Multi-Label Compound Expression Recognition: C-EXPR Database & Network

Dimitrios Kollias
Queen Mary University of London, UK
d.kollias@qmul.ac.uk

Abstract

Research in automatic analysis of facial expressions mainly focuses on recognising the seven basic ones. However, compound expressions are more diverse and represent the complexity and subtlety of our daily affective displays more accurately. Limited research has been conducted for compound expression recognition (CER), because only a few databases exist, which are small, lab controlled, imbalanced and static. In this paper we present an in-the-wild A/V database, C-EXPR-DB, consisting of 400 videos of 200K frames, annotated in terms of 13 compound expressions, valence-arousal emotion descriptors, action units, speech, facial landmarks and attributes. We also propose C-EXPR-NET, a multi-task learning (MTL) method for CER and AU detection (AU-D); the latter task is introduced to enhance CER performance. For AU-D we incorporate AU semantic description along with visual information. For CER we use a multi-label formulation and the KL-divergence loss. We also propose a distribution matching loss for coupling CER and AU-D tasks to boost their performance and alleviate negative transfer (i.e., when MT model's performance is worse than that of at least one single-task model). An extensive experimental study has been conducted illustrating the excellent performance of C-EXPR-NET, validating the theoretical claims. Finally, C-EXPR-NET is shown to effectively generalize its knowledge in new emotion recognition contexts, in a zero-shot manner.

1. Introduction

For the past twenty years research in automatic analysis of facial behaviour was mainly limited to the recognition of the so-called six universal expressions (e.g., anger, happiness), plus the neutral state, influenced by the seminal work of Ekman [7]. However, the affect model based on basic expressions is limited in the ability to represent the complexity and subtlety of our daily affective displays [19]. Many more facial expressions exist and are used regularly by humans. The compound expressions are a better representation of affective displays in everyday interactions. Compound means that the expression category is constructed as a combination of two basic expression categories. Obviously, not all combinations are meaningful for humans. Twelve compound expressions are most typically expressed by humans, e.g., people regularly produce a happily surprised expression and observers do not have any problem distinguishing it from an angrily surprised expression.

The design of systems capable of understanding the community perception of expressional attributes and affective displays is receiving increasing interest. Benefited from the great progress in deep learning research, the performance of expression recognition has greatly improved. However, deep-model based methods are starved for labeled data, whereas the annotation is a highly labor intensive and time consuming process and the complexity of expression categories obscures the labelling procedure. Initially, research was mainly limited to posed behavior captured in highly controlled conditions. Some representative datasets are CK+ [17], MMI [26], CFEE [6] and iCV-MEFED [18].

However, it is now widely accepted that progress in a particular application domain is significantly catalysed when a large number of datasets are collected in-the-wild (i.e., in unconstrained conditions). Thus, expression analysis could not only focus on spontaneous behaviors, but also on behaviours captured in unconstrained conditions. Hence, two in-the-wild databases have been generated, EmotioNet [1] and RAF-DB [15]. These databases, although in-the-wild, are: i) very small in terms of size (RAF-DB contains around 4,000 images; EmotioNet contains around 1,500 images); ii) very imbalanced (in RAF-DB one category consists of 1,700 images; in EmotioNet one category contains half the samples); iii) static (i.e., they contain only images); iv) lacking a training-validation-test set split.

It is evident that these databases are very small and do not contain sufficient data for both training and evaluating deep learning systems, so that the results are meaningful and illustrate good generalization. Compound expression recognition (CER) is in its infancy due to the above limitations. To this end, we collected the largest, diverse, in-the-wild audiovisual database, C-EXPR-DB, reliably annotated for 12 compound expressions plus a category referring to

other affective states. C-EXPR-DB is also annotated for: i) continuous dimensions of valence-arousal (how positive-negative, active-passive the emotional state is); ii) speech detection; iii) facial landmarks and bounding boxes; iv) action units (activation of facial muscles); v) facial attributes.

Recently, some works have utilized multi-task learning (MTL) for basic expression recognition and AU detection (AU-D) [2, 3]. They have shown that MTL helps improve the performance across tasks. Inspired by this, we proposed a novel methodology, C-EXPR-NET, for MTL of CER and AU-D; we are interested in CER but we use AU-D as auxiliary task to enhance CER performance. We utilize SEV-Net [30] for AU-D. The FACS manual [7] provides a complete set of textual descriptions for AU definitions; such a set of AU descriptions provides rich semantic information (about facial area/position, action, motion direction/intensity, relation of AUs). The model introduces such AU semantic descriptions as auxiliary information and processes them via inter- and intra-transformer modules and a cross modality attention module for AU-D.

For CER we use a multi-label formulation, where the output classes are 6 (the basic expressions) and each datum contains annotations for 2 of these 6 categories. We use a softmax activation in the output layer of our method that tackles CER and the Kullback-Leibler divergence (KL-div) as its loss function. The concept for this formulation aligns with what compound expressions are, i.e., expressions that can be constructed as combinations of basic categories. In addition, this formulation allows our method to be additionally trained with images annotated with the 6 basic expressions (as a form of data augmentation or to learn to differentiate between basic expressions as well). This formulation deviates from traditional CER approaches that use a multi-class formulation, with the compound expressions being mutually exclusive classes (i.e., each datum is annotated in terms of only one compound expression).

