Frame Level Emotion Guided Dynamic Facial Expression Recognition with Emotion Grouping

Bokyeung Lee Hyunuk Shin Bonhwa Ku Hanseok Ko
Department of Electrical Engineering, Korea University
Seoul, South Korea
{bksain, hushin, hush999, hsko}@korea.ac.kr

Abstract

Facial expression recognition (FER) has received considerable attention in computer vision, with "in-the-wild" environments such as human-computer interaction and video understanding. Recognizing dynamic facial expressions in videos is generally considered a more practical and reliable approach than still images. However, the dynamic FER problem in videos has challenges in terms of both data acquisition and the structural aspects of the learning model. In particular, video frames that deviate from the target facial expression class can significantly degrade the performance of dynamic FER. In this paper, we present an affectivity extraction network (AEN) for dynamic FER. AEN combines features of different semantic levels and classifies both sentiment and specific emotion categories with emotion grouping. To address the challenges of dynamic FER, we propose frame-level emotion-guided loss functions and a structural aspect of the learning model. The AEN has two branches: a bottom-up branch that learns facial expressions representation at different semantic levels and outputs pseudo labels of facial expressions for each frame using a 2D FER model, and a top-down branch that learns discriminative representations by combining feature vectors of each semantic level for recognizing facial expressions at the corresponding emotion group. Additionally, the proposed frame-level emotion-guided loss functions encourage AEN to prevent the loss of emotional information and retain the emotional probability of a video clip. Experimental results on various video datasets show that the proposed AEN consistently outperforms the state-of-the-art in Ekman and sentiment FER. Representative results demonstrate the promise of the proposed AEN for dynamic FER in the video.

1. Introduction

Facial expression recognition (FER) is a high-level computer vision task that classifies emotion class from images or videos. Facial emotional information is essential in next-generation computer vision such as human-computer interaction and video understanding, [1, 19, 41, 48], etc. FER in static/still images and "in-the-lab" environment have been actively studied and shown good results [12, 21, 22, 24–26, 29, 35, 40, 43, 46, 52, 58, 59, 65–69, 71, 72]. However, existing methods showed limitations from a generalization point of view. So, researchers in emotion recognition consider recognizing dynamic facial expressions in video rather than static/still images as a more practical or reliable approach. To improve the generalization performance of FER, video-based FER in an "in-the-wild" environment is increasing attention [20, 23, 27, 33, 34, 47, 61].

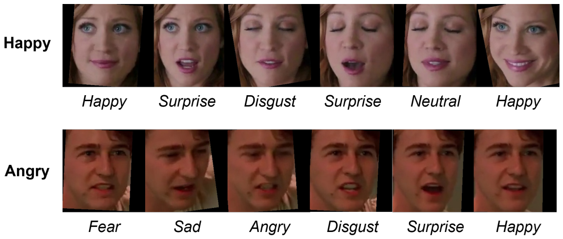

Figure 1. The examples of dynamic facial expression dataset. The emotion label to the left of the frames corresponds to the entire clip. The emotion label located at the bottom of each frame is the estimation results of the 2D FER model trained with the RAF dataset. In general, the emotions shown in the video are inconsistent.

Several datasets, such as CAER, AFEW, DEFW, and FERV39K, have been released to address the problem of dynamic FER [6, 17, 31, 60]. Dynamic FER suffers from the presence of irrelevant frames to the target emotion of the video clip as well as occlusion and non-frontal pose. As shown in Figure 1, a *happy* video clip may contain another facial expression such as *Disgust* that is not related to happiness. Furthermore, the same video clip may contain multiple facial expressions with different emotions that are slightly related to the target emotion label because of facial changes resulting from conversation or eye blinking. To tackle these challenges, various approaches have been proposed [10,39,44,73], including 3D convolutional neural networks (CNNs), recurrent neural networks, and temporal transformers. These methods leverage the temporal characteristics of facial expressions to capture the sequence of emotional changes in videos. However, since the methods compress spatio-temporal data with a single emotion label of a video clip, a loss of emotional information occurs, and this may degrade the performance of emotion recognition.

