Large-Scale Facial Expression Recognition Using Dual-Domain Affect Fusion for Noisy Labels

Dexter Neo, Tsuhan Chen, Stefan Winkler
School of Computing
National University of Singapore
e0534450@u.nus.edu, {winkler,tsuhan}@nus.edu.sg

Abstract

Building models for human facial expression recognition (FER) is made difficult by subjective, ambiguous and noisy annotations. This is especially true when assigning a single emotion class label to facial expressions for large in-the-wild FER datasets. Human facial expressions often contain a mixture of different mental states, which exacerbates the problem of single labels when used to categorize emotions. Dimensional models of affect – such as those using valence and arousal – provide significant advantages over categorical models in terms of representing human emotional states but have remained relatively under-explored. In this paper, we propose an approach for dual-domain affect fusion which investigates the relationships between discrete emotion classes and their continuous representations. In order to address the underlying uncertainty of the labels, we formulate a set of mixed labels via a dual-domain label fusion module to exploit these intrinsic relationships. Finally, we show the benefits of the proposed approach using AffectNet, Aff-Wild, and MorphSet, in the presence of natural and synthetic noise.

1. Introduction

Reading emotions from facial expressions is a vital step towards building machines that can better understand human behaviour and interact appropriately. In recent years, automatic facial expression recognition (FER) has made tremendous progress and become an important problem in machine learning and computer vision applications.

Affective facial expressions can be represented using various different approaches:

- Facial Action Coding System (FACS) [6], where facial actions and muscle movements are described as a combination of different Action Units (AUs). The FACS model however does not directly describe the affective state.

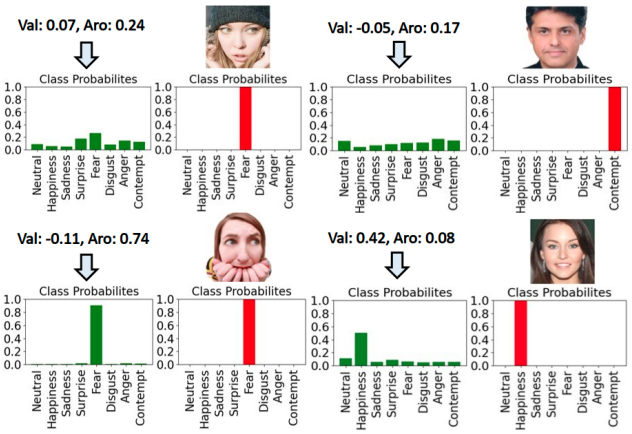

Figure 1. The nature of facial expressions is complex, making it difficult to categorize expressions into single emotion categories (red - ground truth). We propose to derive these variability from the dimensional affect space into soft mixed labels (green - proposed).

- Categorical models [5], where the facial expression is commonly represented as one of six basic emotions (happiness, surprise, sadness, anger, disgust, fear) or neutral.

- Dimensional models as for example [27] that use a continuous multi-variate space, such as valence (pleasantness of the stimulus) and arousal (intensity for the emotion caused by the stimulus).

Despite the different models of affect, most FER systems are based on the categorical approach, which have yet to achieve the capabilities necessary for building meaningful human-machine interactions. This is due to the challenge of categorizing facial expressions into a single basic emotion: humans rarely display high-intensity prototypical facial expressions, whereas mixed emotional states are far more common in everyday situations. This causes the categorical model to suffer greatly from both intra-class vari-

ations and inter-class similarities in the presence of unreliable labels. These unreliable labels do not provide proper feedback during training and often impede the learning of discriminative features, causing many supervised learning algorithms to fail. This problem is especially prevalent in large scale in-the-wild FER datasets, where noisy, ambiguous, poor-quality images and low annotator agreement are common [32]. This makes annotating facial expressions with a single class label very difficult.

In contrast, the dimensional model of affect [27] provides clear advantages over the categorical approach when it comes to representing the full spectrum of human affective states. Other similar models of affect include adding a third dimension of dominance (degree of control caused by stimulus) [21]. Alternative formulations include emotion distribution learning where the probabilities for each emotion class are modelled, thus allowing discrete labels to better represent the ambiguities of facial expressions [36]. These methods include using multi-label classification [15] or multi-class classification [17] to represent the different blends of emotion classes present in the underlying image. Unfortunately, manually annotating a large amount of in-the-wild facial images with multiple labels is expensive and labour-intensive.

In this work, we further build on the fact that ambiguous facial expressions contain mixtures of the different basic emotions [17, 26], and that these uncertainties can be directly derived from the dimensional affect model. Our dual-domain affect fusion accounts for both types of affect representation, by explicitly considering the interpretations of emotions in both image and label space.