However, when we compared the performance of our multi-task method with that of single-task (ST) methods for CER and AU-D, we observed that MTL increased CER performance, but it harmed AU-D performance. Thus negative transfer occurred, as CER task dominated the training process. Inspired by [11, 12], we propose a distribution matching approach based on task relatedness, i.e., knowledge exchange between CER and AU-D tasks is enabled via distribution matching over their predictions. We demonstrate empirically that this distribution matching approach alleviates negative transfer and further boosts CER performance. The main contributions of this work are summarized below:

- We generate the largest, diverse, in-the-wild A/V database, C-EXPR-DB, annotated for compound expressions, valence-arousal, AUs, facial attributes, speech detection, facial landmarks and bounding boxes;
- We propose the novel C-EXPR-NET, a MT method for CER and AU-D; the latter task acts as an auxiliary one for enhancing the former task's performance (we are the first to prove this). For AU-D, our method incorporates visual information, as well as AU descriptors (that act as auxiliary, rich, semantic information) and processes them via inter- and intra-transformer modules and a cross-modality attention module. For CER our method uses a multi-label formulation and KL-div loss. Our method finally contains a distribution matching loss, based on task relatedness, for coupling the tasks to alleviate negative transfer and further boost their performance.
- We conduct an extensive experimental study which shows that: i) C-EXPR-NET outperforms the state-of-the-art (sota) both for CER and BER on RAF-DB, regardless if trained from scratch or pre-trained on C-EXPR-DB; ii) C-EXPR-NET outperforms the sota regardless if AU annotations are manual or automatic; iii) C-EXPR-NET can effectively generalize its knowledge in new emotion recognition contexts, in a zero-shot manner.

2. Related Work

DLP-CNN+mSVM [15] is pretrained for BER and used as a feature extractor for CER; it is trained using the softmax loss and a locality preserving loss - that pulls the locally neighboring faces of the same class together. In [16], a loss is introduced - for BER and CER - that consists of a separate loss and the classical cross entropy; the separate loss consists of an intra-class and an inter-class loss, both based on normalized cosine similarity. ReCNN [29] uses VGGFACE for extracting features cropped into sub-feature maps and processed by other layers. A weighted cross entropy loss is optimized during training for BER and CER. ResNet-18 (ARM) [25] is proposed for BER, consisting of a backbone network that extracts visual features, an auxiliary block that rearranges the features and two functional blocks, for realizing the features' weight distribution by means of convolution and for simplifying the representation learning by splitting the features to two parts.

DACL [8] consists of a backbone network that extracts facial features, a module which eliminates features' noise and irrelevant information and a multi-head binary classification module that calculates the attention weights for the 'cleaned' features; DACL is trained for BER with sparse center loss and softmax cross entropy loss. PSR [27] is a pyramid architecture of a spatial transformer (for alignment), a scaling module (for processing input on different scales), a low- and high-level extractor, a classifier and a concatenation block for BER. [10] is a methodology for generating realistic images with valence-arousal and 6 basic expressions. The authors train VGG-FACE on many databases, augmenting their training set with generated images from the particular database. Finally [11, 12] introduce FaceBehaviorNet, a multi-task model targeting valence-

arousal estimation, BER and AU-D. The network was further utilized in a zero-shot and few-shot setting for CER.

3. C-EXPR-DB Database

Data collection All videos of Compound-Expression-DataBase (C-EXPR-DB) have been downloaded from YouTube. For finding videos with compound expressions, we searched YouTube with different expression related keywords (one of the basic or compound categories or synonym words for them); activities, reactions and actions-causes that trigger or induce these expressions. More details regarding the data and their collection are included in the supplementary material.

Data Properties We downloaded 400 videos of people exhibiting compound behaviors in arbitrary recording conditions (in-the-wild, with high variations in poses, lightning-illumination, background noise levels etc). The total length of the videos is more than 13 hours and the total number of frames is around 200,000. Most spontaneous/in-the-wild A/V databases in affective computing do not contain as many subjects as C-EXPR-DB: DISFA [20] (27), BP4D [32] (41), RECOLA [23] (46), GFT [9] (96), BP4D+ [33] (140), Aff-Wild [13, 31] (200), AMFED [21] (242), AFEW [5] (330), SEWA [14] (398). The subjects in C-EXPR-DB come from different cultural backgrounds and ethnicities with a large age range. Table 1 shows a summary of database's statistics. Images from the database can be seen in supplementary material.

Table 1. C-EXPR-DB's Statistics; CE: Compound Expressions

Attribute	Value
# Videos	400
# Frames	198,978
# Annotators	7
# Modes of Affect	CE & valence-arousal & AUs

Data Annotation C-EXPR-DB contains per-frame annotations for: i) 13 expression categories, ii) valence-arousal, iii) action units (AUs), iv) speech detection, v) facial landmarks and face bounding boxes, vi) facial attributes.

Each frame of the database has been annotated by seven expert annotators for twelve compound expressions, plus the "other" state. The 12 compound expressions are shown in Table 2, along with their total number of annotated frames. The "other" state includes all affective states that are not one of the twelve compound expressions. The same experts further annotated each frame of the database in terms of the continuous dimensions of valence-arousal. For accurately performing frames' annotation, experts exploited all available modalities, namely facial expressions, audio, context, body pose and gesture. When the annotations have been completed, we applied a post-processing step. For each expert, we removed their annotations for frames for which there was a mismatch between valence-arousal and compound expression labels. Table 2 shows the valence-arousal expected values for each compound category.