Recently, psychology-based emotion recognition models have been proposed in various modalities such as image, audio, and text [29, 30, 35, 37, 42, 63]. These models consider the fact that emotions share similar characteristics and use a hierarchical emotion grouping approach, which represents fine-grained emotions (*angry, disgust, fear, happy, sad,* and *surprise*) as a few basic categories (*positive* and *negative*). Especially, hierarchical emotion grouping-based models are attracting attention because they lead to avoiding false alarm errors as well as improving detection performance. Inspired by this, our paper presents a dynamic FER that is based on a hierarchical emotion grouping approach while reducing a loss of emotion information in the feature extraction process of facial expression.

This paper presents an affectivity extraction network (AEN) that combines features of different semantic levels for a hierarchical emotion grouping approach. AEN consists of two branches with 2D CNN, temporal transformers, semantic-to-affective converters (S2ACs), and classifiers. The bottom-up branch learns facial expressions at the different semantic levels and outputs probabilities for a facial expression class for each frame using a 2D FER model. Feature maps extracted from convolution layers in CNN have different semantic levels, and the feature map of higher semantic levels is extracted at deeper layers. While feature maps at lower semantic levels are spatially fine but semantically weak, those at higher semantic levels are spatially coarse but semantically strong [38]. According to the method proposed by FPN [38], a fusion of feature maps at low-level and high-level is an essential factor in detecting small objects. If the concept of a small object in FPN is viewed from emotion recognition, it can be replaced with the granularity of emotions. That is, in a two-level hierarchical emotion grouping model, a high affective level increases the granularity of emotions, and in order to recognize the fine-grained emotion categories well, the combining of a low-level semantic feature map with a high-level semantic feature map is essential. The top-down branch learns discriminative feature representation by combining feature vectors of each semantic level and a high semantic level for recognizing facial expressions at the corresponding affective level. To generate effectively combined feature vectors, we introduce an attention-based semantic-to-affective converter. To reduce the loss of emotional information in AEN, we propose two frame-level emotion-guided loss functions guided by the emotional probabilities of each frame. The frame-level emotion-guided loss functions consist of a temporal affectivity extraction loss and a global affectivity extraction loss. The temporal affectivity extraction loss function allows the temporal transformer to maintain emotional feature representation corresponding to the target emotion while compressing the changes in facial expression. The global affectivity extraction loss function aims that the emotional probability of each affective level follows that of each semantic level. The proposed two loss functions allow AEN to understand what emotions are included in the video clip.

2. Related Work

2.1. Dynamic Facial Expression Recognition in Videos

The recent advancements in deep learning and the arrangement of dynamic facial expression recognition challenges [5] have paved the way for the development of spatio-temporal deep network algorithms to understand facial expression sequences. Prior studies, such as [7, 32, 45, 49, 62], utilized 2D CNN and temporal-based deep networks, such as recurrent neural network (RNN), Long Short-Term Memory (LSTM), or Gated Recurrent Unit (GRU), to learn spatial and temporal information. To jointly extract spatio-temporal features from video, researchers have adopted 3D CNN, which can capture temporal correlation, in studies such as [11,28,57,70]. Additionally, some studies, such as [32, 49], used audio signals in conjunction with facial information and designed multi-modal network architectures to generate discriminative features.

More recently, transformer-based methods have been proposed for dynamic facial expression recognition due to their powerful learning ability in various computer vision tasks. Former-DFER [73] was the first study to apply the temporal transformer for dynamic FER and enable the network to learn contextual information. This study serves as the baseline method for our proposed method. STT [47] designed a spatio-temporal transformer to capture discriminative features by utilizing spatial attention and temporal attention and employing a compact cross-entropy loss function, which trains the close intra-class correlation and the large inter-class distance. NR-DFERNet [34] presented a noise-robust dynamic FER network to suppress interference of target irrelevant frames at the temporal stage using self-attention. DPCNet [61] proposed a dual path multi-excitation collaborative network to produce a practical spatio-temporal relation from limited input frames, and it shows that only the crucial regions of input frames and keyframes determine accurate expressions of video sequences. GCA+IAL [33] designed an intensity-aware loss