Our contributions can be summarized as follows:

- We bridge the gap between dimensional and categorical affect by developing a novel dual-domain label fusion method, which models uncertainty in FER classification tasks by exploiting the underlying relationships with valence/arousal values.
- We propose two strategies (Mixture of Gaussians and Euclidean distance based) to derive such relationships and compare the two.
- We show that our framework provides lower calibration error and easily complements other calibration strategies.
- Finally, we validate our framework on a synthetic laboratory dataset (MorphSet) as well as two of the largest in-the-wild FER datasets (AffectNet and Aff-Wild). Our approach achieves state-of-the-art performance on AffectNet for both dimensional and categorical tracks.

2. Related Work

2.1. Facial Expression Recognition

Traditionally, FER systems comprise three different phases namely, facial detection, feature extraction, and finally regression or classification. Current state-of-the-art methods usually involve deep learning, where a convolutional neural network (CNN) is used to automatically extract deep features and trained in an end-to-end fashion. This includes methods such as network ensembles with shared representations [29], which hierarchically combine abstract features such as lines, edges and colors with local features such as nose, mouth and eyes. CNNs are sensitized to the size of the input images, as in-the-wild images typically vary in size; [30] used super-resolution to alleviate this issue. [12] applied 3D morphable models to affect synthesis on neutral images to the desired target affect. [33] developed region attention networks (RAN) for facial images with occlusions and varying pose. For more details on other methods used in categorical FER tasks, please refer to recent survey papers [18, 36].

2.2. Learning with Uncertainties

Ambiguous facial expressions often cause uncertainty amongst annotators. Traditionally FER datasets were collected procedurally in controlled laboratory environments, where participants were asked to categorize photographs of acted expressions into basic emotions [7, 16, 25]. This would allow expressions to be represented as a distribution rather than a single emotion, as different annotators may feel differently about each photograph, and even the same annotator may have multiple mixed emotions to each photograph [26]. However, as FER datasets grow larger, this procedural method of collecting annotations becomes very expensive.

Popular in-the-wild FER datasets such as AffectNet [22] and Aff-Wild2 [14] typically only provide one-hot labels for categorical affect. However, in-the-wild images often contain a mixture of different emotions, and one-hot encodings are insufficient to capture the underlying diverse emotions of these images. Furthermore, the subjectivity of emotions makes it difficult to annotate certain facial expressions, as some emotions are highly similar to each other. For example, expressions of "Contempt" and "Anger" are visually indistinguishable and should not be hard-assigned to a single emotion class.

Regularization techniques such as uniform label smoothing have been shown to help the model become less overconfident by regarding each incorrect class as equally probable [23]. However, some classes in FER are intrinsically closer than others. For example, the correct class label "Fear" may be mistaken for "Surprise" more often than for "Disgust". Clearly, the smoothing between the incor-