Table 2. C-EXPR-DB's Annotations and their valence-arousal expected range

Expression	# Frames	Valence-Arousal Range
Sadly Fearful	10,112	$V < 0, A > 0$
Sadly Surprised	10,780	$V < 0, A > 0$
Sadly Disgusted	10,765	$V < 0, A > 0$
Sadly Angry	8,878	$V < 0, A > 0$
Fearfully Angry	11,591	$V < 0, A > 0$
Fearfully Surprised	14,445	$V < 0, A > 0$
Fearfully Disgusted	10,356	$V < 0, A > 0$
Angrily Surprised	10,535	$V < 0, A > 0$
Angrily Disgusted	9,415	$V < 0, A > 0$
Disgustedly Surprised	10,637	$V < 0, A > 0$
Happily Surprised	24,915	$V > 0, A > 0$
Happily Disgusted	8,885	$A > 0$
Other	44,456	V/A $\in [-1, 1]$

Additionally, a FACS trained AU coder annotated C-EXPR-DB with 17 AUs (AU 1,2,4,5,6,7,9,10,11,12,15,17, 20,23,24,25,26). The experts also labeled the frames on which the subject is speaking. More details regarding the annotators and their annotations in terms of compound expressions, valence-arousal, AUs and speech can be found in the supplementary material. Additionally, facial landmarks were annotated for all frames. Manual annotation of facial landmarks is highly labour intensive. Based on [24], trained annotators can only achieve a sustained annotation speed of 30 frames per hour and thus it would be impractical to manually annotate all of the frames of C-EXPR-DB. Therefore, the annotation was performed semi-automatically. Experts manually labeled with landmark points (and bounding box) the first frame of each video in which the subject appeared; then these landmarks were provided to the MDNET face tracker [22] that automatically annotated the rest of the frames. In parallel, we used the RetinaFace [4] to detect facial landmarks in all frames; we compared the outputs of the tracker and detector to find wrong detections, which were manually corrected. Finally, the experts manually annotated facial attributes in all videos. The specific attributes can be found in the supplementary material.

Major Contribution of C-EXPR-DB It is the compound expression annotations along with its in-the-wild nature, the fact that it is A/V as well as large and that each frame has been annotated by 7 experts. We also need to stress the fact that the database contains so many different annotations and especially its annotations in terms of three different modes of affect. C-EXPR-DB's multimodality (faces, body and pose, gestures, audio/speech) is important as true emo-

tion inference requires multiple modalities to disambiguate emotions that map to similar expressions in one modality. It contains video sequences that show the evolution of the compound expressions through time with all its development (onset/apex/offset); these video sequences contain different compound expressions and intensities, different body posture and hand gestures per identity. These can facilitate and foster research on image/video generation (e.g. GANs).

4. The Proposed Method: C-EXPR-NET

Fig. 1 gives an overview of our proposed framework, Compound-Expression-Network (C-EXPR-NET) for recognizing compound expressions.

Problem Statement For a given image $x \in \mathcal{X}$, we can have label annotations:

i) in terms of 6 basic expressions $y_{expr} \in \{0,1\}^6$, where the 6 classes are mutually exclusive (multi-class problem); their notation is in one-hot encoding, e.g, "happy" is denoted as $[0,0,0,1,0,0]$; or

ii) in terms of 12 compound expressions $y_{expr} \in \{0,0.5\}^6$, where the 12 classes are non mutually exclusive (multi-label problem), e.g, the "happily surprised" class (consisting of "happy" and "surprise" classes) is denoted as $[0,0,0,0.5,0,0.5]$; and

iii) in terms of 17 binary action units $y_{AU} \in \{0,1\}^{17}$

Generation of Missing Labels for the AU Auxiliary Task If no AU labels exist, then we generate automatic ones. We merge many AU annotated databases so as: i) to have samples annotated in terms of all 17 AUs and ii) to have an adequate amount of samples that can produce a good AU-D performance. Then we train a CNN-RNN model on them. This will act as the Teacher Pre-trained Network of Fig. 1 that will provide the AU labels to our proposed method. More details exist in the supplementary material.

Method's Components C-EXPR-NET consists of 5 parts: Backbone Network (BNet); Expression Branch (ExprB); AU Branch (AUB); Distribution Matching (DM) module; Data Augmentation (DA) module.

Backbone Network (BNet): The input image is encoded by the Backbone Network to spatial visual features $\mathcal{V} \in \mathcal{R}^{H \times W \times D}$; W, H, D are the feature map's width, height, depth. Any CNN/CNN-RNN can be Backbone Net.

Expression Branch (ExprB): The visual features $\mathcal{V}$ are fed into the Expression Block (i.e., resnet block) that produces features $\mathcal{V}' \in \mathcal{R}^{H' \times W' \times D'}$, where W', H', D' are the feature map's width, height, depth; these features are followed by a classifier with softmax that produces the expression probabilities $p_{expr} \in \{0,0.5\}^6$. The loss function, related to this block, is KL-div that minimizes the distance between expression labels' and predictions' distributions. KL-div is defined for two probability distributions y_{expr} and p_{expr} corresponding to the expression labels and predictions -N being the total number of images fed to the module- as:

$$\begin{aligned}\mathcal{L}_{expr} &= \frac{1}{N}\sum_{k=1}^{N} \mathrm{KL}(y_{expr}^k || p_{expr}^k) \\ &= \frac{1}{N}\sum_{k=1}^{N} \left\{ \sum_{(y_{expr}^k, p_{expr}^k)} y_{expr}^k \cdot \log\left(\frac{y_{expr}^k}{p_{expr}^k}\right) \right\} \end{aligned} \quad (1)$$

AU Branch (AUB): It consists of two parts: the AU Semantic Encoding part and the Cross-Modality Attention one, as is the case with SEV-Net. At first, based on the FACS manual, a summary of the 17 AU semantic descriptions is created; some of these can be seen in the bottom of Fig. 1. These descriptions denote the process to spot and annotate the particular action associated with each AU activation. Each description consists of multiple sentences and each sentence consists of words.

