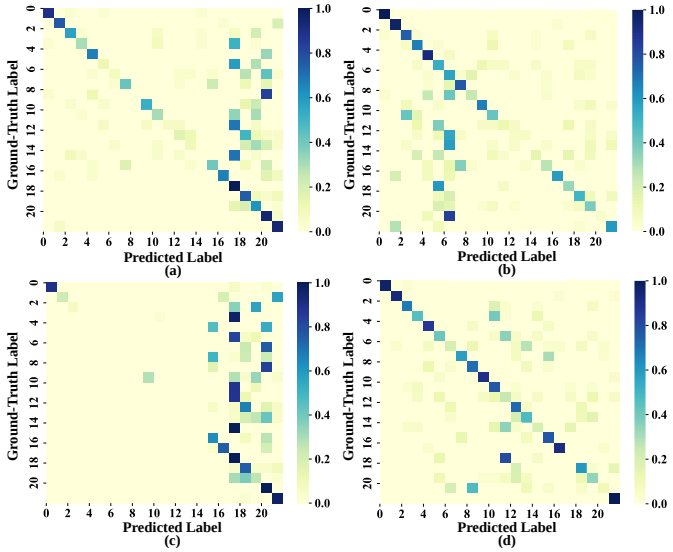

Fig. 6. **Visualization of confusion matrices** obtained by (a) Baseline, (b) FOSTER, (c) SCN, and (d) our VTA-Net at the last incremental task on CFEE (C=5).

visualize the influence of the number of incremental classes (denoted as C) on RAF-DB. As shown in Fig. 4, the different values of C have great influence on the final performance. When the number of incremental classes becomes large, the influence caused by forgetting is alleviated. Meanwhile, our method greatly outperforms than Baseline. This shows the effectiveness of our method. Therefore, we choose two specific values of C to evaluate the performance of CIL FER tasks when the number of incremental classes is small (C=3) and large (C=5) in our experiments.

Performance on the Traditional FER Task. To identify the relationship between our task and the traditional FER task (i.e., the classification of basic expressions), we train and test our model on the basic expression images (such a task is the same as the initial task $\mathcal{D}_0$) and compare our method with several state-of-the-art FER methods on two popular basic expression databases (MMI [51] and Oulu-CASIA [52]). The results are given in Table II. We can see that our proposed method, although designed for class-incremental FER rather than traditional FER, outperforms some state-of-the-art FER methods (such as DDL, FDRL, and CBLSTM). This shows the feasibility of our model on the traditional FER task. Note that our method achieves worse performance than the recently proposed method ADDL. ADDL leverages complicated network design to alleviate the influence of multiple disturbing factors for FER. In contrast, our method aims to address the stability-plasticity dilemma of class-incremental FER (note that our method is not designed for the traditional FER task). Our method explores the relationship between basic and compound expressions by attribute learning. Moreover, we model the dependency between textual attributes and category names of old and new expressions in TCM.

TABLE III
THE LABEL INFORMATION OF DIFFERENT EXPRESSIONS IN CFEE.

Label	Expressions	Label	Expressions
0	Neutral	11	Appealled
1	Happy	12	Fearfully Angry
2	Sad	13	Sadly Fearful
3	Happy	14	Hatred
4	Angry	15	Sadly Surprised
5	Surprised	16	Happily Surprised
6	Disgusted	17	Angrily Disgusted
7	Angrily Surprised	18	Fearfully Disgusted
8	Awed	19	Sadly Disgusted
9	Disgustedly Surprised	20	Fearfully Surprised
10	Sadly Angry	21	Happily Disgusted

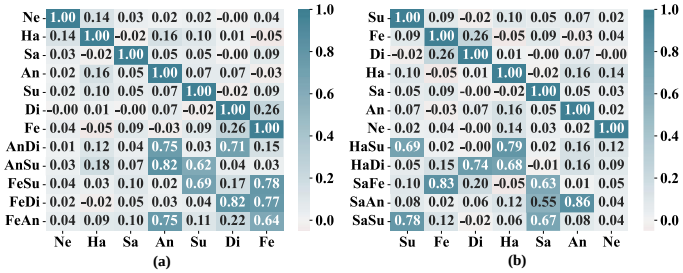

Fig. 7. **Visualization of correlations** between textual word vectors corresponding to some category names on (a) CFEE and (b) RAF-DB.