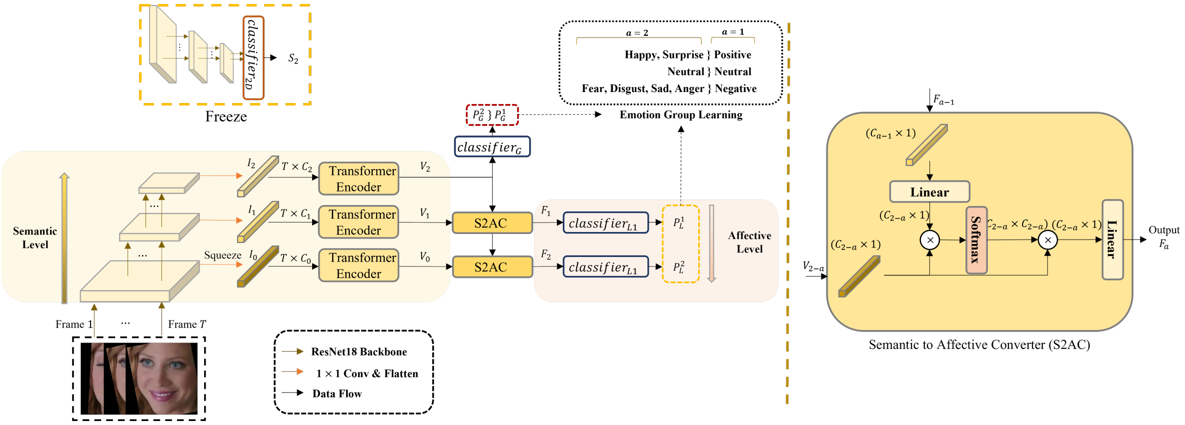

Figure 2. An overview of AEN and its sub-network architecture. The proposed AEN is based on two-level emotion hierarchical assumption. The backbone network (ResNet18) is pretrained with the FER dataset for a single image. The two pre-trained networks are employed for training AEN, one is the backbone network, which is finetuned with dynamic facial expression video, and the other one provides emotion probability $S_2 \in \mathbb{R}^{\mathrm{T} \times \mathrm{K}}$ of each frame without finetune. K_2 denotes the number of fine-grained emotion labels. s and a are semantic and affective levels, respectively.

function inspired by the fact that all non-neutral expressions tend to approach neutral expressions when the emotional intensity converges to zero. This loss function maximizes the probability of the target class while minimizing the largest logit excluding the target class, and the trained network can understand the intensity of expression.

2.2. Emotion Group Learning

Several methods to adopt an easy-to-difficult strategy in emotion recognition are inspired by the cognitive model of human beings. The approach first judges coarsely the categories of emotions, and then determines finely the categories of emotions. Facial expressions often require similar muscle movements causing small muscle contractions [9]. For instance, anger and sadness produce highly similar changes in the eyes and mouth, while disgust often changes the corner of the mouth slightly. KTN [35] grouped and redefined coarse labels (*positive, negative, neutral,* and *surprise*) based on basic Ekman emotions (*angry, disgust, fear, happy, sad, surprise*, and *neutral*) using the coarse-to-fine labels strategy. They proposed fine-stream, which focuses on directly learning the original fine label information of facial expressions, and coarse-stream, which obtains the coarse label information. The authors of KTN then leverage knowledge distillation using information from the coarse-stream to improve the representation power of fine-stream. [64] proposed coarse-to-fine cascaded network to address the label ambiguity problem in facial expression recognition in video, which consists of the coarse-net and the negative-net. The negative-net focus on classifying four negative expressions and coarse-net predicts other emotion classes. They can capture both universal and unique features of each emotion using smooth predicting. In visual emotion recognition, MDAN [63] defined a multi-level emotion hierarchy and grouped fine-grained emotions according to [50]. MDAN also includes a local classifier at each semantic level of the top-down branch, with each local classifier focusing on learning the discrimination among emotions at a particular emotion level. Emotion group learning allows networks to learn important expression information and has been demonstrated to be effective.

