Exploiting Emotional Dependencies with Graph Convolutional Networks for Facial Expression Recognition

Panagiotis Antoniadis, Panagiotis Paraskevas Filntisis, Petros Maragos
School of E.C.E., National Technical University of Athens, Greece

Abstract— **Over the past few years, deep learning methods have shown remarkable results in many face-related tasks including automatic facial expression recognition (FER) in-the-wild. Meanwhile, numerous models describing the human emotional states have been proposed by the psychology community. However, we have no clear evidence as to which representation is more appropriate and the majority of FER systems use either the categorical or the dimensional model of affect. Inspired by recent work in multi-label classification, this paper proposes a novel multi-task learning (MTL) framework that exploits the dependencies between these two models using a Graph Convolutional Network (GCN) to recognize facial expressions in-the-wild. Specifically, a shared feature representation is learned for both discrete and continuous recognition in a MTL setting. Moreover, the facial expression classifiers and the valence-arousal regressors are learned through a GCN that explicitly captures the dependencies between them. To evaluate the performance of our method under real-world conditions we perform extensive experiments on the AffectNet and Aff-Wild2 datasets. The results of our experiments show that our method is capable of improving the performance across different datasets and backbone architectures. Finally, we also surpass the previous state-of-the-art methods on the categorical model of AffectNet.**

I. INTRODUCTION

Facial expressions are one of the most powerful nonverbal ways for human beings to convey their emotional state [11]. Facial expression recognition (FER) has been a topic of study for decades due to its potential applications in various fields including human-computer interaction, digital entertainment, advertisement, health care and intelligent robot systems [10], [3], [43], [46], [20]. However, recognizing facial expressions in the wild is still very challenging due to variations, occlusions and the ambiguity of human emotion [37], [55].

While the cultural and ethnic background of a person can affect his expressive style, Ekman indicated that humans perceive certain basic emotions in the same way regardless of their culture [18], [17]. These six universal facial expressions (happiness, sadness, surprise, fear, disgust and anger) constitute the categorical model. Contempt was subsequently added as one of the basic emotions [42]. Due to its direct and intuitive definition of facial expressions, the categorical model is used in the majority of FER algorithms ([4], [1], [28], [26], etc) and large-scale databases (AffectNet [44], RAF-DB [36], SFEW [13], FER-2013 [22], EmotionNet [19], etc). However, the subjectivity and ambiguity of restricting human emotion to discrete categories result in large intra-class variations and small inter-class differences.

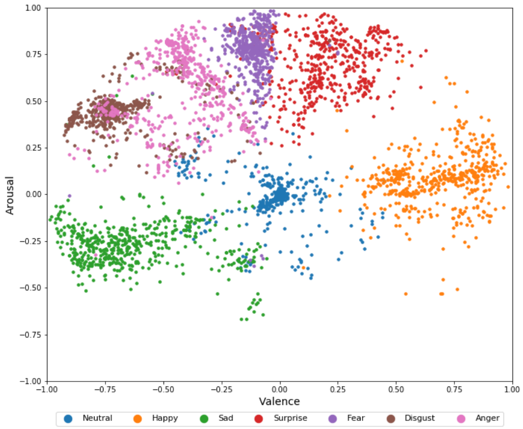


Fig. 1: Distribution of the basic expressions in the VA space using the validation set of AffectNet that illustrates the emotional dependencies between the categorical and the dimensional model. The basic emotions are located around the neutral emotion that appears when valence and arousal are close to zero.

Recently, the dimensional model proposed by Russell [51] has gained a lot of attention where emotion is described using a set of 2 latent dimensions that are valence (how pleasant or unpleasant a feeling is) and arousal (how likely is the person to take action under the emotional state). Another dimension called dominance is used sometimes to know whether the person is controlling the situation or not. Since a continuous representation can distinguish between subtly different displays of affect and encode small changes in the intensity, some recent algorithms [47], [7], [31] and databases (Aff-Wild [60] and AFEW-VA [32]) have utilized the dimensional model for uncontrolled FER. Even so, predicting a 2-dimensional continuous value instead of a category increases a lot the task complexity and lacks intuitiveness.

Altogether, the categorical and the dimensional model have their respective benefits and drawbacks [28]. Therefore, recent studies try to leverage both representations, along with Action Units (AU) detection, through multi-task learning (MTL) [7], [31], [58]. However, the strong dependence between the categorical and the dimensional model is not fully exploited when they only share a common feature representation. Fig. 1 illustrates this relation using the validation set of AffectNet.

In multi-label classification, there has been a lot of re-

P.P. Filntisis and P. Maragos are also affiliated with Athena Research Center. This work is partially supported by the H2020 EASIER Project under GA n° 101016982.

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

Proc. 16th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India (Virtual Event), Dec. 2021

search on how to properly capture and explore the correlation between labels [38], [35]. In [9], Chen proposed ML-GCN model for multi-label image recognition that explicitly learns the labels correlation by generating the object classifiers via a Graph Convolutional Network (GCN) [30]. Inspired by this architecture, we propose Emotion-GCN a novel GCN based MTL framework for FER in the wild. The main idea in this paper is to to generate dependent expression classifiers and valence-arousal (VA) regressors though a GCN based mapping function instead of learning them as separate parameter vectors. The generated vectors are then applied to an extracted image representation to enable end-to-end training. Hence, the dependence between the categorical and the dimensional emotion models is explicitly captured through both a shared feature representation and the dependent classifiers and regressors. Experiments on the AffectNet and Aff-Wild2 datasets indicate that Emotion-GCN increases the performance across different datasets and backbone networks managing to achieve state-of-the-art results on the categorical model of AffectNet.