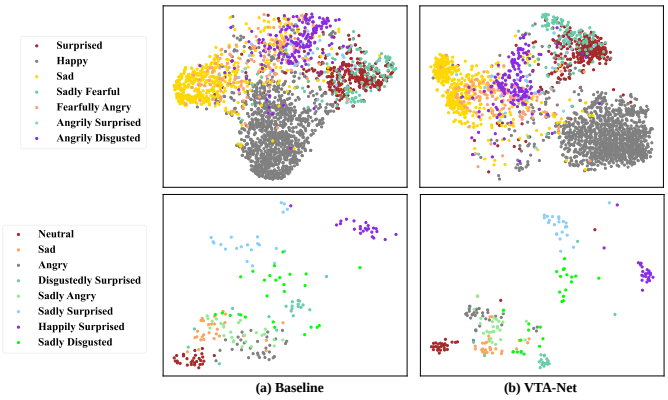

Fig. 8. **Visualization of expression features** obtained by (a) Baseline and (b) VTA-Net. We randomly select several basic and compound expressions from different incremental tasks. The upper row and the lower row of panels show the feature distributions on RAF-DB and CFEE, respectively.

D. Visualization

Visualization of Attention Maps. Fig. 5 visualizes the attention maps of 7 visual attributes corresponding to 7 basic expression images and some compound expression images (from K=7 classes) on RAF-DB. Our model can effectively learn visual attributes for different basic expression images. Each visual attribute can pay attention to the most related regions which contain rich and fine-grained information for identifying the corresponding expression. Based on the learned visual attributes from basic expressions, compound expressions can be represented in terms of visual attributes. Thus, we can represent an old or new expression (a basic or a compound expression) in a simple and unified way, serving as the basis of our method for the class-incremental FER task.

Moreover, we can see that each compound expression associates with its relevant visual attributes which pay attention to the most related regions for this expression. For example, as shown in Fig. 5, the visual attribute of the expression 'Angrily Surprised' (corresponding to the 'Angry' and 'Surprised' expressions) can focus on the informative region (e.g., the eye and the mouth) and suppress other uninformative regions. Although some expressions could share very similar facial actions, the most related expressions should be emphasized with co-occurrence loss. The same phenomenon also can be shown in the compound expression 'Angrily Disgusted', as a combination of the basic expressions 'Angry' and 'Disgusted', the relevant visual attributes related to 'Angrily Disgusted' are more highlighted.

Visualization of Confusion Matrices. We visualize the confusion matrices obtained by several methods at the last incremental task on CFEE (C=5), as shown in Fig. 6. The label information of different expressions in CFEE is given in Table III. The Baseline method is prone to focus on learning new classes even with a distillation loss, and it forgets the learned knowledge severely. The stability (remembering old classes) of FOSTER is better than other methods while its plasticity (learning new classes) is not good. Moreover, these methods do not fully consider both the rich textual information (which characterizes the intrinsic dependency between expressions) and the visual information. Due to the bias towards new classes, SCN is likely to relabel the samples to new classes, leading to the forgetting of old classes. On the contrary, our VTA-Net effectively mitigates the stability-plasticity dilemma by considering both visual and textual modalities.

Visualization of Correlations. As shown in Fig. 7, we visualize the correlations (we use the cosine similarity) between textual word vectors corresponding to some category names on CFEE and RAF-DB, showing the close relationship between basic and compound expressions. The correlations between a compound expression and its two relevant basic expressions are higher than those between a compound expression and irrelevant basic expressions. Therefore, the textual modality can provide rich prior information and offer guidance to visual attribute learning in TVM.

Visualization of Expression Features. In Fig. 8, we visualize the expression features of test data by t-SNE on RAF-DB (C=3) and CFEE (C=3). Compared with the features obtained by Baseline, the features obtained by our VTA-Net method are more discriminative on different classes. Our method can effectively learn features with better inter-class separability and intra-class compactness for identifying different expressions.