3. Proposed Method

In this section, we describe the AEN architecture with emotion grouping and two loss functions guided by the emotional probabilities of each frame. The AEN consists of a bottom-up branch for providing features of different semantic levels and a top-down branch for generating discriminative representations at each affective level. To ensure that the proposed model structure is suitable for dynamic FER, we present two frame-level emotion-guided loss functions and emotion group learning.

3.1. Affective Extraction Networks

As shown in Figure 2, AEN consists of two hierarchical branches: a bottom-up and a top-down branch. The bottom-up branch generates spatio-temporal feature representations of different semantic levels and a top-down branch provides discriminative features based on the affective levels by combining feature vectors of each semantic level and a high semantic level. We employ ResNet18 [15] as a backbone network of the bottom-up branch to extract static information. Facial images are input to ResNet18 to output feature maps of different semantic levels. The semantic levels consist of three stages such as low, middle, and high layers. The feature maps at each semantic level are transformed to

$I_s \in \mathbb{R}^{\mathrm{T}\times \mathrm{C_s}}$, $s = 0, 1, 2$, using squeeze operation, which consists of point-wise convolution and flattening. The transformed feature maps are forwarded into the temporal transformer of each semantic level to capture temporal correlation. The output of temporal transformers at each semantic level, $V_s \in \mathbb{R}^{\mathrm{C_s}\times 1}$, $s = 0, 1, 2$, denotes spatio-temporal feature vectors representing each semantic level. We also employ another ResNet18 network pretrained with a facial expression recognition dataset of 2D images to make probabilities of T input facial images. The probabilities are used as a pseudo label in section 3.3.

A top-down branch consists of two S2AC and three classifiers, and V_s is forwarded to the S2AC. The high semantic information V_2 is forwarded to the global classifier. To compensate for semantic information of semantically low feature, V_{s-1} is fed to a linear layer, and the number of dimensions is expanded to fit V_s as shown in the right of Figure. 2. We discover all element-wise dependencies between features at the neighboring affective level. For affective level a, S2AC is formulated as:

$$F_a = W_2 \rho\left(\frac{W_1 F_{a-1} V_{2-a}^{\mathsf{T}}}{\sqrt{C_{2-a}}}\right) V_{2-a}, a = 1, 2 \qquad (1)$$

where W_1 and W_2 are weights of linear layer and ρ denotes row-wise softmax function. F_0 is set to V_2. A fusion of low-level features with high-level semantic information increases the granularity of emotions. Through S2AC, we convert feature representation to be interpretable at the affective level. Since affective level of F_2 is higher than F_1, F_2 is useful to determine specific emotion categories. Combined features F_1 and F_2 are forwarded to each local classifier, respectively.

3.2. Emotion Group Learning

Emotion group learning encourages AEN to mimic the cognitive mode of human beings and to learn important expression information. As shown in Figure 2, we grouped emotion categories of $a = 2$ into emotion classes of $a = 1$. For example, we grouped *happy* and *surprise* categories with $a = 2$ into ”positive” with $a = 1$. In order for the proposed AEN to be used effectively, we need a multi-class loss function that reflects the hierarchical emotion group learning. Recall that [63] trains deep networks with a multi-class cross-entropy loss to learn global and local discrimination. We define a multi-class cross-entropy loss function that reflects the predictions of the global emotion classifier and the local emotion classifier for each affective level. Unlike [63] which recognizes visual emotion in a single image, the aim of our emotion group learning is to find common representations of facial regions in videos. The multi-class loss function for emotion group learning can be formulated as