The rest of this paper is organized as follows. Section II presents the recent work on FER systems and MTL. Section III describes the deep learning method that this paper proposes. Section IV discusses the experimental results, providing a clear view of the accuracy improvements introduced by our method. This section also includes a description of the data used during the experiments. Finally, Section V summarizes the key aspects of our work and concludes the paper.

II. RELATED WORK

Image-based FER has been extensively studied for many years. Typically, a FER system consists of three stages: face detection, feature extraction and classification. Traditional approaches tend to conduct FER by using handcrafted features, such as Local Binary Patterns [53], Gabor wavelets [40], [45] and Histogram of Oriented Gradients [5]. While there is a lot of intuition behind these features and their performance on several lab-controlled databases is impressive, they lack generalizability and sufficient learning capacity [37].

Later, many in-the-wild facial expression databases were developed that enabled the research of FER in more challenging environments. Deep Convolutional Neural Networks (CNNs) have achieved promising recognition performance by learning powerful high-level features [61], [15], [59], [29], [24], [14]. In [39] and [57] region-based attention networks were designed for pose and occlusion aware FER, where the regions are either cropped from landmark points or fixed positions. Facial Motion Prior Networks were proposed in [8] that generate a facial mask to focus on facial muscle moving regions. In [25] deep Siamese neural networks equipped with a supervised loss function were used to reduce the high intra-class variation of the task. In [21] deep and handcrafted features were combined and a local learning framework was used at test time based on SVMs. Recently, Self-Cure network [56] was proposed that suppresses the uncertainties caused by ambiguous expressions and the suggestiveness of the annotators.

MTL was first explored in [6] based on the idea that learning complementary tasks in parallel while using a shared representation improves the generalization performance. Since then, MTL has been used in many areas of computer vision, such as classification and detection [52] or geometry and regression tasks [16]. In the facial analysis domain, a multi-purpose algorithm for seven face-related tasks was proposed in [50]. FATAUVA-Net [7] performs sequential facial attribute recognition, AU detection, and VA estimation on videos. In [31] a holistic framework was proposed that jointly learns three facial behavior tasks (recognition of basic emotions, VA estimation and AUs detection) and two simple strategies were used for coupling the tasks during training. Closer to our work, in [58] a two-level attention with a two-stage MTL (2Att-2Mt) framework was proposed for facial emotion estimation on static images. However, the work was focused mainly on the estimation of VA and the dependencies between the emotion representations was not further explored.

Apart from MTL, research on the dependencies between the categorical and the dimensional emotion representations is limited. Recently, in CAKE [28] the link between the two representations was explored and a 3-dimensional representation of emotion learned in a multi-domain fashion was proposed.

III. OUR PROPOSED METHOD

In this section the architecture of the proposed network, the dependent classifiers and regressors and the MTL setting for recognition are introduced. The overall architecture of the proposed Emotion-GCN is shown in Fig. 2.

A. Feature extraction

Given an input image $\mathbf{I}$ of size 227×227 pixels, we obtain $1024 \times 7 \times 7$ feature maps using a Dense Convolutional Network (DenseNet) [27]. Each layer in DenseNet obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers to ensure maximum information flow between layers. Then, a global max-pooling function is applied to get the feature vector $\mathbf{x}$ of the image:

$$\mathbf{x} = f_{gmp}(f_{cnn}(\mathbf{I})) \in \mathbb{R}^D \tag{1}$$

where f_{cnn} corresponds to the convolution layers of the DenseNet, f_{gmp} to the the global max-pooling function and $D = 1024$.

B. Dependent classifiers and regressors

Given the feature vector $\mathbf{x}$ of the image, we want to learn one classifier for each facial expression (classification task) and one regressor for each dimension in the VA space (regression task). Each classifier acts as a weight vector that is multiplied with the feature $\mathbf{x}$ and then passes though a softmax activation function assigning a probability score to its emotion category. In parallel, each regressor follows