Then we split each AU semantic descriptor into tokens by using the WordPiece tokenizer [28] and assign positional encoding to each word. Thus, for each token, its input representation is the sum of its trainable word, segment and positional embedding. Each AU semantic descriptor is fed to an Intra-Encoder module that consists of a multi-layer transformer network that encodes contextual information for tokens within each sentence. The output of the Intra-Encoder is a set of embeddings for each AU semantic descriptor; these embeddings are fed to the Inter-Encoder module, which is a multi-layer transformer encoder that captures the inter-AU relations (among multiple descriptor embeddings). The Inter-Encoder module outputs embeddings, $\mathcal{E}_i$, $i = 1, .., 17$, one embedding per AU descriptor.

All the AU embeddings $\mathcal{E}_i$ are then fed to the Cross-Modality Attention module, along with the output of the AU Block, which is a resnet block taking as input the visual features $\mathcal{V}$ and produces output features $\mathcal{V}'' \in \mathcal{R}^{H'' \times W'' \times D''}$, where W'', H'' and D'' are the width, height and depth of the feature map. The Cross-Modality Attention module outputs attention maps; each category-specific cross-modality attention map a_i^j for AU_i at location j is defined as:

$$a_i^j = \mathrm{ReLU}\Big(\cos\theta_i^j\Big) \Big/ \sum_{j=1}^{W'' \times H''} \mathrm{ReLU}\Big(\cos\theta_i^j\Big) \quad (2)$$

where $\cos\theta_i^j$ is the cosine similarity between the output features $\mathcal{V}_j''$ of the AU Block and the embeddings $\mathcal{E}_i$ of the Inter-Encoder module (the latter is needed to be at first linearly projected to the same dimensions of the former). Then, the attention maps are multiplied with the output features of the AU Block, thus acting as weights for the ag-