Test Accuracy on CFEE and EmotioNet. We show the test accuracy vs. the number of classes by different methods on CFEE and EmotioNet in Fig. 9. Our method performs better than the other competing methods at each incremental task and obtains better results with the accuracy improvements of 1.01%/0.80% and 1.13%/0.91% on CFEE (C=3/C=5) and EmotioNet (C=3/C=5), respectively. This demonstrates the

TABLE IV

PERFORMANCE COMPARISONS (THE AVERAGE ACCURACY (%) AND THE STANDARD DEVIATION OVER THE INCREMENTAL TASKS) BETWEEN OUR PROPOSED METHOD AND SEVERAL STATE-OF-THE-ART METHODS WITH THE DIFFERENT NUMBERS OF INCREMENTAL CLASSES C=3 AND C=5 ON CFEE, RAF-DB, AND EMOTIONET. THE BEST RESULTS ARE MARKED IN **BOLD**.

Methods	CFEE		RAF-DB		EmotioNet	
	C=3	C=5	C=3	C=5	C=3	C=5
iCaRL [9]	$67.39_{\pm 1.25}$	$68.27_{\pm 1.64}$	$63.33_{\pm 0.79}$	$63.96_{\pm 0.22}$	$59.48_{\pm 0.44}$	$61.40_{\pm 0.77}$
PODNet [10]	$63.82_{\pm 1.85}$	$66.31_{\pm 1.55}$	$58.36_{\pm 1.20}$	$61.02_{\pm 0.92}$	$56.11_{\pm 0.57}$	$59.73_{\pm 1.32}$
COIL [56]	$56.35_{\pm 1.26}$	$58.25_{\pm 0.47}$	$47.73_{\pm 2.65}$	$48.34_{\pm 1.13}$	$52.85_{\pm 2.21}$	$56.38_{\pm 1.62}$
AFC [27]	$65.54_{\pm 1.75}$	$66.81_{\pm 1.49}$	$68.59_{\pm 1.11}$	$66.96_{\pm 0.47}$	$59.79_{\pm 1.50}$	$61.75_{\pm 0.91}$
FOSTER [11]	$62.12_{\pm 1.60}$	$62.39_{\pm 1.17}$	$69.11_{\pm 0.58}$	$70.04_{\pm 0.27}$	$60.90_{\pm 2.06}$	$62.85_{\pm 0.40}$
MEMO [29]	$66.01_{\pm 2.28}$	$67.95_{\pm 1.97}$	$63.22_{\pm 1.47}$	$62.49_{\pm 0.72}$	$57.87_{\pm 1.85}$	$58.73_{\pm 0.93}$
SCN [2]	$46.62_{\pm 0.23}$	$53.73_{\pm 1.29}$	$43.34_{\pm 2.53}$	$40.34_{\pm 1.45}$	$50.21_{\pm 1.84}$	$55.40_{\pm 1.43}$
Baseline	$60.30_{\pm 1.41}$	$61.30_{\pm 0.95}$	$45.53_{\pm 1.49}$	$48.77_{\pm 1.67}$	$55.48_{\pm 0.75}$	$57.23_{\pm 1.08}$
VTA-Net (Ours)	$\mathbf{68.40}_{\pm 1.46}$	$\mathbf{69.07}_{\pm 1.59}$	$\mathbf{71.05}_{\pm 1.20}$	$\mathbf{72.17}_{\pm 0.38}$	$\mathbf{62.03}_{\pm 1.50}$	$\mathbf{63.76}_{\pm 0.76}$