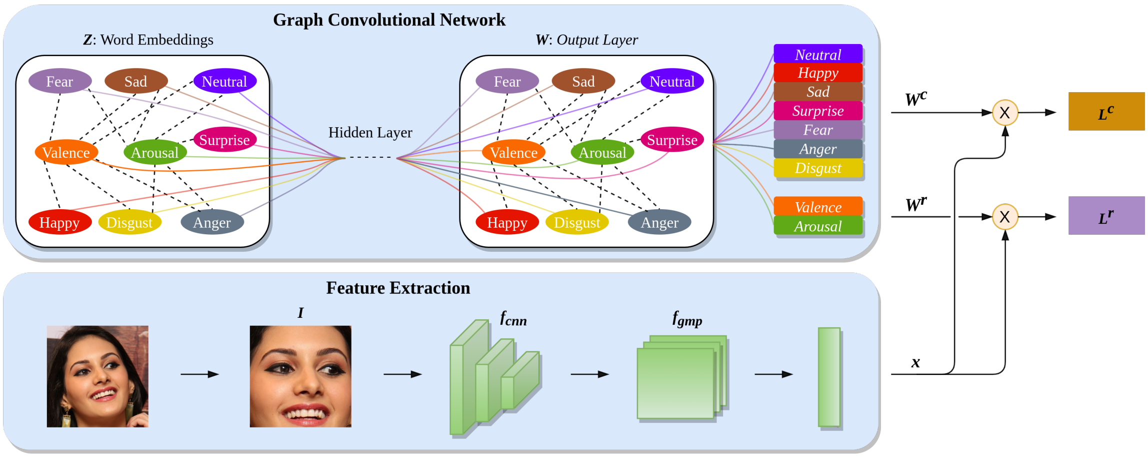


Fig. 2: Overall architecture of our Emotion-GCN model for FER in the wild. Our graph contains the seven expression labels and the two VA dimensions that are connected to each other based on the adjacency matrix $\mathbf{A}$. Stacked GCNs are learned over the graph to map the word embeddings of the nodes $\mathbf{Z}$ into a set of dependent classifiers and regressors, $\mathbf{W^c}$ and $\mathbf{W^r}$ respectively. These vectors are then applied to an image representation $\mathbf{x}$ extracted from the input image $\mathbf{I}$ via a DenseNet $\mathbf{f_{cnn}}$ followed by a global max-pooling function $\mathbf{f_{gmp}}$. The whole network is trained end-to-end for both basic expression classification ($\mathbf{L^c}$) and VA regression ($\mathbf{L^r}$).

the same procedure without passing through an activation function since its output is a predicted continuous value for either valence or arousal. We denote the expression classifiers by $\mathbf{W^c} \in \mathbb{R}^{7\times D}$ and the VA regressors by $\mathbf{W^r} \in \mathbb{R}^{2\times D}$ where each row of the matrices corresponds to a classifier and a regressor respectively. Traditionally, these vectors are optimized as individual parameters of the deep learning network. To capture the correlation between the categorical and the dimensional model, we propose to learn these vectors using a GCN that maps their word embeddings to dependent classifiers and regressors retaining the shared information between the two tasks.

1) GCN Overview: Generally, the goal of a GCN is to learn a function f on a graph $G = (V, E)$ that takes as input a feature description for each node of the graph $\mathbf{H^l} \in \mathbb{R}^{n\times d}$ $(n = |V|)$ and a correlation matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and produces new node-level features $\mathbf{H^{l+1}} \in \mathbb{R}^{n\times d'}$. The update rule is formulated as follows:

$$\mathbf{H^{l+1}} = h(\mathbf{\hat{A}H^lW^l}) \tag{2}$$

where $\mathbf{W^l} \in \mathbb{R}^{d\times d'}$ is the weight matrix for the l-th GCN layer, $\mathbf{\hat{A}}$ is the normalized version of matrix $\mathbf{A}$ such that all rows sum to one and $h(\cdot)$ is LeakyRELU [41]. For more information on GCN we refer readers to [30].

In our case, the nodes of the graph correspond to the seven expression labels and the two VA dimensions i.e. V = {Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger, Valence, Arousal}. Hence, the input is $\mathbf{Z} \in \mathbb{R}^{9\times d}$ that contains the word embedding of each node (d is the dimensionality of the embeddings). Each GCN layer l takes the node representations from the previous layer $\mathbf{H^l}$ as inputs and outputs new node representations $\mathbf{H^{l+1}}$ using (2). Finally, the output representation of the last layer $\mathbf{W} \in \mathbb{R}^{9\times D}$ contains the dependent classifiers and regressors. The first seven rows of the output matrix $\mathbf{W}$ constitute the classification part $\mathbf{W^c}$ and the rest two the regression part $\mathbf{W^r}$.

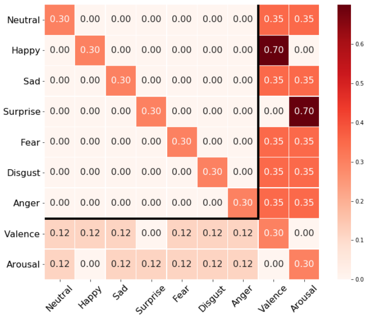
2) Adjacency Matrix Design: According to the update rule of a GCN (2) the feature of a node in the graph is the weighted sum of its own feature and the adjacent nodes' features. Since the purpose behind the use of a GCN is to exploit the dependencies between the categorical and the dimensional model, we should design the adjacency matrix $\mathbf{A}$ to this direction. As before, we assume that the first seven rows of $\mathbf{A}$ correspond to the basic expressions and the last two are the continuous dimensions i.e. valence and arousal. Since we deal with a multi-task and not a multi-label recognition problem, we are interested only in the correlation between the categorical and the dimensional model. Hence, we set the other pairs of $\mathbf{A}$ to zero except for the diagonal to allow self-loops. Also, we take the absolute value of the correlation to ignore its type (positive or negative) and focus on its amplitude. The correlation matrix $\mathbf{A} \in \mathbb{R}^{9\times 9}$ can be written as:

$$A_{ij} = \begin{cases} 1, & \text{if } \; i = j \\ |c_{ij}|, & \text{if } \; i \in \mathit{Cat} \;\wedge\; j \in \mathit{Dim} \\ |c_{ij}|, & \text{if } \; j \in \mathit{Cat} \;\wedge\; i \in \mathit{Dim} \\ 0, & \text{else} \end{cases} \tag{3}$$

where *Cat* and *Dim* are the set of indices of the categorical and the dimensional labels respectively. As a correlation metric, we use the Spearman's rank correlation coefficient [54] that for two variables $\mathbf{X} = \{x_1, ..., x_N\}$ and $\mathbf{Y} = \{y_1, ..., y_N\}$ is defined as:

$$c_{xy} = \frac{\sum_{k=1}^{N} x_{k,r} y_{k,r}}{\sqrt{\sum_{k=1}^{N} x_{k,r}^2 \sum_{k=1}^{N} y_{k,r}^2}} \tag{4}$$

where $x_{k,r}$ and $y_{k,r}$ are the rank transformation of the initial values x_k and y_k. Following the ideas in ML-GCN [9], we