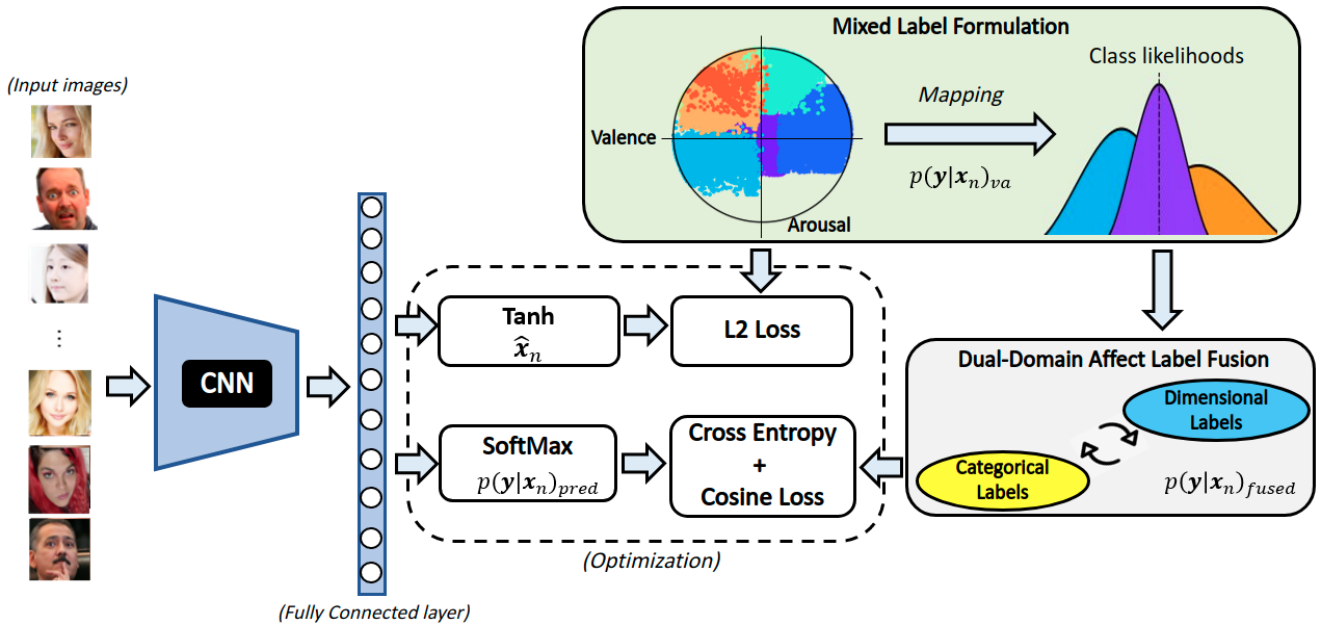

Figure 2. Learning representations for both dimensional and categorical affect are shared through an encoder, class likelihoods are derived from the dimensional space and dual-domain label fusion is performed on-the-fly to introduce uncertainties into the categorical labels.

rect classes of FER is non-uniform and should be reliably represented as a probability distribution [19, 26]. For example, [37] proposed Emotion Distribution Learning to map a face image to an emotion probability distribution. [30] considered the prior distribution of each class in the training set to smooth the labels. [32] used Self Cure Networks to suppress the impact of uncertain samples, thus preventing overfitting on uncertain images. [3] proposed Auxiliary Label Space graphs to leverage topological information from labels.

3. Proposed Methodology

In contrast to prior works, we investigate the underlying relationships between dimensional and categorical affect. As shown in Figure 2, our proposed dual-domain affect fusion framework directly formulates mixed-class likelihoods from the dimensional valence/arousal (VA) space and performs label fusion in order to obtain probabilities that best represent each facial expression. Our framework jointly considers the categorical and dimensional models of affect and comprises the following two modules, which are described in more detail below:

1. *Mixed Label Formulation* extracted from the dimensional affect space using either Mixture of Gaussians or Euclidean distance-based methods.
2. *Dual-Domain Affect Label Fusion* combines dimensional labels with categorical labels for FER classification tasks.

3.1. Mixed Label Formulation

3.1.1 Gaussian Mixture Labels

A probabilistic approach to deriving a set of soft mixed labels for facial expression classification would be to model the dimensional affect space with a Gaussian Mixture Model (GMM). The GMM is a linear superposition of different Gaussians which create highly complex probability density functions that can be used to approximate any continuous density with arbitrary accuracy.

We use the notations as per [2] and formulate the GMM for dimensional affect as follows:

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1}$$

where $\boldsymbol{x_n}$ represents each VA point in dimensional affect space, $\mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ describes each Gaussian and $p(\boldsymbol{x_n})$ denotes the marginal distributions/normalizer. The sets of parameters used to define the properties of each Gaussian component in the GMM are represented with $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$. Specifically π_k represents the prior distribution, $\boldsymbol{\mu}_k$ represents the centroids and $\boldsymbol{\Sigma}_k$ represents the covariance matrices for each of the k classes or components respectively. These sets of parameters can be computed directly from the VA annotations present in the dimensional affect space. A set of conditional probabilities $p(\boldsymbol{y}|\boldsymbol{x}_n)_{GMM}$ (mixed labels) can be obtained using Bayes' Theorem:

$$p(\boldsymbol{y}|\boldsymbol{x}_n)_{GMM} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \quad (2)$$

The GMM for the training set of AffectNet is visualized in Figure 3, where each of the k classes belongs to a different Gaussian in the GMM. As we traverse along different areas of the GMM, changes for each point $\boldsymbol{x}_n$ in the dimensional affect space would also provide different conditional class probabilities when transitioning from one Gaussian component to another.

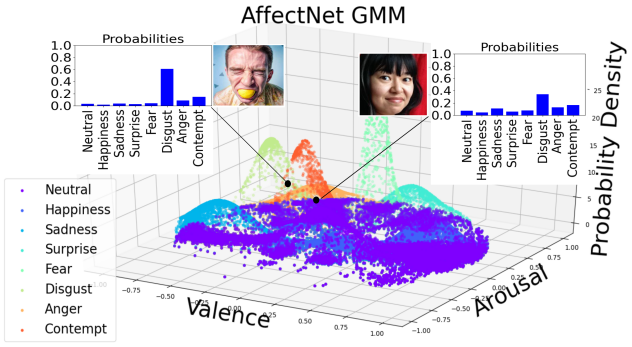

Figure 3. GMM formulation for AffectNet, showing different conditional class probabilities corresponding to different points in the dimensional affect space.

3.1.2 Euclidean Mixture Labels

Apart from the probabilistic approaches, another intuitive method would be to directly rank classes that are semantically closer with higher probabilities and reduce the probabilities of classes that are further away. This would allow the model to also consider the next most likely classes. Inspired by the work on soft ordinal labels [4], we propose a distance based method to use the relationships of VA annotations and map these values into a probability distribution across k classes.