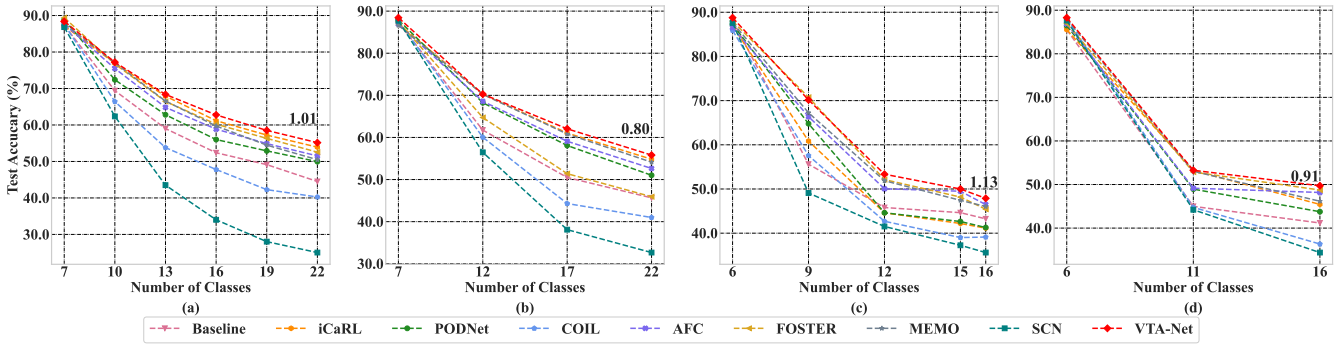

Fig. 9. **Test accuracy (%) vs. the number of classes** obtained by different methods for (a) C=3 and (b) C=5 on CFEE, while (c) C=3 and (d) C=5 on EmotioNet.

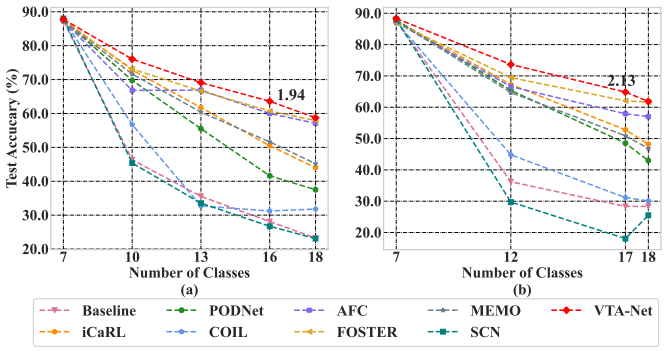

Fig. 10. **Test accuracy vs. the number of classes** obtained by different methods for (a) C=3 and (b) C=5 on RAF-DB.

effectiveness of our method.

FOSTER achieves slightly worse performance than our method on EmotioNet, but its performance greatly drops on CFEE. This is mainly due to the different characteristics of the databases. CFEE involves more subtle differences between the compound expression categories (such as hatred and appalled) than EmotioNet. Note that FOSTER leverages the dynamic structures and requires a large number of samples to accurately identify these subtle distinctions, leading to insufficient training in CFEE. In contrast, our method can achieve the best results on both CFEE and EmotioNet among all the competing methods.

E. Comparison with State-of-the-Art Methods

Table IV shows the performance comparisons between our proposed VTA-Net method and several state-of-the-art CIL methods and an FER-based method on three facial expression databases. We also give the test accuracy vs. the number of classes by different methods on RAF-DB in Fig. 10. Note that some methods [30], [31] also explore both visual and textual information for image recognition. However, they focus on few-shot class-incremental learning, which intrinsically differs from our settings. Therefore, these methods are not evaluated in our experiments.

Our VTA-Net method achieves the best average accuracy among all the competing methods while its standard deviation is comparable to those of other methods on all the databases. Specifically, VTA-Net obtains the highest accuracy of 68.40% (69.07%) on the in-the-lab CFEE database, 71.05% (72.17%), and 62.03% (63.76%) on the in-the-wild RAF-DB and EmotioNet databases, respectively, when the number of incremental classes is C=3 (5). iCaRL, PODNet, and AFC leverage different distillation strategies to alleviate the stability-plasticity dilemma, while FOSTER and MEMO investigate the dynamic structures to balance the learning of

new classes and the retaining of old classes. However, these methods do not fully consider both the rich textual information and the visual information. For the FER-based method, SCN performs well on basic expressions but it may relabel the samples to new classes due to the bias towards new classes. On the contrary, based on the relationship between basic and compound expressions, our proposed VTANet considers both visual and textual modalities for class-incremental FER. The above results show the effectiveness of our proposed method,