	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Valence	Arousal
Neutral	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.35	0.35
Happy	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.70	0.00
Sad	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.35	0.35
Surprise	0.00	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.70
Fear	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.35	0.35
Disgust	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.35	0.35
Anger	0.00	0.00	0.00	0.00	0.00	0.00	0.30	0.35	0.35
Valence	0.12	0.12	0.12	0.00	0.12	0.12	0.12	0.30	0.00
Arousal	0.12	0.00	0.12	0.12	0.12	0.12	0.12	0.00	0.30

Fig. 3: Adjacency matrix of our Emotion-GCN.

use a threshold τ to filter the noisy edges as follows:

$$A'_{ij} = \begin{cases} 1, & \text{if } A_{ij} \geq \tau \\ 0, & \text{if } A_{ij} < \tau \end{cases} \tag{5}$$

where $\tau = 0.1$ to enable the propagation of information between weakly correlated nodes. As shown in [34], the learned features of each node may be over-smoothed and become indistinguishable when a binary correlation matrix is used. To alleviate the over-smoothing problem, we apply the re-weighted scheme of ML-GCN that is defined as:

$$A''_{ij} = \begin{cases} (p / \sum_{\substack{j=1 \\ i \neq j}}^{9} A'_{ij}) \times A'_{ij}, & \text{if } i \neq j \\ 1 - p, & \text{if } i = j \end{cases} \tag{6}$$

where the variable p determines the weights assigned to a node itself and its adjacent nodes. We set $p = 0.7$ to increase the influence of the neighborhood nodes. Fig. 3 shows the output adjacency matrix $\mathbf{A}''$ of Emotion-GCN using the training set of AffectNet. As we expected, pairs like happy-valence and surprise-arousal are strongly connected since the presence of the one usually denotes the presence of the other (Fig. 1).

C. Multi-task learning setting

After we generate the dependent matrices $\mathbf{W^c}$, $\mathbf{W^r}$ and the feature vector $\mathbf{x}$, we perform FER simultaneously on the categorical and the dimensional model. For the classification task, the predicted scores are computed as:

$$\hat{y}_i = \frac{e^{\mathbf{w_i^c} \cdot \mathbf{x}}}{\sum_{k=1}^{7} e^{\mathbf{w_k^c} \cdot \mathbf{x}}} \tag{7}$$

where $\hat{y}_i$ is the probability of emotion i and $\mathbf{w_i^c}$ is the classifier of emotion i. The network is trained using a weighted version of the traditional categorical cross entropy loss $\mathbf{L^c}$ since the dataset is highly imbalanced [44]. In other words, the network is penalized more for misclassifying samples from under-represented classes than from well-represented classes as follows:

$$L^c = -\sum_{i=1}^{7} w_i \; y_i \log(\hat{y}_i) \tag{8}$$

$$w_i = \frac{f_i}{f_{min}} \tag{9}$$

where $y_i = 1$ if class i is the ground truth expression, f_i is the number of samples of the i-th class and f_{min} is the number of samples in the most under-represented class i.e. Disgust. For the regression task, the predicted values are computed as:

$$\hat{p}_i = \mathbf{w_i^r} \cdot \mathbf{x} \tag{10}$$

where $\hat{p}_1, \hat{p}_2$ are the predicted values of valence and arousal respectively and $\mathbf{w_1^r}, \mathbf{w_2^r}$ correspond to the VA regressors. Most studies in the literature use the Mean Squared Error (MSE) as a loss function to train regression models. However, more and more works on emotion recognition use a Correlation-based loss function to measure the agreement between the true emotion dimension and the predicted emotion degree [31], [23]. The Concordance Correlation Coefficient (CCC) is often used since it takes the bias into Pearson's correlation coefficient [2]. It is defined as:

$$\rho_c = \frac{2 s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \tag{11}$$

where s_x and s_y denote the variance of the predicted and ground truth values respectively, $\bar{x}$ and $\bar{y}$ are the corresponding mean values and s_{xy} is the respective covariance value. The range of CCC is from -1 (perfect disagreement) to 1 (perfect agreement). Hence, in our case we define $\mathbf{L^r}$ as:

$$L^r = 1 - \frac{\rho_v + \rho_a}{2} \tag{12}$$

where ρ_v and ρ_a are the respective CCC of valence and arousal. The overall loss function of our network training is defined as $L = L^c + L^r$.

IV. Experiments

In this section, we present the experimental evaluation of the proposed Emotion-GCN model (Fig. 2). First, we briefly describe the facial expression datasets that we use and the details of our implementations. Then, an ablation study of our models follows and finally a comparison with the current state-of-the-art methods.

A. Datasets

1) AffectNet: AffectNet is by far the largest database of facial expressions that provides both categorical and VA annotations. It contains more than one million facial images collected from the Internet by querying three major search engines using 1,250 emotion-related keywords in six different languages [44]. About half of the retrieved images are manually annotated for the presence of seven discrete facial expressions (categorical model) and the intensity of valence and arousal (dimensional model). It is a very challenging database as it contains images of people from different races and ethnic groups as well as high variety in the background, lighting, pose, point of view, etc. To be consistent with the majority of past studies [61], [39], [8], [28], [14], [21], [25], we exclude the emotion of contempt and train our models on the images with neutral and the 6 basic emotions (approximately 280,000 training samples). Since the test set of AffectNet is not publicly available, we evaluate our

TABLE I: Performance of our models on the categorical model of AffectNet and Aff-Wild2.