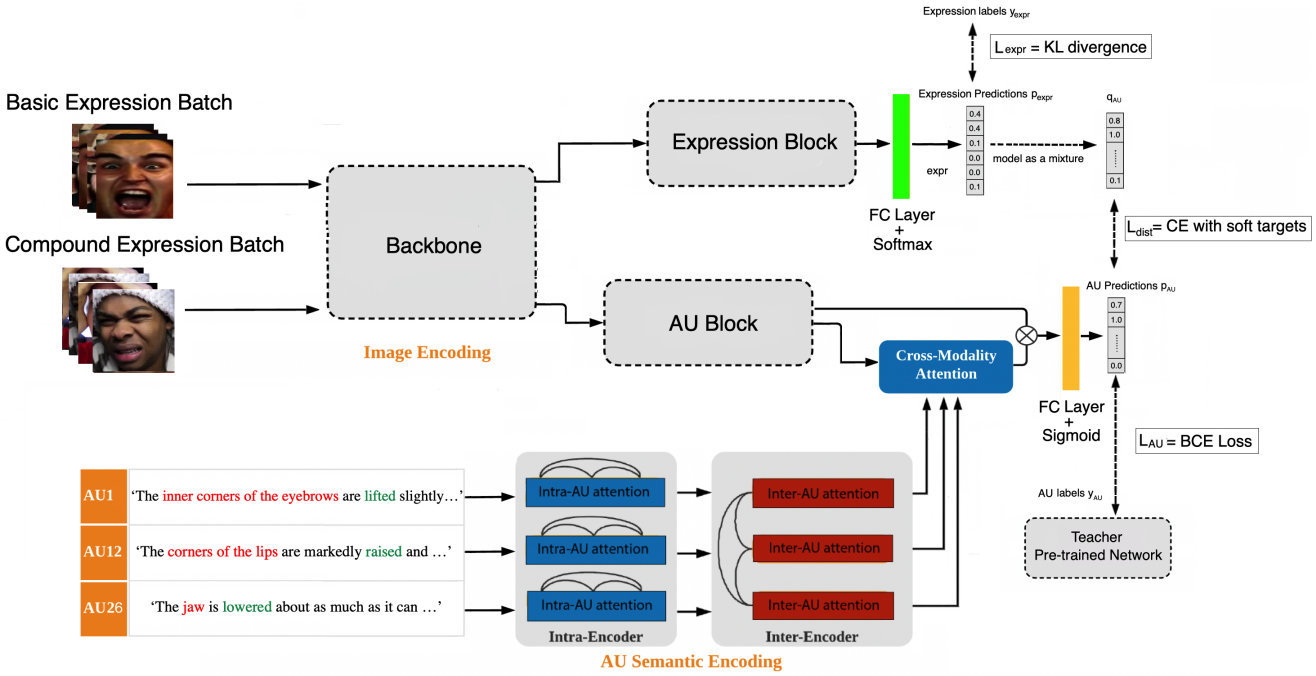

Figure 1. The proposed C-EXPR-NET. The visual features are extracted by a Backbone network and fed to an Expression and an AU block for producing new features. The Expression Block's features are fed to a classifier that produces compound expression predictions p_{expr}. By utilizing knowledge about the relatedness between expressions and AUs, q_{AU} is derived from p_{expr} (i.e, AUs are modeled as a mixture over the expression predictions). AU semantic embeddings are obtained through the Intra-AU and Inter-AU attention modules; the former captures the relation among words within each sentence that describes individual AUs, and the latter focuses on the relation among those sentences. The learned AU semantic embeddings and the features of the AU Block are used to generate attention maps via a cross-modality attention module. The attention maps are used as weights for the aggregated features; their product is fed to a classifier that produces AU predictions p_{AU}. Finally, we match p_{AU} with q_{AU} to make the predicted AUs consistent with the activated AUs of the predicted expressions.

gregated features. A high value in a specific location j of AU_i denotes that the location j is more important than other locations for recognizing AU_i; thus the model needs to pay more attention to that location when detecting that specific AU. Finally the output of this multiplication is fed to a classifier with sigmoid activation function. This branch produces the predictions $p^i_{AU}, i = 1, \ldots, 17$ for the AUs. The loss function related to this is a binary cross entropy (where N is the total number of images):

$$\mathcal{L}_{AU} = \frac{1}{17 \cdot N} \sum_{k=1}^{N} \sum_{i=1}^{17} \left[y^{i,k}_{AU} \log p^{i,k}_{AU} + (1 - y^{i,k}_{AU}) \log (1 - p^{i,k}_{AU}) \right] \tag{3}$$

To sum up, this branch applies the attention in 3 levels to capture different AU semantic relations: i) words level (location, action type/intensity, etc); ii) sentence level (AU relations, can 2 AUs happen concurrently?); iii) modality level (connecting AU semantic embeddings to visual features). As a result, the model is able to learn more discriminative features from more meaningful areas.

Distribution Matching (DM) module: We generate a new distribution $q^i_{AU}, i = 1, \ldots, 17$ where the AUs are modeled as a mixture over the expression categories. The new distribution is derived from the predictions of the expression branch p_{expr} according to a pre-defined relatedness between expressions and AUs. This relatedness can be extracted from the psychological study of [6], where it is found that, e.g., AU12 is activated when someone is happy (related to "happily surprised" and "happily disgusted" expressions), thus $q^{12}_{AU} = p_{\text{happy}}$. In another case, AU4 is activated in sadness, fear, anger and disgust (related to all compound expressions, apart from "happily surprised" and "happily disgusted"), thus $q^4_{AU} = p_{\text{sadness}} + p_{\text{fear}} + p_{\text{anger}} + p_{\text{disgust}}$. The new distribution is therefore defined as:

$$q^i_{AU} = \sum_{p_{expr}} p_{expr} \cdot p_{AU_i|expr} \tag{4}$$

where $p_{AU_i|expr}$ is defined deterministically from Table 1 of [6] (in supplementary); it is 1 if AU_i is activated for the particular "expr" and 0 otherwise.

We match distributions p^i_{AU} and q^i_{AU} -so as to make the predicted AUs consistent with the activated AUs of the predicted expressions- by minimizing the cross entropy with the soft target loss term (N is the total number of images):

$$\mathcal{L}_{dist} = \frac{\sum_{k=1}^{N}\sum_{i=1}^{17}\left[p^{i,k}_{AU}\log q^{i,k}_{AU} + (1-p^{i,k}_{AU})\log(1-q^{i,k}_{AU})\right]}{17 \cdot N} \tag{5}$$

Data Augmentation (DA) module: To increase the size of the training set, we further add images annotated in terms of the 6 basic expressions (e.g. from RAF-DB).

Overall Loss Function The overall objective function is the sum of the losses defined previously (α_1, α_2 and α_3 control the relative importance of each term):

$$\mathcal{L}_{overall} = \alpha_1\mathcal{L}_{expr} + \alpha_2\mathcal{L}_{AU} + \alpha_3\mathcal{L}_{dist} \tag{6}$$

When experimenting with the A/V C-EXPR-DB, we utilize the visual modality by feeding images to C-EXPR-NET (C-EXPR-NET-V). However, C-EXPR-NET is modality agnostic; thus we extract spectrograms from the audio modality and feed them to C-EXPR-NET (C-EXPR-NET-A). Finally, we perform late fusion on the expression and AU logits of C-EXPR-NET-V and C-EXPR-NET-A. More details on this as well as training implementation details for our proposed method can be found in supplementary material.

5. Experimental Results

To evaluate the proposed method, we perform extensive experiments on RAF-DB and C-EXPR-DB. The Unweighted Average Recall (UAR) and F1 score are the performance metrics for RAF-DB and C-EXPR-DB, respectively. Details regarding RAF-DB and pre-processing performed, as well as definitions of the performance evaluation metrics, can be found in the supplementary material.

Comparison with State-of-the-Art for CER on RAF-DB At first, we train C-EXPR-NET on RAF-DB for Compound Expression Recognition (CER) and compare its performance to the state-of-the-art. Table 3 shows that C-EXPR-NET outperforms by: i) 7% the best performing method FaceBehaviorNet (pretrained with millions of images from 10 databases for multi-task learning of valence-arousal, basic expressions, AUs and fine-tuned on RAF-DB for CER); ii) 9.2% the ReCNN (pretrained on million faces for face recognition and then trained on RAF-DB for CER); iii) 10.7% the DLP-CNN + mSVM (pretrained for BER on RAF-DB and then fine-tuned on RAF-DB for CER); iv) 12.1% the ResNet-18 + separate loss.