The Baseline method can adapt to new classes but it cannot remember the learned knowledge well. FOSTER gives good performance on old expressions but obtains low accuracy on new expressions. COIL develops a semantic mapping to transfer old classifiers to new classes with optimal transport and transfer new classifiers to old classes symmetrically. However, the semantic mapping in COIL does not fit FER CIL very well. iCaRL, PODNet, and AFC explore different distillation strategies, where iCaRL leverages a distillation loss via old exemplars while PODNet and AFC utilize the distillation loss to prevent the model from forgetting important information of old classes. For the FER method, SCN achieves good accuracy on the classification of basic expressions but fails to identify old expressions in the incremental tasks. Although the distillation loss and exemplars from old classes are used to train SCN, it still suffers from catastrophic forgetting since the bias towards new classes makes the model easily relabel new classes.

Existing methods ignore the importance of textual information in the class-incremental FER. In contrast, our VTA-Net method explores the effective learning way of textual modality and FER class-incremental tasks. In such a way, VTA-Net can model the intrinsic relationship between basic and compound expressions based on visual-textual attribute learning through TVM and TCM. Meanwhile, VTA-Net utilizes an effective Visual-Textual Distillation Loss to complement and calibrate the two branches. Among all the competing methods, VTA-Net can effectively balance the trade-off between old and new expressions and achieve the best performance in terms of average accuracy on three databases.

V. Conclusion and Future Work

In this paper, we develop a novel VTA-Net for class-incremental FER by taking advantage of both visual and textual modalities. To fit to the compound FER task, we take well-studied and easily-accessible basic expressions as initial classes while treating new compound expressions as incremental classes. By elaborately designing TVM and TCM, old/new expression images and their expression names can be represented in terms of visual and textual attributes. In this way, the textual modality provides auxiliary supervision to identify fine-grained expressions while the representations of visual attributes are continuously enhanced during each incremental task, alleviating the stability-plasticity dilemma. Experiments show the effectiveness of our method in comparison with several state-of-the-art methods.

Our method mainly focuses on addressing the stability-plasticity dilemma of class-incremental FER. Hence, we neither utilize data augmentations to balance classes nor employ a LOSO or LOVO decomposition to avoid representation bias. However, the unbalanced class distribution or representation bias can also affect the performance of class-incremental FER. These problems merit further research in future work. Apart from this, in real-world applications, some new compound expressions may not be clear combinations of basic expressions. A more flexible way to deal with such a case also deserves our future research. Our exploration suggests that we can leverage a similarity function to explore the top similar basic expressions of a new expression via the word embedding in the textual space, and then leverage similar basic expressions to guide visual-textual attribute learning.

References

[1] Y. Yan, Y. Huang, S. Chen, C. Shen, and H. Wang, "Joint deep learning of facial expression synthesis and recognition," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2792–2807, 2020.

[2] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6897–6906.

[3] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7660–7669.

[4] D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang, "Face2Exp: Combating data biases for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20 291–20 300.

[5] 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 Trans. Multimedia*, vol. 24, pp. 3327–3339, 2022.

[6] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[7] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Natl. Acad. Sci.*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[8] S. Grossberg, "Adaptive resonance theory: How a brain learns to consciously attend, learn, and recognize a changing world," *Neural Netw.*, vol. 37, pp. 1–47, 2013.

[9] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2001–2010.

[10] A. Douillard, M. Cord, C. Ollion, T. Robert, and E. Valle, "PODNet: Pooled outputs distillation for small-tasks incremental learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 86–102.

[11] F.-Y. Wang, D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "FOSTER: Feature boosting and compression for class-incremental learning," *arXiv preprint arXiv:2204.04662*, 2022.

[12] C. Fabian 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5562–5570.

[13] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2852–2861.

[14] J. Guo, S. Zhou, J. Wu, J. Wan, X. Zhu, Z. Lei, and S. Z. Li, "Multi-modality network with visual and geometrical information for micro emotion recognition," in *Proc. IEEE Int. Conf. Auto. Face Gesture Recognit.*, 2017, pp. 814–819.

[15] Z. Zhang, M. Yi, J. Xu, R. Zhang, and J. Shen, "Two-stage recognition and beyond for compound facial emotion recognition," in *Proc. IEEE Int. Conf. Auto. Face Gesture Recognit.*, 2020, pp. 900–904.