Method	AffectNet	Aff-Wild2
Single-task	64.37	45.06
Multi-task + MSE	64.8	43.1
Multi-task + CCC	65.69	43.33
Emotion-GCN	**66.46**	**48.92**

TABLE II: Classification accuracy of Emotion-GCN on the categorical model of AffectNet using different values for τ, p and L (number of GCN layers). In the first table $p = 0.7$ and in the second table $\tau = 0.1$.

L \ τ	0	0.05	0.1	0.15
1	53.09	43.85	53.71	54.94
2	65.29	65.31	**66.46**	65.69
3	65.77	64.94	64.74	65.74

L \ p	0.2	0.3	0.4	0.5	0.6	0.7	0.8
1	49.6	50.94	52.57	52.66	52.86	53.71	59.31
2	65.71	66.14	65.29	66.29	66.09	**66.46**	65.06
3	65.29	65.54	65.89	66.09	65.54	64.74	65.83

TABLE III: Performance of our models on the dimensional model of AffectNet and Aff-Wild2.

Method	AffectNet		Aff-Wild2	
	CCC-V	CCC-A	CCC-V	CCC-A
Single-task	0.761	0.628	0.416	0.501
Multi-task + MSE	0.752	0.572	0.435	0.378
Multi-task + CCC	**0.768**	**0.651**	0.408	0.481
Emotion-GCN	0.767	0.649	**0.457**	**0.514**

approaches on the validation set that contains 500 images for each emotion. The mean class accuracy is used as the evaluation metric for the classification task because the validation set is balanced and the CCC for the regression task.

2) Aff-Wild2: Aff-Wild2 is the first ever database annotated for valence-arousal estimation, action unit detection and basic expression classification [31]. It consists of 548 videos collected from YouTube and shows both subtle and extreme human behaviours in real-world settings. We trained and evaluated our models on a subset of the database that contains only the frames with both categorical and VA annotations, as required by our method. As proposed by the authors of Aff-Wild2 a weighted average between the F1 score and mean class accuracy is used as the evaluation metric for the classification task ($0.67 \times F1 + 0.33 \times Acc$).

B. Implementation details

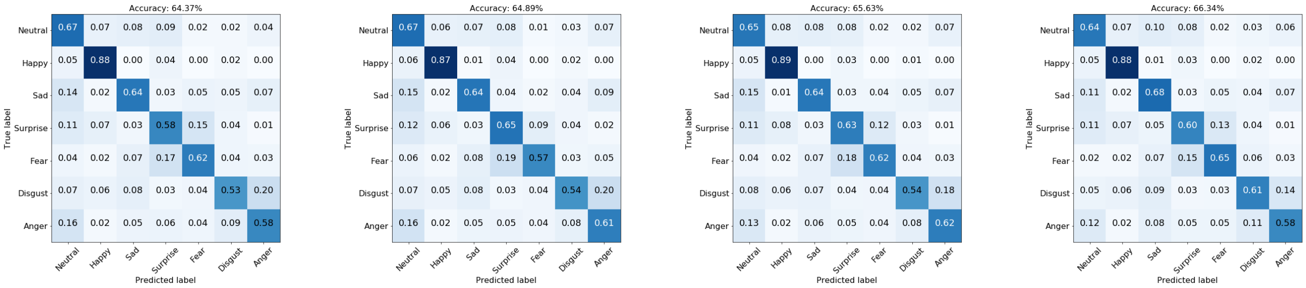
Before training the network, the face region of each image is cropped using the provided bounding box by the database and scaled to 227×227 pixels. Also, we perform face alignment based on the position of the eyes to obtain a normalized representation of each face. Specifically, we compute the location of the eyes by taking the mean value of the six detected landmarks in each eye. Then, we rotate the image by θ that is defined as $\theta = tan^{-1}\left(\frac{r_y-l_y}{r_x-l_x}\right)$, where $(r_x,\ r_y)$ and $(l_x,\ l_y)$ are the coordinates of the right and the left eye respectively. To augment the data, six types of augmentation techniques are used (flip, rotation and changes in brightness, contrast, hue and saturation) [25]. The GCN module consists of two layers defined with output dimensionality of 512 and 1024 respectively. For the word representations, we use 300-dimensional GloVe [49] trained on the Wikipedia dataset. The networks are trained for 10 epochs using a batch size of 35 and a learning rate of 0.001. Stochastic Gradient Descent is adopted as the optimization algorithm with a momentum of 0.9 and PyTorch [48] is used as our deep learning framework.

C. Ablation Study

First, we investigate the performance of four different networks on the categorical and the dimensional model of affect to present the improvements that Emotion-GCN introduces.