Domain Adaptation Experiment on RAF-DB for CER The proposed C-EXPR-DB is currently the largest in-the-wild database annotated for compound expressions. Thus we train C-EXPR-NET on this database and then use it in a domain adaptation context. In more detail, we consider C-EXPR-NET (pretrained on C-EXPR-DB) as a prior and fine-tune it on RAF-DB for CER. We compare its performance to the state-of-the-art. Table 3 shows that C-EXPR-NET when pretrained on C-EXPR-DB outperforms by: i) 4.8% the same method that is not pretrained on C-EXPR-DB, but is directly trained on RAF-DB; ii) 11.8% the FaceBehaviorNet. Our method yields state-of-the-art results, illustrating the strength of using C-EXPR-DB for CER.

Comparison with State-of-the-Art for BER on RAF-DB Since C-EXPR-NET targets multi-label CER (rather than multi-class CER), it can also solve the basic expression recognition problem (BER). Thus we utilize the C-EXPR-NET that is directly trained on RAF-DB for CER and test its performance on RAF-DB for BER. We then compare its performance to the state-of-the-art. Table 3 shows that C-EXPR-NET outperforms by: i) 5% the best performing method ResNet-18 (ARM) (pretrained for BER on AffectNet and then fine-tuned on RAF-DB for BER); ii) 5.9% the DACL (pretrained on millions of faces for face recognition and then trained on RAF-DB for BER including data augmentation); iii) 6% the PSR; iv) 7% the ReCNN (pretrained on million faces for face recognition and trained on RAF-DB for BER); v) 8.2% the VGGFACE + augmented (trained on a combination of artificially generated images and real images of RAF-DB for BER); vi) 8.4% the FaceBehaviorNet; vii) 9.1% the ResNet-18 + separate loss; viii) 12% the DLP-CNN + mSVM. When C-EXPR-NET is pretrained on C-EXPR-DB and then fine-tuned on RAF-DB for CER, its performance for BER is 2.3% higher than the performance of the same method that is not pretrained on C-EXPR-DB, but is directly trained on RAF-DB. All these verify that C-EXPR-NET trained for CER indirectly learns as well to distinguish between the basic expressions.

Zero-Shot Learning on RAF-DB Here we train C-EXPR-NET without the data augmentation component on RAF-DB for CER. We test its performance for BER in a zero-shot setting. The fact that C-EXPR-NET targets CER as a multi-label problem (rather than multi-class one) results in C-EXPR-NET being able to effectively generalize its knowledge in new emotion recognition contexts, in a zero-shot manner. Table 3 shows that zero-shot C-EXPR-NET: i) outperforms some methods that have been trained for 6 basic expression recognition (VGG + mSVM and baseD-CNN + mSVM); ii) has a comparable performance to the other state-of-the-art methods (between 1.8% and 8.8%), although all methods have been trained for BER and the total number of images used in their training was an order

of magnitute bigger than the total number of images used in zero-shot C-EXPR-NET's training.

Table 3. Performance comparison between the state-of-the-art and C-EXPR-NET for 11-compound and 6-basic expression recognition on RAF-DB; performance metric is UAR

RAF-DB	Compound	Basic
zero-shot C-EXPR-NET	0.517	*0.714*
C-EXPR-NET	**0.553**	**0.852**
C-EXPR-NET pretrained on C-EXPR-DB	**0.601**	**0.875**
FaceBehaviorNet [11, 12]	-	0.768
fine-tuned FaceBehaviorNet	*0.483*	-
VGG + mSVM [15]	0.316	0.579
baseDCNN + mSVM [15]	0.402	0.706
DLP-CNN + mSVM [15]	0.446	0.732
ResNet-18 + separate loss [16]	0.432	0.759
ReCNN [29]	0.461	0.782
VGG-FACE + augmented [10]	-	0.770
ResNet-18 (ARM) [25]	0.471	*0.802*
PSR [27]	0.465	0.792
DACL [8]	0.466	0.793

Ablation Study for each proposed Component on RAF-DB & C-EXPR-DB At first, we keep only the Backbone Network and Expression Block of C-EXPR-NET, which predicts the compound expressions when fed with images annotated in terms of compound expressions. We utilize three widely used DNNs to act as these components, namely ResNet50, VGG16 and DenseNet121. We compare their performance when CER is formulated as a: i) multi-class (MC) problem, in which the compound expression output classes are mutually-exclusive; the output layer has 12 units, softmax activation and the loss function is a categorical cross entropy one; ii) multi-label (ML) problem, in which the compound expression output classes are not mutually-exclusive; the output layer has 6 units, sigmoid activation and the loss function is binary cross entropy (CE); iii) multi-label problem, in which the output layer has 6 units, softmax activation and the loss function is KL-divergence.

Table 4 (rows 1-9) shows that the case when CER is formulated as multi-label problem utilizing softmax and KL-divergence provided the best results on both tested databases, RAF-DB and C-EXPR-DB. It provided the best performance regardless of which network was used, making our methodology model agnostic. Comparing the 3 aforementioned DNNs, ResNet50 achieved the best performance consistently among all settings and databases; thus we decided to use ResNet50 in our proposed method.