[16] X. Zou, Y. Yan, J.-H. Xue, S. Chen, and H. Wang, "When facial expression recognition meets few-shot learning: A joint and alternate learning framework," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 5367–5375.

[17] R. Wu, G. Zhang, S. Lu, and T. Chen, "Cascade EF-GAN: Progressive facial expression editing with local focuses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5021–5030.

[18] W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, and N. Sebe, "Every smile is unique: Landmark-guided diverse smile generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7083–7092.
[19] N. Otberdout, C. Ferrari, M. Daoudi, S. Berretti, and A. Del Bimbo, "Sparse to dense dynamic 3d facial expression generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20 385–20 394.
[20] T. Ma, B. Li, Q. He, J. Dong, and T. Tan, "GaFET: Learning geometry-aware facial expression translation from in-the-wild images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 7115–7125.
[21] N. Otberdout, C. Ferrari, M. Daoudi, S. Berretti, and A. Del Bimbo, "Generating multiple 4d expression transitions by learning face landmark trajectories," *IEEE Trans. Affect. Comput.*, 2023.
[22] J. Zhu, B. Luo, S. Zhao, S. Ying, X. Zhao, and Y. Gao, "IExpressNet: Facial expression recognition with incremental classes," in *Proc. ACM Int. Conf. Multimed.*, 2020, pp. 2899–2908.
[23] K. Chen, X. Yang, C. Fan, W. Zhang, and Y. Ding, "Semantic-rich facial emotional expression recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1906–1916, 2022.
[24] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proc. Natl. Acad. Sci.*, vol. 114, no. 13, pp. 3521–3526, 2017.
[25] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 532–547.
[26] W. Chen, Y. Liu, N. Pu, W. Wang, L. Liu, and M. S. Lew, "Feature estimations based correlation distillation for incremental image retrieval," *IEEE Trans. Multimedia*, vol. 24, pp. 1844–1856, 2022.
[27] M. Kang, J. Park, and B. Han, "Class-incremental learning by knowledge distillation with adaptive feature consolidation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16 071–16 080.
[28] S. Yan, J. Xie, and X. He, "DER: Dynamically expandable representation for class incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3014–3023.
[29] D.-W. Zhou, Q.-W. Wang, H.-J. Ye, and D.-C. Zhan, "A model or 603 exemplars: Towards memory-efficient class-incremental learning," *arXiv preprint arXiv:2205.13218*, 2022.
[30] A. Cheraghian, S. Rahman, P. Fang, S. K. Roy, L. Petersson, and M. Harandi, "Semantic-aware knowledge distillation for few-shot class-incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2534–2543.
[31] J. Li, Y. Bai, Y. Lou, X. Linghu, J. He, S. Xu, and T. Bai, "Memory-based label-text tuning for few-shot class-incremental learning," *arXiv preprint arXiv:2207.01036*, 2022.
[32] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.
[33] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 831–839.
[34] Z. Wu and J. Cui, "LA-Net: Landmark-aware learning for reliable facial expression recognition under label noise," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 20 698–20 707.
[35] D. Poux, B. Allaert, N. Ihaddadene, I. M. Bilasco, C. Djeraba, and M. Bennamoun, "Dynamic facial expression recognition under partial occlusion with optical flow reconstruction," *IEEE Trans. Image Process*, vol. 31, pp. 446–457, 2021.
[36] F. Xue, Q. Wang, and G. Guo, "TransFER: Learning relation-aware facial expression representations with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3601–3610.
[37] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process*, vol. 29, pp. 4057–4069, 2020.
[38] S. Xie, H. Hu, and Y. Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," *Pattern Recogn.*, vol. 92, pp. 177–191, 2019.
[39] S. Purushwalkam, M. Nickel, A. Gupta, and M. Ranzato, "Task-driven modular networks for zero-shot compositional learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3593–3602.
[40] X. Li, X. Yang, K. Wei, C. Deng, and M. Yang, "Siamese contrastive embedding network for compositional zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9326–9335.
[41] N. Patricia and B. Caputo, "Learning to learn, from transfer learning to domain adaptation: A unifying perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1442–1449.
[42] I. Biederman, "Recognition-by-components: A theory of human image understanding." *Psychological Review*, vol. 94, no. 2, pp. 115–147, 1987.
[43] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
[44] S. Chen, Z. Hong, G.-S. Xie, W. Yang, Q. Peng, K. Wang, J. Zhao, and X. You, "MSDN: Mutually semantic distillation network for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7612–7621.
[45] D. Huynh and E. Elhamifar, "Fine-grained generalized zero-shot learning via dense attribute-based attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4483–4493.
[46] D.-W. Zhou, F.-Y. Wang, H.-J. Ye, and D.-C. Zhan, "PyCIL: A python toolbox for class-incremental learning," *arXiv preprint arXiv:2112.12533*, 2021.
[47] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.
[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[49] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang, "DeViSE: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1–9.
[50] Z. Ji, X. Liu, Y. Pang, and X. Li, "SGAP-Net: Semantic-guided attentive prototypes network for few-shot human-object interaction recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11 085–11 092.
[51] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," *IEEE Int. Conf. Multimedia Expo.*, vol. 2005, pp. 317 – 321, 2005.
[52] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikainen, "Facial expression recognition from near-infrared videos," *Image Vision Comput.*, vol. 29, no. 9, pp. 607 – 619, 2011.
[53] D. Ruan, Y. Yan, S. Chen, J. H. Xue, and H. Wang, "Deep disturbance-disentangled learning for facial expression recognition," in *Proc. ACM Int. Conf. Multimed.*, 2020, pp. 2833–2841.
[54] D. Ruan, R. Mo, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, "Adaptive deep disturbance-disentangled learning for facial expression recognition," *Int. J. Comput. Vision*, vol. 130, no. 2, pp. 455–477, 2022.
[55] L. Chen, M. Li, M. Wu, W. Pedrycz, and K. Hirota, "Convolutional features-based broad learning with LSTM for multidimensional facial emotion recognition in human–robot interaction," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 54, no. 1, pp. 64–75, 2024.
[56] D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Co-transport for class-incremental learning," in *Proc. ACM Int. Conf. Multimed.*, 2021, pp. 1645–1654.