1) Categorical Model: We trained (i) a single-task network for discrete FER using the weighted CE loss of (8). Then, two multi-task networks were trained for discrete and continuous FER using a weighted CE loss for the classification task and a (ii) MSE or (iii) CCC loss (12) for the regression task. Finally, we trained (iv) the proposed Emotion-GCN model that generates dependent classifiers and regressors using a 2-layer GCN as presented in Fig. 2. Table I shows the results of our experiments. In AffectNet we can see that learning to predict the VA values as an additional task in a MTL framework increases the accuracy by **1.32%** since the shared representation improves the generalization of the network. Also, in agreement with recent studies, using the CCC loss for the regression task increases the classification accuracy of the model verifying that in MTL a correlation-based regression loss performs better. Finally, Emotion-GCN achieves a total accuracy of **66.46%**, since the dependent classifiers manage to effectively capture the dependencies between the two emotion representations. The confusion matrices for these models are shown in Fig. 4. It can be seen that our proposed method increases the accuracy for most classes while the single-task network performs better only in the neutral class. Similarly, in Aff-Wild2 the total evaluation metric increases by **3.86%** indicating that the benefits of Emotion-GCN generalize in more datasets. To investigate how different choices for the parameters of the GCN affect the performance, we perform additional experiments in Table II. We observe that the number of GCN layers (L) is the most crucial parameter since using only one layer decreases the accuracy a lot.

2) Dimensional Model: For the evaluation of our models on the VA space, the networks (ii), (iii) and (iv) are the same since they are trained for both discrete and continuous FER. Additionally, a single-task network for VA regression was trained using the CCC loss of (12). Table III shows the performance evaluation of our experiments on the dimensional model. In AffectNet we can see that the multi-task network with the CCC loss improves the regression performance along with the classification one verifying that both tasks benefit from the shared feature representation. Actually, we



(a) Single-task (b) Multi-task + MSE (c) Multi-task + CCC (d) Emotion-GCN

Fig. 4: Confusion matrices of our models on the validation set of AffectNet.

observe that the increase in the performance of arousal prediction is higher that that of valence. This is due to the fact that most emotions have positive value of arousal (Fig. 1) and a simultaneous emotion classification provides more useful information in the task of arousal regression. Finally, our GCN based approach achieves similar performance with that of the multi-task network on the dimensional model. In Aff-Wild2, our proposed model surpasses the performance of both the single-task and multi-task networks. Overall, our method presents significant improvements in both the categorical and the dimensional model. In Fig. 5 we present some positive and negative results of our Emotion-GCN model on AffectNet.

D. *Visualization*

To further analyze the effectiveness of our approach, we investigate the similarity between the dependent classifiers and regressors. In Fig. 6 the cosine similarity of the learned vectors on AffectNet by our single-task networks, our best multi-task network and our Emotion-GCN are presented. As we can see, learning a shared representation for both tasks through MTL slightly increases the similarity between the expression classifiers and the VA regressors. Our proposed method increases their similarity even more in consistence with the dependencies presented in Fig. 1. Specifically, the regressor of valence comes closer to the classifier of happy and the regressor of arousal closer to the classifiers of anger, disgust, fear and surprise. These emotions appear in regions of VA space where the values of valence or arousal are high. Therefore, the proposed network has successfully captured the dependence between the categorical and the dimensional emotion representation.

We also observe an increase in the similarity between the pairs valence-surprise and happy-surprise while their respective nodes are not adjacent in the graph (Fig. 3). These dependencies are successfully captured by our network since in a 2-layer GCN a node incorporates information from a 2-hop neighborhood [12]. To further examine whether the dependence between the emotions of happiness and surprise is reasonable, we compute their co-occurrence in a multi-label emotion dataset. The EMOTIC dataset [33] is a collection of images of people in unconstrained environments annotated according to their apparent emotional states. Each person is annotated for 26 discrete categories, with multiple labels assigned to each image. About 25% of the samples labeled as happy are labeled as surprise too that indicates that these two emotions are strongly related indeed.

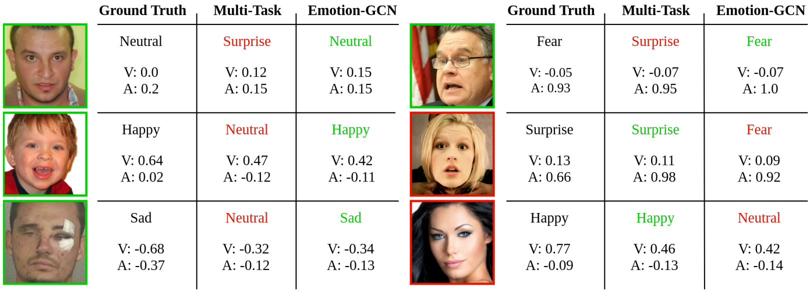
E. *Dependence between classifiers*

Inspired by the fact that there are dependencies between the basic emotions, we trained a similar GCN based network but we enabled this time the direct propagation of information between the seven emotion classifiers. Instead of setting their correlation to zero as in Emotion-GCN, we compute the conditional probability matrix of the basic emotions like in ML-GCN. The model achieves a total accuracy of 66.23% on the categorical model of AffectNet. Despite the intuition behind this approach, the recognition performance is slightly lower than that of our proposed method. We believe that by enabling the propagation of information between the basic emotions the dependence between the categorical and dimensional model is ignored. Also, we deal with a single-label recognition task and the possible benefits of this approach are more suitable in a multi-label recognition task.