We compute the probabilities $p(\boldsymbol{y}|\boldsymbol{x}_n)_{Euclid}$ using the Euclidean distances from the centroids $\boldsymbol{\mu}_k$ to rank each VA point in the training set:

$$p(\boldsymbol{y}|\boldsymbol{x}_n)_{Euclid} = \frac{\max ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||_2^2}{||\boldsymbol{x}_n - \boldsymbol{\mu}_k||_2^2} \quad (3)$$

This allows the dimensional relationships between the emotion classes to be mapped into a probability distribution. These distributions are then further normalized into valid likelihoods such that $\sum^{K} p(\boldsymbol{y}|\boldsymbol{x}_n)_{Euclid} = 1$. For instance, data points which are closer to a certain centroid $\boldsymbol{\mu}_k$ should be mapped to that class.

Since certain emotion classes or facial representations may be intrinsically closer than others, our proposed formulations for mixed labels are non-uniform and do not regard the incorrect classes with equal probabilities, unlike label smoothing.

3.2. Dual-Domain Affect Label Fusion

Due to the ambiguity of facial expressions in certain images, some VA values may have been incorrectly annotated, leading to a loss in training accuracy when solely using mixed labels derived from the dimensional affect space. To counteract this problem, we propose to fuse the ground truth one-hot labels with the distributions obtained from the dimensional affect space.

Prior research [13, 34] has shown that facial expressions share both the categorical and dimensional affect space. Performing label fusion would further allow both models of affect to complement one another. We propose that the degree of fusion between the two distributions is coupled by a hyperparameter α:

$$p(\boldsymbol{y}|\boldsymbol{x}_n)_{fused} = \alpha p(\boldsymbol{y}|\boldsymbol{x}_n)_{GT} + (1-\alpha) p(\boldsymbol{y}|\boldsymbol{x}_n)_{VA} \quad (4)$$

where $p(\boldsymbol{y}|\boldsymbol{x}_n)_{fused}$ represents the resultant fused categorical distributions, $p(\boldsymbol{y}|\boldsymbol{x}_n)_{VA}$ are either the Euclidean or GMM distributions obtained directly from the dimensional space, and $p(\boldsymbol{y}|\boldsymbol{x}_n)_{GT}$ denotes the one-hot encoded vector representing the ground truth class label. An example of our label fusion method is further illustrated in Figure 4. The fused result is a soft mixed label which contains a combination of different probabilities derived from the ground truth label and VA annotations.

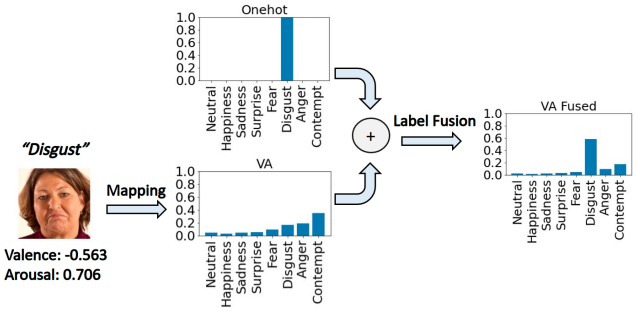

Figure 4. Dual-domain fusion accounts for labels from both categorical and dimensional affect spaces. The distributions obtained from dimensional affect may not always coincide with the true class; fusion is proposed to preserve training accuracy.

3.3. Loss Functions

In order to tackle the class-imbalances commonly found in FER tasks, we adopt the same weighted cross entropy loss as per [22]. The class weights W_k can be computed using:

$$W_k = \frac{N_{\max}}{N_k}, \quad (5)$$

where $N_{\max}$ represents the number of training samples in the majority class. In the event of an evenly balanced class distribution, the class weights W_k are equal to one.

The weighted cross entropy loss between the softmax predictions and fused labels $\boldsymbol{y}$ can be computed as:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{n}^{N} W_k \boldsymbol{y} \log \frac{\exp(\boldsymbol{\theta}^T f_n)}{\sum_j^K \exp(\boldsymbol{\theta}_j^T f_n)} \tag{6}$$

where $\boldsymbol{\theta}$ represents the weights and f_n represents the input features to the fully connected layer.