$$L_{mc} = -\sum_{a=1}^{H_a} \sum_{k=1}^{|K_a|} Y_k^a P_{O,k}^a, \qquad (2)$$

where H_a means the number of affective levels. $|K_i|$ is the number of emotion groups at affective level a, and Y_k^a denotes the ground truth value of the emotion class belonging to the affective level. P_O^a refers to the overall prediction at each affective level and is defined as

$$P_O^a = \alpha \times \rho(P_L^a) + (1 - \alpha) \times \rho(P_G^a), \qquad (3)$$

where P_L^a and P_G^a are the outputs of the local classifiers and the global classifier, respectively. α is a fusion parameter, which controls the relative importance between P_L and P_G. Then, the global prediction for group j at $a-1$, $\rho(P_{G,j}^{a-1})$ is acquired by summing the global probability of all sub-categories k at a of group j at $a-1$, $\rho(P_{G,j}^{a-1}) = \sum_{k\in j} \rho(P_{G,k}^a)$. In Equation 3, the multi-class loss is simply the cross entropy between a one-hot distribution Y_k^a and estimated probability $\rho(P_{O,k}^a)$. By minimizing L_{mc}, AEN is simultaneously optimized to learn discriminative feature representations at each affective level, i.e., an image is classified in *positive* at $a = 1$ and is classified in *happy* at $a = 2$.

3.3. Frame Level Emotion Guided Loss Function

In dynamic FER, there are several problems, which cause performance degradation. The temporal transformer plays the role of converting spatio-temporal information into the discriminant feature vector. However, it is difficult for the temporal transformer to convert feature maps at each semantic level into discriminant feature vectors without frame-level guide information since compressed semantic information is used as input. Also, if the video input data contains frames different from the emotion of the video clip, the temporal transformer cannot guarantee the acquisition of the discriminant feature vector. That is, if only the cross-entropy loss function is used in model training, there is a limit to extracting discriminant features for emotion recognition because cross-entropy does not consider the ambiguity that comes from video data actually containing multiple emotions. To address this issue, we propose frame-level emotion-guided loss functions induced by the emotional probabilities of each frame, which consists of a temporal affectivity extraction loss and a global affectivity extraction loss. The pre-trained 2D FER model is used not only as a backbone network in AEN but also as a guide network for a pseudo-label generation. In training process, we acquire the emotional probabilities $S_2 \in \mathbb{R}^{\mathrm{T}\times \mathrm{K_2}}$ and use one-hot encoded target $Y^{H_a} \in \mathbb{R}^{\mathrm{K_2}\times 1}$.

The temporal affectivity extraction loss allows transformer encoders to reduce the loss of information related to the target emotion of the video clip. The temporal affectivity extraction loss can be formulated as

$$L_{ta} = \sum_{i=0}^{2} ||V_i - I_i^{\intercal} S_2 Y^{H_a})||_1^1, \tag{4}$$

where $S_2 Y^{H_a} \in \mathbb{R}^{\mathrm{T}\times 1}$ denotes the probability of emotion of each frame related to the ground truth emotion of the video clip and is considered as the importance weights of the input of the temporal transformer. L_{ta} is defined the difference between V_s and the weighted sum of I_s. By minimizing L_{ta}, the transformer encoder directly learns to ignore irrelevant frames while highlighting frames corresponding to the emotion of the video clip. Then, the output of the temporal transformer V_s has the information with dominant emotion as well as the correlation between frames.

To apply a global affectivity extraction loss function, we independently translate each output of local-classifiers P_L^a into the range 0-1 using the sigmoid(σ) function. We can interpret the output of sigmoid as the emotional probability of the input video. We aim that the emotional probability of each affective level in AEN follows that of each semantic level. So, we add a global affectivity extraction loss that encourages AEN to predict emotional distribution at each local classifier as follows:

$$L_{ga} = \sum_{a=1}^{H_a} \beta_i ||\sigma(P_L^a) - \rho(S_a^{\intercal} \mathbf{1})||_2^2, \tag{5}$$