F. *Comparison with the State of the Art*

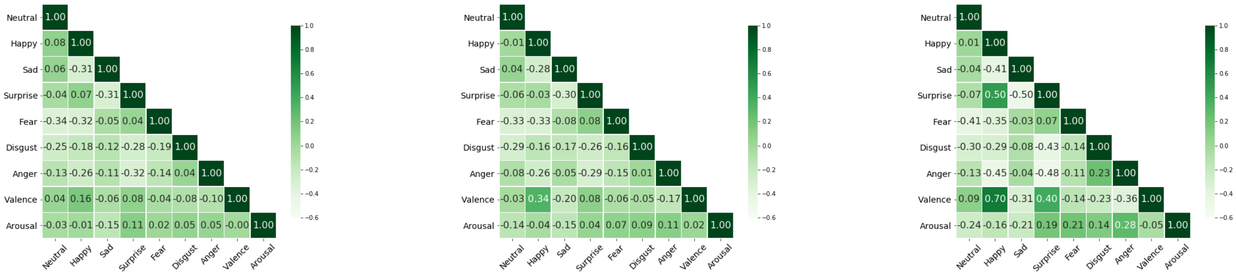
In Table IV, we compare the performance of our Emotion-GCN model with several state-of-the-art methods for FER on the categorical model of AffectNet. Regarding Aff-Wild2, we only used the subset which contains both categorical and continuous annotations, while methods in the bibliography typically report results on the whole dataset, making direct comparisons not possible. In IPA2LT [61], the authors designed LTNet to discover the latent truths from the human annotations and the machine annotations trained from different FER datasets. In gACNN [39] and OADN [14] attention networks were proposed for occlusion aware FER. Facial Motion Prior Network in [8] generates a facial mask so as to focus on facial muscle moving regions. In [21] deep (CNNs) and handcrafted features (BOVW) were combined and in [25] a deep Siamese network along with a supervised loss function was used to reduce the intra-class variation of the task. Closer to our work, CAKE [28] proposed a 3-dimensional representation of emotion learned in a multi-domain fashion. Our Emotion-GCN model considerably outperforms these recent state-of-the-art methods, achieving an accuracy of **66.46%**.

In [24] BReG-NeXt achieved state-of-the-art accuracy in 8-way classification on AffectNet (including contempt)



Ground Truth	Multi-Task	Emotion-GCN
Neutral V: 0.0 A: 0.2	Surprise V: 0.12 A: 0.15	Neutral V: 0.15 A: 0.15
Happy V: 0.64 A: 0.02	Neutral V: 0.47 A: -0.12	Happy V: 0.42 A: -0.11
Sad V: -0.68 A: -0.37	Neutral V: -0.32 A: -0.12	Sad V: -0.34 A: -0.13
Fear V: -0.05 A: 0.93	Surprise V: -0.07 A: 0.95	Fear V: -0.07 A: 1.0
Surprise V: 0.13 A: 0.66	Surprise V: 0.11 A: 0.98	Fear V: 0.09 A: 0.92
Happy V: 0.77 A: -0.09	Happy V: 0.46 A: -0.13	Neutral V: 0.42 A: -0.14

Fig. 5: Predictions of our models on samples of AffectNet. The first column indicates the ground truth values. The second and the third column present the predictions of the multi-task network trained with CCC loss and our Emotion-GCN respectively. To examine which network better exploits the emotional dependencies, we selected samples where the predictions of the networks on the dimensional model are close. In the first four samples our Emotion-GCN model successfully recognizes the depicted emotion while the multi-task network fails indicating that our proposed model effectively captured the dependencies presented in the VA space (Fig. 1). However, there are cases where our network fails (last two samples).



(a) Single-task

(b) Multi-task + CCC

(c) Emotion-GCN

Fig. 6: Visualization of the cosine similarity between the learned classifiers and regressors by our models on AffectNet.

TABLE IV: Comparison with state-of-the-art methods on AffectNet (7-way classification).

Method	Accuracy
IPA2LT [61]	57.31
gACNN [39]	58.78
Facial Motion Prior Network [8]	61.52
CAKE [28]	61.7
OADN [14]	61.89
CNNs and BOVW + global SVM [21]	63.31
Siamese [25]	64
Emotion-GCN (ours)	**66.46**

by introducing a residual-based network architecture. To investigate the ability of Emotion-GCN to generalize over other model architectures as well, we replaced our DenseNet backbone network with BReG-NeXt using the publicly available code[1]. Since the provided code does not include the data preprocessing strategy, we followed our preprocessing and training pipeline described before (same with DenseNet), which explains the different results compared to [24]. As we can see in Table V, our proposed model outperforms the single-task and multi-task models when using BReG-NeXt as the backbone network indicating that Emotion-GCN generalizes across different model architectures as well.

V. Conclusion

In this work, a novel GCN based MTL framework is proposed for in-the-wild FER. Specifically, our Emotion-GCN model learns a shared feature representation for both discrete and continuous expression recognition to exploit the dependencies between the categorical and the dimensional model of affect. To further capture these dependencies, the expression classifiers and the VA regressors are learned though a GCN that maps their word representation to dependent vectors inspired by recent work in multi-label image recognition. Experimental results on AffectNet, the largest facial expression database, have demonstrated that our Emotion-GCN outperforms the performance of the recent state-of-the-art methods for discrete FER.

TABLE V: Performance of Emotion-GCN using BReG-NeXt as the backbone network on AffectNet.

Method	Network	Accuracy
Single-task	BReG-NeXt	60.49
Multi-task+CCC	BReG-NeXt	61.14
Emotion-GCN	BReG-NeXt	**61.94**

[1] `https://github.com/behzadhsni/BReG-NeXt`

References

[1] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool. Covariance pooling for facial expression recognition. In *Proc. CVPRW*, 2018.