We follow [1] and further maximise the cosine similarity (sim) between softmax predictions and fused labels:

$$\mathcal{L}_{CS} = W_k(1 - sim(\boldsymbol{y}, \frac{\exp(\boldsymbol{\theta}^T f_n)}{\sum_j^K \exp(\boldsymbol{\theta}_j^T f_n)})) \tag{7}$$

For affect regression, we minimize the Euclidean distance between the valence/arousal predictions and annotations. Since the training samples for each class are imbalanced, we also apply the same set of weights W_k during training.

$$\mathcal{L}_2 = \frac{1}{N}\sum_{n}^{N} W_k ||\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n||_2^2 \tag{8}$$

In order to jointly learn both dimensional and categorical affect, the final objective function is given by:

$$\mathcal{L} = \mathcal{L}_2 + \mathcal{L}_{CE} + \mathcal{L}_{CS} \tag{9}$$

4. Experiments and Results

4.1. Datasets

AffectNet [22] is the largest image database for in-the-wild facial expression recognition. It comprises 450,000 manually annotated images collected from the Internet. The dataset provides VA annotations and eight class labels for categorical affect (the "Contempt" class is added to the typical seven basic emotions). We randomly split a subset of training images for validation while selecting hyperparameters, thereafter we retrain the model using the entire training set. As the test set is not released, we follow the evaluation protocol proposed by the authors in order to keep comparisons fair. We report the performance of our method on the validation set, which contains 4,000 images for categorical affect and 4,500 for dimensional affect.

AffWild [35] is the largest video database for in-the-wild facial expression recognition which contains VA annotations. To derive class labels for a corresponding VA label and frame, we borrow the parameters $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ computed from the training set of AffectNet and assign categorical labels from the dimensional space. Aff-Wild contains 252 videos for training and 46 videos for testing. This translates to a total of roughly 1M+ frames with valence/arousal annotations and class expressions. We performed a random 80-20 training/test split on the Aff-Wild training videos for our experiments. We were unable to use the given Aff-Wild test set since the official test set does not contain any VA annotations. We also did not use Aff-Wild2 [14] since the VA labels do not intersect with the class labels.

MorphSet [31] is a synthetically augmented dataset generated from a collection of laboratory datasets [7, 16, 25]. It comprises seven different categorical class labels and a relatively balanced distribution between classes. The dataset provides multiple expressions per identity and comes with highly consistent VA annotations. We randomly split the dataset based on the different subjects into roughly 63,000 images for training and 16,000 images for testing.

4.2. Implementation Details

For our experiments we use Resnet-18 CNN [8] as the backbone architecture. The input images are resized to 224×224 pixels, along with typical data augmentation strategies such as colour jitter, random horizontal flips and random affine transformations, which would allow the model to regularize better to unseen samples. The parameters $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ from dimensional affect were computed within one standard deviation.

We use a batch size of 256 and an initial learning rate of 2.5e-4 with a decay factor of 0.5 every 30 epochs. The fusion hyper-parameter α is empirically set to 0.5. Effects of the fusion hyper-parameter will be discussed further in the Ablation Studies. The network is optimized using the Adam optimizer [11] and Stochastic Weighted Averaging (SWA) [10] from the Pytorch library. All experiments are manually seeded for reproducibility and trained on a 8GB NVIDIA GeForce RTX 2070 GPU for 50 epochs. The code is made available at *https://github.com/dexterdley/RC-AffectNet* .

4.3. AffectNet Benchmarking

4.3.1 Categorical Affect Classification

We compare our results for 8 category classification against the state-of-the art methods in Table 1. For fair comparisons, we only include methods which do not use additional training data. We find that learning with dual-domain affect label fusion greatly improves the performance for categorical affect classification. This is largely due to the ability of mixed VA labels to encapsulate information about the other classes from the dimensional affect space.

4.3.2 Dimensional Affect Regression

Predicting values for dimensional affect is commonly regarded as a regression problem. Due to the nature of highly skewed training samples, we find an improvement in overall

Algorithm 1: The Dual-Domain Affect Fusion Framework

Data: Training Set D;

Initialize model parameters θ
for *each mini-batch* $N \in D$ **do**
  for *each sample* $n \in \{1, ..., N\}$ **do**
  return $p(\boldsymbol{y}|\boldsymbol{x}_n)_{va}$ − `getMixedLabels()`
  return $p(\boldsymbol{y}|\boldsymbol{x}_n)_{fused}$ − `LabelFusion()`
  Compute $\mathcal{L}_{CE} + \mathcal{L}_{CS}$ from $p(\boldsymbol{y}|\boldsymbol{x}_n)_{fused}$
  Compute $\mathcal{L}_2$ from $\boldsymbol{x}_n$
  Update θ by gradient descent
return θ

Function `LabelFusion()`**:**
  $\alpha p(\boldsymbol{y}|\boldsymbol{x}_n)_{GT} + (1-\alpha)p(\boldsymbol{y}|\boldsymbol{x}_n)_{va}$
  return $p(\boldsymbol{y}|\boldsymbol{x}_n)_{fused}$

Function `getMixedLabels()`**:**
  return $p(\boldsymbol{y}|\boldsymbol{x}_n)_{va}$

Method	Accuracy	F1
AffectNet [22]	58.00	58.00
Wide Ensemble [29]	59.30	-
Auxillary Label [3]	59.35	-
RAN [33]	59.50	-
SCN [32]	60.23	-
Pyramid super res. [30]	60.68	-
Deep 3DMM [12]	60.00	59.00
Weighted Cluster [24]	60.70	60.49
Ours	**61.93**	**61.93**

Table 1. Classification performance of different methods on AffectNet (8 categories).

regression errors and correlation values when adding class weights W_k to the weighted regression loss as formulated in Equation 8.