The above results verify our intuition that formulating CER as a multi-label problem is more realistic and makes it easier for the model to decompose and recognise expressions more accurately as it takes into account the information that two classes are correlated. Compound means that the expression is constructed as a combination of two basic expressions. "Happily surprised" & "happily disgusted" expressions involve facial muscles typically used in the production of expressions of happiness and surprise & happiness and disgust, respectively; for both compound expressions, the basic expression "happy" is involved. Thus an expression recognition system identifies some similar patterns on the two compound expressions, that confuse the system if the problem is formulated as multi-class.

Next, we keep only the Backbone Network and Expression Block of the proposed method and feed it with both compound expression annotated images and basic expression annotated images from RAF-DB. Table 4 (rows 9-10) illustrates that CER performance is enhanced by 4% and 2% on RAF-DB and C-EXPR-DB, respectively, when the extra basic expression annotated images are fed to the model as a form of data augmentation. This data augmentation is plausible due to formulating CER as a multi-label problem.

Next, we target multi-task (MT) learning by incorporating an AU detection task (AU-D) in the above setting, with the aim to further increase CER performance. Thus we add the AU branch to the previous model. Table 4 (rows 10-11) indicates that CER performance increases by 5% and 7% on RAF-DB and C-EXPR-DB, respectively. This result is the first ever proof that AU-D can act as an auxiliary task to enhance CER performance, regardless whether the AU annotations are manual (C-EXPR-DB) or automatic (RAF-DB). Additionally we train a single-task (ST) model (ResNet) only with the Backbone Network and AU Branch for AU-D. We compare its AU-D performance with that of the MT model described in the previous paragraph. Table 5 (rows 1-2) shows that the single-task model's performance is higher by 2% and 1% on RAF-DB and C-EXPR-DB, respectively. The fact that MT model's performance for AU-D is worse than that of ST model indicates that negative transfer occurs, because CER dominates the training process.

Consequently, we add the Distribution Matching module in the MT model, forming C-EXPR-NET, and compare its performance to that of the ST model for AU-D. Table 5 (rows 2-3) shows that C-EXPR-NET (denoted as C-EXPR-NET-V since the visual modality is utilized as input to the method) outperforms the single-task model by 2% & 3% on RAF-DB and C-EXPR-DB, respectively. We also notice that C-EXPR-NET's performance for CER is further increased by 3% and 5% on RAF-DB and C-EXPR-DB, respectively (shown on row 12 of Table 4).

The above presented results refer to cases where the visual modality was used; images were provided as input to the models. Let us mention that the proposed methodology is modality agnostic, meaning that it works for any type of modality (not only the visual one). Since C-EXPR-DB is an A/V database, it further contains the audio modality. Therefore we utilize the audio modality in the form of spectrograms and provide them as input to our method. In the case

of A/V C-EXPR-DB, we extract spectrograms from the audio modality and use them as input to C-EXPR-NET; the resulting model is denoted C-EXPR-NET-A in Tables 4 and 5. Alternatively, instead of extracting spectrograms from the audio modality, we could use the raw signal-waveform as input to our method. In that case we would have to choose appropriate feature extractors as Backbone Network, Expression Block and AU Block. Finally, we perform late fusion on the expression and AU logits of C-EXPR-NET-V and C-EXPR-NET-A. Tables 4 and 5 illustrates that the late fusion method enhances both CER and AU-D performance by 3% and 5% from the visual-only (C-EXPR-NET-V) and audio-only (C-EXPR-NET-A) models.

Table 4. Ablation Study: Performance comparison for CER; in parenthesis is the problem formulation: multi-class (MC); multi-label with sigmoid cross entropy (ML-CE); multi-label with softmax KL-div (ML-KL); ExprB/AUD: Expression/AU Branches; DM/DA: Distribution Matching/Data Augmentation modules; V/A: visual/audio modalities given as input to system; LF: late fusion; metrics: UAR for RAF-DB, F1 for C-EXPR-DB

Methods	Components				Databases	
	ExprB	AUB	DM	DA	RAF-DB	C-EXPR-DB
VGG (MC)	✓				0.29	0.35
DenseNet (MC)	✓				0.33	0.36
ResNet (MC)	✓				0.35	0.39
VGG (ML-CE)	✓				0.32	0.36
DenseNet (ML-CE)	✓				0.35	0.37
ResNet (ML-CE)	✓				0.37	0.40
VGG (ML-KL)	✓				0.38	0.41
DenseNet (ML-KL)	✓				0.40	0.41
ResNet (ML-KL)	✓				0.43	0.46
ResNet (ML-KL)	✓			✓	0.47	0.48
ResNet (ML-KL)	✓	✓		✓	0.52	0.55
C-EXPR-NET-V	✓	✓	✓	✓	0.55	0.60
C-EXPR-NET-A	✓	✓	✓	✓	-	0.58
C-EXPR-NET-LF	✓	✓	✓	✓	-	0.63

Table 5. Performance comparison for AU-D; ML-KL: multi-label with softmax & KL-div; ExprB/AUD: Expression/AU Branches; DM/DA: Distribution Matching/Data Augmentation modules; V/A: visual/audio modalities; LF: late fusion; metric: F1

Methods	Components				Databases	
	ExprB	AUB	DM	DA	RAF-DB	C-EXPR-DB
ResNet		✓			0.58	0.52
ResNet (ML-KL)	✓	✓		✓	0.56	0.51
C-EXPR-NET-V	✓	✓	✓	✓	0.60	0.55
C-EXPR-NET-A	✓	✓	✓	✓	-	0.53
C-EXPR-NET-LF	✓	✓	✓	✓	-	0.58