where β_i is weights of loss corresponding to affective level, $\mathbf{1} \in \mathbb{R}^{\mathrm{T}\times 1}$ denotes one-vector that all elements are one, and frame-level emotional probabilities at each affective level can be represented $S_{j,a-1} = \sum_{k \in j} S_{k,a}$ according to emotion grouping. P_L^a is the output of local classifier at affective level a, and it is transformed between 0 and 1 using sigmoid function σ. $\rho(S_a^{\intercal}\mathbf{1})$ means the mean of frame-level probabilities for the emotion class in each affective level. L_{ga} is the difference between the output of local classifier and pseudo probabilities generated by frame-level emotional probabilities.

AEN is trained with the following total loss function:

$$L = L_{mc} + \lambda_1 L_{ta} + \lambda_2 L_{ga}, \tag{6}$$

where λ_1 and λ_2 are regularization parameters. By optimizing our proposed loss function, AEN is trained like human cognitive models and learns contextual information.

4. Experiments

4.1. Datasets

DFEW: The DFEW [17] consists of over 16,000 video clips from more than 1500 movies, such as tragedies, comedies and romantic, etc. These video clips contain natural facial expressions and then is a significantly challenging dataset because of the unconstrained conditions, illumination, and occlusions. All samples on DFEW have been split into five same-size parts without overlap. So, we implement five-fold cross-validation, which takes one part of the samples for the testing set and the others for the training set.

AFEW: AFEW [6] dataset served as an evaluation platform for the EmotiW challenge from 2013 to 2019. This dataset contains about 1800 video clips collected from TV programs and movies, so AFEW is very close to real-world data. All samples on AFEW have been split into three subsets: training, validation, and testing set. We train AEN on the training set and evaluate results on the validation set as the previous methods did.

FERV39K: The FERV39K [60] dataset is proposed recently, which is the current largest benchmark for dynamic FER in the wild. This dataset contains over 38000 video clips collected from several scenarios, which can be partitioned into various scenes (i.e., daily life, business, and school). All samples on FERV39K have been split into two subsets: training, and testing sets without overlapping.

4.2. Metrics

We use the unweighted average recall (UAR) and the weighted average recall (WAR) as the evaluation metrics. UAR and WAR are generally considered important metrics in almost dynamic FER research including baseline [17,73]. UAR is an unweighted average recall and denotes the accuracy per class divided by the number of classes without consideration of instances per class. WAR is weighted average recall and means general accuracy. We evaluate accuracy for a 3-emotion and 7-emotion category because the aim of our proposed method is not only to classify fine-grained emotions but also to classify coarse labels, and this is an important metric that demonstrates the robustness of a model [63, 64]. The accuracy for a 3-emotion category is calculated by considering it as the correct answer when selecting the emotion class of the same group as the label in the 7 specific classes. High accuracy for 3 emotions denotes that the model reduces the hierarchy violation.

4.3. Implementation Details

Preprocessing For DFEW and FERV39K datasets, the video frames' face region is publicly available, then we train and evaluate the processed data directly. For the AFEW dataset, the FFmpeg toolkit [53] is used to extract frames from the raw videos. The face region of the video frame is detected using the Dlib toolbox [18]. All the face images are resized to 112×112 pixels to input AEN.

Training and Inference Details We train AEN with the Pytorch platform with an NVIDIA RTX 3090 GPU. We first pre-train the backbone network with RAF dataset [36] to ex-

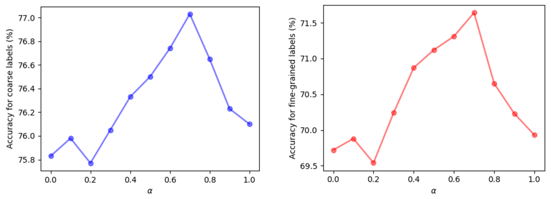

Figure 3. Effect of fusing parameter α for P_G and P_G on 3 and 7 emotion classification WAR. α is the weight of P_L