	RMSE		CCC	
Method	Valence	Arousal	Valence	Arousal
Mobile CNN [9]	0.41	0.37	-	-
AffectNet [22]	0.37	0.41	0.60	0.34
CAAE [20]	0.45	0.41	0.49	0.41
Wide Ensemble [29]	0.36	0.33	-	-
Deep 3DMM [12]	0.37	0.38	0.62	0.54
Ours	**0.34**	**0.33**	**0.66**	**0.55**

Table 2. Regression performance for valence and arousal predictions on AffectNet. RMSE values closer to zero are better, whereas CCC values closer to one are better.

We compare our proposed method against the current state-of-the-art methods in Table 2, by achieving RMSE of **0.34** and **0.33** for valence and arousal respectively. We also obtain higher concordance correlation coefficients (CCC) for both valence and arousal predictions.

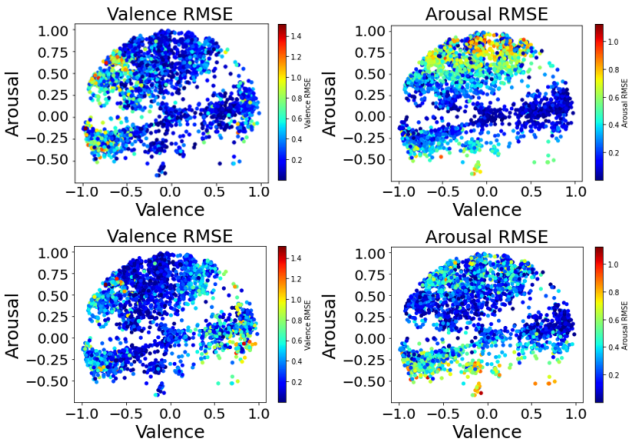

Figure 5. Error plots for valence and arousal using L2 loss (top) with RMSE of 0.427 & 0.39 vs. weighted L2 loss (bottom) with RMSE of 0.33 & 0.30.

These improvements in the performance are further illustrated using Figure 5. The overall root mean squared error (RMSE) for both valence and arousal scatter plots is generally lower when using weighted L2 loss as compared to using regular L2 loss. The error reduction is most noticeable in the top left quadrant in the 2D space, where minority classes such as "Contempt" and "Fear" reside. This coincides with the intention of applying weights to cause the network to pay more attention to samples that belong to the minority classes.

5. Analysis and Discussion

In this section, we conduct in-house experiments on AffectNet, Aff-Wild, and MorphSet. Firstly, we evaluate the robustness of our proposed method under synthetic label noise. Secondly, we show the effects of the fusion hyperparameter α. Lastly, we visualize the features learnt by the network trained using different methodologies.

5.1. Evaluation under Synthetic Label Noise

We evaluate our proposed method under different settings of categorical label noise, whereby a percentage of the training labels (10%, 20% or 30%) are randomly flipped to other classes. For the Aff-Wild experiments, we assign $p(\boldsymbol{y}|\boldsymbol{x}_n)_{GMM}$ to each corresponding VA annotation. The one-hot formulations are assigned to the argmax of $p(\boldsymbol{y}|\boldsymbol{x}_n)_{GMM}$. For these experiments the fusion hyperparameter α is set to 0.5

In Table 3, we compare the typical one-hot label formulation against both the Euclidean and GMM fusion formulations and find that label fusion from the dimensional affect space is especially helpful for FER classification tasks even under the injection of synthetic categorical label noise. In the case where label noise is not added, the improvements of label fusion are marginal ∼0.49%. The benefits of label fusion become more pronounced when label noise is increased to 30%, outperforming the baseline one-hot formulation by up to 14.9%, 0.46% and 5.06% on AffectNet, Aff-Wild, and MorphSet respectively. This suggests that dual-domain label fusion helps the network generalize to unseen samples on both in-the-wild and laboratory datasets by preventing it from making over-confident predictions.