Yuanling Lv is currently pursuing the master's degree with the School of informatics, Xiamen University, China. Her research interests include deep learning and facial expression recognition.

Guangyu Huang is currently pursuing the master's degree with the School of informatics, Xiamen University, China. Her research interests include deep learning and computer vision.

Yan Yan (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Tsinghua University, China, in 2009. He worked as a Research Engineer with the Nokia Japan Research and Development Center from 2009 to 2010. He worked as a Project Leader with the Panasonic Singapore Laboratory in 2011. He is currently a Full Professor with the School of Informatics, Xiamen University, China. He has published around 100 papers in the international journals and conferences, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *IJCV*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, CVPR, ICCV, ECCV, AAAI, and ACM MM. His research interests include computer vision and pattern recognition.

Jing-Hao Xue (Senior Member, IEEE) received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998, and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is currently a Professor with the Department of Statistical Science, University College London. His research interests include statistical pattern recognition, machine learning, and computer vision. He received the Best Associate Editor Award of 2021 from the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the Outstanding Associate Editor Award of 2022 from the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Si Chen (Member, IEEE) received the Ph.D. degree from the School of Informatics, Xiamen University, China, in 2014. She is currently a Professor with the School of Computer and Information Engineering, Xiamen University of Technology, China. In recent years, she has published more than 40 papers in international journals and conferences, including IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, Pattern Recognition, Knowledge-Based Systems, CVPR, and ACM MM. Her research interests include computer vision, machine learning, and data mining.

Hanzi Wang (Senior Member, IEEE) is currently a Distinguished Professor of Minjiang Scholars in Fujian province and a Founding Director of the Center for Pattern Analysis and Machine Intelligence (CPAMI) at Xiamen University in China. He received his Ph.D. degree in Computer Vision from Monash University. His research interests are concentrated on computer vision and pattern recognition including visual tracking, robust statistics, object detection, video segmentation, model fitting, optical flow calculation, 3D structure from motion, image segmentation and related fields.