Method	L_{cr}	L_{mc}	L_{ta}	L_{ga}	WAR (%)
Former-DFER [73]	✓	×	×	×	66.57
DPCNet [61]	✓	×	×	×	65.78
AEN	✓	×	×	×	68.10
AEN	×	✓	×	×	67.97
AEN	×	×	✓	×	70.23
AEN	×	×	×	✓	69.76
AEN	×	✓	✓	×	71.13
AEN	×	✓	✓	✓	71.64

Table 1. Ablation study for the proposed loss functions on DFEW fold1. L_{cr} denotes that AEN is trained with loss function that considers $a = 2$.

tract emotion probability unlike STT [47], which pretrained backbone network with MS-Celeb-1M [13]. The settings for training and testing our model are the same as for the baseline method [73] So, we can acquire $T = 16$ facial frames as the input. The highest affective level H_a is set 2. We empirically set $\alpha = 0.7$. β_0 and β_1 are 1.4 and 0.6, respectively. The regularization parameters λ_1 and λ_2 are 0.4 and 0.5, respectively.

4.4. Ablation Study

In this section, we show the impact both of AEN and proposed loss functions will be verified, and then the performance differences according to the emotion groups and hyper-parameter are presented. All the experiments are conducted on the DFEW dataset with fold1 as one of five cross-validation sets.

Evaluation of α. Figure 3 shows the WAR variation at coarse and fine-grained labels (3 and 7 emotion class) when α ranges from 0.0 to 1.0. As α increases from 0 to 0.7, we can see a trend of improved performance. We acquire the best performances for 3 and 7 emotions at $\alpha = 0.7$ and then the accuracy decreases rapidly for both 3 and 7 emotion labels. This result proves the effect of local classifiers at different affective levels and implies that the balance between global and local classifiers should be adjusted appropriately.

Evaluation of different loss functions. Table 1 demonstrates that our proposed loss functions are effective for dynamic FER. L_{cr} denotes the general cross-entropy loss function for seven emotion classes in the above experiment.

Emotion grouping			WAR (%)
Positive	Neutral	Negative	
Happy, Anger	Neutral, Sad	Fear, Disgust, Surprise	67.79
Happy	Neutral, Surprise	Fear, Disgust, Sad, Anger	71.23
Happy, Surprise	Neutral	Fear, Disgust, Sad, Anger	71.64

Table 2. Evaluation of AEN with different emotion groups on DFEW fold1. The first row is a completely misgrouped case. The second and third rows are categories that are generally considered in psychology.

In the fourth row, solely optimizing multi-class loss function L_{mc} for emotion group learning degrades the performance of AEN. In contrast, the emotion group learning strategy succeeded by solving problems of dynamic FER using L_{ta} and L_{ga} in the 7th and 8th rows. The results imply that utilizing frame-level emotion-guided loss produces more discriminative features by reducing the loss of emotional information. Then, this maximizes the effect of emotion group learning.

Evaluation of different emotion groups We experimented with three categorical emotion groups: (1) completely mixed group (first row), (2) sentiment group considered at [4] (second row), (3) emotion grouping of Ekman emotions [8] (third row) as shown in Table 2. AEN with a group (1) leads to significant performance drop and instability during training. The difference between group (2) and group (3) is a small difference in whether the *surprise* class is included in *neutral* or *positive* in the parent category, but using group (3) outperforms AEN trained with a group (2) for emotion group learning strategy. Therefore, we use group (3) in all experiments. This result indicates that the proper emotion group learning strategy encourages AEN to learn highly discriminative feature representation. In addition, we can see that *surprise* has many characteristics more similar to *happy* than *neutral* in dynamic facial expression.

4.5. Comparison with State-of-the-art Methods

Our proposed method is compared with the state-of-the-art on three datasets. The comparison methods can be divided into 3D and 2D CNN-based methods. The best performance is marked in bold.