	Noise (%)	One-hot	Fusion (Euclidean)	Fusion (GMM)
AffectNet	0	0.6017	0.6066	0.5945
	10	0.5842	0.5943	0.5920
	20	0.5193	0.5890	0.5748
	30	0.4336	0.5822	0.5746
Aff-Wild	0	0.2227	0.2235	0.2325
	10	0.2130	0.2322	0.2174
	20	0.2184	0.2322	0.2305
	30	0.2221	0.2242	0.2267
MorphSet	0	0.8999	0.8912	0.8922
	10	0.8939	0.8649	0.9030
	20	0.8709	0.8514	0.8866
	30	0.8377	0.8224	0.8883

Table 3. F1 scores for different methods on AffectNet, Aff-Wild, and MorphSet subjected to different values of synthetic categorical label noise.

5.2. Effects of Fusion Hyperparameter α

The differences between one-hot labels and fused labels can be further evaluated by comparing the effects of the hyperparameter α in our label fusion module. The hyperparameter α directly influences the degree of fusion between the dimensional and categorical labels.

As shown in Table 4, the best performing value of α is found to be 0.5, where both the highest classification scores and lowest regression errors are observed on AffectNet and MorphSet. Notably, solely using the mixed labels from the dimensional affect space would result in poorer performance as compared to directly using the categorical labels. However, the decrease in performance is only marginal ∼1%, suggesting that mixed labels drawn from the dimensional space are enough for classification without the need for additional class annotations.

When label fusion is performed, both accuracy and F1 are improved by 1-2% when $\alpha = 0.5$. For the regression tasks on AffectNet, the fused labels only help to reduce the overall regression errors slightly, and they do not play a big part in further lowering the RMSE values for valence or arousal. Surprisingly, the regression errors on MorphSet are significantly lower when label fusion is performed, by a margin of 0.011 and 0.016 for valence and arousal respectively. We suspect the differences in improvements to be due to the fact that AffectNet is collected in-the-wild and generally contains noisier annotations than MorphSet, which is much more controlled.

	Fusion	Regression		Classification	
	α	RMSE V	RMSE A	Accuracy	F1
AffectNet	0.0	0.344	0.301	0.5715	0.5643
	0.3	0.337	0.298	0.5958	0.5950
	0.5	**0.333**	0.302	**0.6048**	**0.6066**
	0.7	0.337	**0.296**	0.6046	0.6040
	1.0	0.337	0.302	0.5845	0.5843
MorphSet	0.0	0.063	0.069	0.9414	0.9417
	0.3	0.063	0.067	0.9500	0.9503
	0.5	**0.060**	**0.065**	**0.9557**	**0.9561**
	0.7	0.063	0.072	0.9457	0.9464
	1.0	0.071	0.081	0.9511	0.9511

Table 4. Model performance for different values of fusion hyperparameter α on AffectNet and MorphSet

5.2.1 Qualitative Analysis

We show qualitative examples using GradCAM [28] for each of the different methods across each dataset in Figure 6. GradCAM produces a visual heatmap which highlights the important features used for predictions by utilizing the gradients propagating through the final convolutional layer. Each row in Figure 6 shows a randomly sampled image from AffectNet, Aff-Wild, and MorphSet.

The corresponding columns show features learnt by the network trained with different labels, followed by training with additional label noise. We observe that the models trained using one-hot labels have less consistent heatmaps across different datasets. When additional synthetic label noise is injected into the training set, we observe that the learnt heatmaps of the one-hot variant to fluctuate significantly as compared to the Euclidean or GMM fusion variant. The additional information drawn from the dimensional space directly helps the network to reduce overconfidence, thereafter improving generalization performance to unseen samples. Features learnt using our proposed fusion mainly focus on the important regions of the face and remain relatively consistent even under the influence of additional label noise.

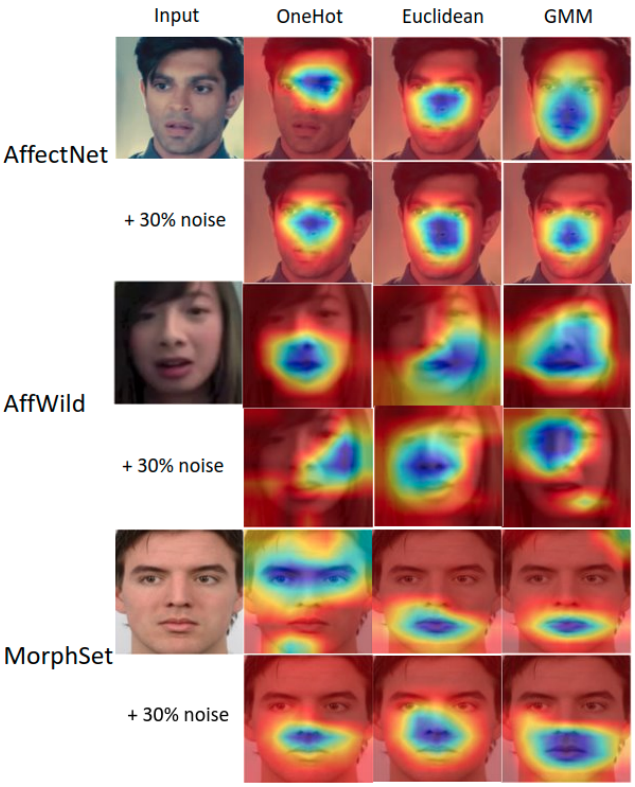

Figure 6. Comparisons of the features learnt using different fusion methods under GradCAM analysis.