[2] B. T. Atmaja and M. Akagi. Evaluation of error and correlation-based loss functions for multitask learning dimensional speech emotion recognition. *arXiv preprint arXiv:2003.10724*, 2020.

[3] P. M. Blom, S. Bakkes, C. Tan, S. Whiteson, D. Roijers, R. Valenti, and T. Gevers. Towards personalised gaming via facial expression recognition. In *Proc. AIIDE*, 2014.

[4] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong. Island loss for learning discriminative features in facial expression recognition. In *Proc. FG*, 2018.

[5] P. Carcagnì, M. Del Coco, M. Leo, and C. Distante. Facial expression recognition and histograms of oriented gradients: a comprehensive study. *SpringerPlus*, 4, 2015.

[6] R. Caruana. Multitask learning. *Machine learning*, 28, 1997.

[7] W.-Y. Chang, S.-H. Hsu, and J.-H. Chien. Fatauva-net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In *Proc. CVPRW*, 2017.

[8] Y. Chen, J. Wang, S. Chen, Z. Shi, and J. Cai. Facial motion prior networks for facial expression recognition. In *Proc. VCIP*, 2019.

[9] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo. Multi-label image recognition with graph convolutional networks. In *Proc. CVPR*, 2019.

[10] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Trans. on PAMI*, 38, 2016.

[11] C. Darwin. *The expression of the emotions in man and animals*. University of Chicago press, 2015.

[12] T. Derr, Y. Ma, and J. Tang. Signed graph convolutional networks. In *Proc. ICDM*, 2018.

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

[14] H. Ding, P. Zhou, and R. Chellappa. Occlusion-adaptive deep network for robust facial expression recognition. In *Proc. IJCB*, 2020.

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

[16] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proc. ICCV*, 2015.

[17] P. Ekman. Strong evidence for universals in facial expressions: a reply to russell's mistaken critique. *Psychological bulletin*, 1994.

[18] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17, 1971.

[19] 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. CVPR*, 2016.

[20] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos. Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction. *Robotics and Automation Letters*, 4, 2019.

[21] M.-I. Georgescu, R. T. Ionescu, and M. Popescu. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, 7, 2019.

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

[23] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proc. ACM international conference on Multimedia*, 2017.

[24] B. Hasani, P. S. Negi, and M. Mahoor. Breg-next: Facial affect computing using adaptive residual networks with bounded gradient. *IEEE Trans. on Affective Computing*, 2020.

[25] W. Hayale, P. Negi, and M. Mahoor. Facial expression recognition using deep siamese neural networks with a supervised loss function. In *Proc. FG*, 2019.

[26] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proc. ICMI*, 2017.

[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proc. CVPR*, 2017.

[28] C. Kervadec, V. Vielzeuf, S. Pateux, A. Lechervy, and F. Jurie. Cake: Compact and accurate k-dimensional representation of emotion. *CoRR*, 2018.

[29] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proc. ICMI*, 2015.

[30] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *Proc. ICLR*, 2017.

[31] D. Kollias and S. Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *BMVC*, 2019.

[32] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65, 2017.

[33] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Emotic: Emotions in context dataset. In *Proc. CVPRW*, 2017.

[34] Q. Li, Z. Han, and X.-M. Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proc. AAAI Conference on Artificial Intelligence*, 2018.

[35] Q. Li, M. Qiao, W. Bian, and D. Tao. Conditional graphical lasso for multi-label image classification. In *Proc. CVPR*, 2016.

[36] S. Li and W. Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Trans. on Image Processing*, 28, 2018.

[37] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE Trans. on Affective Computing*, 2020.

[38] X. Li, F. Zhao, and Y. Guo. Multi-label image classification with a probabilistic label enhancement model. In *Proc. UAI*, 2014.

[39] Y. Li, J. Zeng, S. Shan, and X. Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Trans. on Image Processing*, 28, 2018.

[40] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. on Image processing*, 11, 2002.

[41] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, 2013.

[42] D. Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16, 1992.

[43] D. McDuff, R. El Kaliouby, J. F. Cohn, and R. W. Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Trans. on Affective Computing*, 6, 2014.

[44] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. on Affective Computing*, 10, 2017.

[45] S. Moore and R. Bowden. Local binary patterns for multi-view facial expression recognition. *Computer vision and image understanding*, 115, 2011.

[46] G. Muhammad, M. Alsulaiman, S. U. Amin, A. Ghoneim, and M. F. Alhamid. A facial-expression monitoring system for improved healthcare in smart cities. *IEEE Access*, 5, 2017.

[47] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. on Affective Computing*, 2, 2011.

[48] 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. In *NIPS-W*, 2017.

[49] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proc. EMNLP*, 2014.

[50] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *Proc. FG*, 2017.

[51] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39, 1980.

[52] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, 2014.

[53] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27, 2009.

[54] C. Spearman. The proof and measurement of association between two things. *International journal of epidemiology*, 1961.

[55] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Proc. FG*, 2011.

[56] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proc. CVPR*, 2020.

[57] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. on Image Processing*, 29, 2020.

[58] W. Xiaohua, P. Muzi, P. Lijuan, H. Min, J. Chunhua, and R. Fuji. Two-level attention with two-stage multi-task learning for facial emotion recognition. *Journal of Visual Communication and Image Representation*, 62, 2019.

[59] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proc. ICMI*, 2015.

[60] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proc. CVPRW*, 2017.

[61] J. Zeng, S. Shan, and X. Chen. Facial expression recognition with inconsistently annotated datasets. In *Proc. ECCV*, 2018.