Ablation Study on AU Detection Performance on RAF-DB and C-EXPR-DB Here we perform ablation experiments to show and verify the value of incorporating AU semantic descriptors as auxiliary information for AU detection, via the AU Semantic Encoding and Cross-Modality Attention parts. At first we train a vanilla ResNet-50 for AU detection (vanilla means that only images are provided as input to the network) and compare its performance to that of a ResNet-50 that includes AU Semantic Encoding and Cross-Modality Attention parts (i.e., the AU Branch of the proposed methodology) and thus takes as input both images and 17 AU semantic descriptors. Table 6 illustrates that the latter network outperformed the vanilla ResNet by 2% and 3% on RAF-DB and C-EXPR-DB, respectively.

Finally, we aim to show that incorporating AU semantic descriptors as auxiliary information for AU detection brings similar and significant performance gains, regardless of which backbone network is used. Previously we showed that this is the case when ResNet-50 is used. We further utilize VGG16 and DenseNet121 as backbone networks. Table 6 illustrates that the performance gain when VGG is used is 2% and 3% for RAF-DB and C-EXPR-DB, respectively. Table 6 also shows that the performance gain when DenseNet is used is 3% and 3% for RAF-DB and C-EXPR-DB, respectively. These results validate that incorporating AU semantic descriptors as auxiliary information to the system enhances its performance for AU detection.

Table 6. Ablation Study: Performance comparison for AU detection when the AU semantic descriptors are vs are not provided as auxiliary information; performance metric: F1 for both databases

Methods	Databases	
	RAF-DB	C-EXPR-DB
vanilla ResNet	0.56	0.49
ResNet with AU Semantic Encoding and Cross-Modality Attention	0.58	0.52
vanilla VGG	0.50	0.45
VGG with AU Semantic Encoding and Cross-Modality Attention	0.52	0.48
vanilla DenseNet	0.54	0.46
DenseNet with AU Semantic Encoding and Cross-Modality Attention	0.57	0.49

6. Conclusions

Limited research has been conducted for CER, because there exist only two, small in-the-wild databases. In this paper, we introduce C-EXPR-DB, the largest A/V in-the-wild database annotated in terms of compound expressions, which can foster further research on the area. We also propose C-EXPR-NET, a novel methodology for CER. We perform extensive experiments that illustrate that the database is needed and that the proposed method outperforms the state-of-the-art both for basic and compound expression recognition. The limitations of this work are the fact that C-EXPR-DB is imbalanced and the fact that there are not many other in-the-wild databases annotated for compound expressions that can be used to further evaluate our method.

References

[1] C.F. Benitez-Quiroz, R. Srinivasan, and A.M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR'16)*, Las Vegas, NV, USA, June 2016.

[2] Zijun Cui, Tengfei Song, Yuru Wang, and Qiang Ji. Knowledge augmented deep neural networks for joint facial expression and action unit recognition. *Advances in Neural Information Processing Systems*, 33, 2020.

[3] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 828–835. IEEE Computer Society, 2020.

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

[5] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *MultiMedia, IEEE*, 19(3):34–41, 2012.

[6] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[7] Paul Ekman. Facial action coding system (facs). *A human face*, 2002.

[8] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2402–2411, 2021.

[9] Jeffrey M Girard, Wen-Sheng Chu, László A Jeni, and Jeffrey F Cohn. Sayette group formation task (gft) spontaneous facial expression database. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 581–588. IEEE, 2017.

[10] Dimitrios Kollias, Shiyang Cheng, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, 128(5):1455–1484, 2020.

[11] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.

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

[13] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, BjÃ¶rn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, feb 2019.

[14] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Bjorn Schuller, Kam Star, et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *arXiv preprint arXiv:1901.02839*, 2019.

[15] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017.

[16] Yingjian Li, Yao Lu, Jinxing Li, and Guangming Lu. Separate loss for basic and compound facial expression recognition in the wild. In *Asian Conference on Machine Learning*, pages 897–911. PMLR, 2019.

[17] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[18] Iiris Lüsi, Julio CS Jacques Junior, Jelena Gorbova, Xavier Baró, Sergio Escalera, Hasan Demirel, Juri Allik, Cagri Ozcinar, and Gholamreza Anbarjafari. Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 809–813. IEEE, 2017.

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

[20] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, 4(2):151–160, 2013.

[21] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 881–888, 2013.

[22] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[23] Fabien Ringeval, Andreas Sonderegger, Jens Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[24] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 50–58, 2015.

[25] Jiawei Shi, Songhao Zhu, and Zhiwei Liang. Learning to amend facial expression representation via de-albino and affinity. *arXiv preprint arXiv:2103.10189*, 2021.

[26] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65, 2010.

[27] Thanh-Hung Vo, Guee-Sang Lee, Hyung-Jeong Yang, and Soo-Hyung Kim. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, 8:131988–132001, 2020.

[28] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[29] Yifan Xia, Hui Yu, Xiao Wang, Muwei Jian, and Fei-Yue Wang. Relation-aware facial expression recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[30] Huiyuan Yang, Lijun Yin, Yi Zhou, and Jiuxiang Gu. Exploiting semantic embedding and visual feature for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491, 2021.

[31] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017.

[32] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[33] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016.