6. Conclusions

Facial expressions are complex, and the typical one-hot encoding may not fully represent the different underlying emotions. We presented a novel method to derive mixed labels via a dual-domain affect fusion method which combines dimensional affect with categorical labels. We proposed two methods to derive mixed labels from dimensional space and show the benefits of dual-domain affect fusion as well as the features learnt using different variants of fusion. Our method outperforms the current state of the art, and experiments yield robust performance against synthetic label noise.

References

[1] Bjorn Barz and Joachim Denzler. Deep learning on small datasets without pre-training using cosine loss. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 5

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 3

[3] Shikai Chen, Jianfeng Wang, Yuedong Chen, Zhongchao Shi, Xin Geng, and Yong Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13981–13990, 2020. 3, 6

[4] Raúl Díaz and A. Marathe. Soft labels for ordinal regression. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4733–4742, 2019. 4

[5] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971. 1

[6] P. Ekman and W. V. Friesen. *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press, 1978. 1

[7] Ellen Goeleven, Rudi De Raedt, Lemke Leyman, and Bruno Verschuere. The Karolinska directed emotional faces: A validation study. *Cognition and Emotion*, 22(6):1094–1118, 2008. 2, 5

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 5

[9] C. Hewitt and H. Gunes. CNN-based facial affect analysis on mobile devices. *arXiv*, abs/1807.08775, 2018. 6

[10] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv:1803.05407*, 2018. 5

[11] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, Dec. 2014. 5

[12] 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. 2, 6

[13] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv 1910.11111*, 2020. 4

[14] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv:1910.04855*, 2019. 2, 5

[15] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[16] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, Skyler T. Hawk, and Ad van Knippenberg. Presentation and validation of the Radboud faces database. *Cognition and Emotion*, 24(8):1377–1388, 2010. 2, 5

[17] Shan Li and Weihong Deng. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 127:884–906, June 2019. 2

[18] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215, 2022. 2

[19] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017. 3

[20] Alexandra Lindt, Pablo Barros, Henrique Siqueira, and Stefan Wermter. Facial expression editing with continuous emotion labels. In *Proc. 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)*, pages 1–8, May 2019. 6

[21] A Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14:261–292, 1996. 2

[22] A. Mollahosseini, B. Hasani, and M. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. 2, 4, 5, 6

[23] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 4696–4705, 2019. 2

[24] Tan Quan Ngo and Seokhoon Yoon. Facial expression recognition based on weighted-cluster loss and deep transfer learning using a highly imbalanced dataset. *Sensors*, 20(9), 2020. 6

[25] Michal Olszanowski, Grzegorz Pochwatko, Krzysztof Kuklinski, Michal Scibor-Rylski, Peter Lewinski, and Rafal K. Ohme. Warsaw set of emotional facial expression pictures: A validation study of facial display photographs. *Frontiers in Psychology*, 5:1516, 2015. 2, 5

[26] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2, 3

[27] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. 1, 2

[28] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 10 2019. 7

[29] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient facial feature learning with wide ensemble-based convolutional neural networks. In *Proc. 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, Feb 2020. 2, 6

[30] T. Vo, G. Lee, H. Yang, and S. Kim. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, 8:131988–132001, 2020. 2, 3, 6

[31] Vassilios Vonikakis, Dexter Neo, and Stefan Winkler. Morphset: Augmenting categorical emotion datasets with dimensional affect labels using face morphing. In *Proc. IEEE International Conference on Image Processing (ICIP)*, 2021. 5

[32] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3, 6

[33] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29(1), 2020. 2, 6

[34] Xiaohua Wang, Muzi Peng, Lijuan Pan, Min Hu, Chunhua Jin, and Fuji Ren. Two-level attention with two-stage multi-task learning for facial emotion recognition. *Journal of Visual Communication and Image Representation*, 62:217–225, 2019. 4

[35] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild' challenge. In *Proc. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1980–1987. IEEE, 2017. 5

[36] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Björn Schuller, and Kurt Keutzer. Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6729–6751, Oct. 2022. 2

[37] Ying Zhou, Hui Xue, and Xin Geng. Emotion distribution recognition from facial expressions. In *Proc. 23rd ACM International Conference on Multimedia*, page 1247–1250, 2015. 3