Our baseline Former-DFER shows lower performance than the latest methods as shown in Table 3. In contrast, our AEN produces the best results for UAR and WAR in 3 and 7 emotion classes. Specifically, GCA-IAL [33] is the state-of-the-art method with the UAR 55.71% and the WAR of 69.24% for 7 emotions, and DPCNet [61] has the highest accuracy among the previous methods with the UAR 71.58% and the WAR 71.95% for 3 emotions. Our AEN outperforms GCA-IAL by 0.95% and 0.13% in terms of the UAR and the WAR for 7 emotions, respectively. Moreover, AEN obtains better results with respect to the UAR and the WAR of 3 emotions compared with DPCNet by 3.02% and

Method	Accuracy of Emotions (%)							3 Emotions (%)		7 Emotions (%)	
	Happiness	Sadness	Neutral	Anger	Surprise	Disgust	Fear	UAR	WAR	UAR	WAR
3D Resnet18 [14]	76.32	50.21	64.18	62.85	47.52	0.00	24.56	-	-	46.52	58.27
Resnet18+LSTM [15, 16]	83.56	61.56	68.27	65.29	51.26	0.00	29.34	-	-	51.32	63.85
Resnet18+GRU [3, 15]	82.87	63.83	65.06	68.51	52.00	0.86	30.14	-	-	51.68	64.02
Former-DFER [73]	84.05	62.57	67.52	70.03	56.43	3.45	31.78	69.96	70.57	53.69	65.70
STT [47]	87.36	67.90	64.97	71.24	53.10	3.49	34.04	-	-	54.58	66.65
NR-DFERNet-v1 [34]	88.47	64.84	70.03	75.09	61.60	0.00	19.43	-	-	54.21	68.19
NR-DFERNet-v2 [34]	86.42	65.10	70.40	72.88	50.10	0.00	45.44	-	-	55.77	68.01
DPCNet [61]	89.93	64.61	67.12	63.18	53.67	15.86	31.56	71.58	71.95	55.13	66.32
GCA+IAL [33]	87.95	67.21	70.10	76.06	62.22	0.00	26.44	-	-	55.71	69.24
AEN	89.24	69.38	70.67	72.08	59.07	4.17	32.00	**74.60**	**74.96**	**56.66**	**69.37**

Table 3. Comparison with state-of-the-art methods on DFEW. Five-fold cross-validation is implemented.

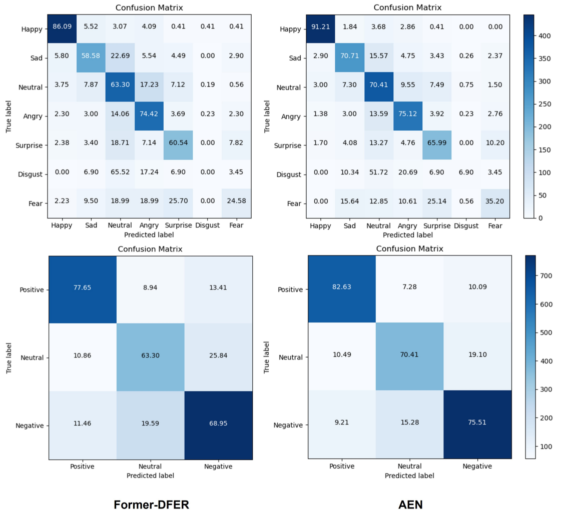

Figure 4. The confusion matrices of our baseline (Former-DFER) and our proposed AEN were evaluated on DFEW fold1. The top and bottom figures are the confusion matrices for fine-grained and coarse emotion labels, respectively.

Method	3 Emotions (%)		7 Emotions (%)	
	UAR	WAR	UAR	WAR
EmotiW-2019 Baseline [5]	-	-	-	38.81
C3D [54]	-	-	43.75	46.72
I3D-RGB [2]	-	-	41.86	45.41
R(2+1)D [55]	-	-	42.89	46.19
3D ResNet18 [14]	-	-	42.14	45.67
ResNet18+LSTM [15, 16]	-	-	43.96	48.82
ResNet18+GRU [3, 15]	-	-	43.75	46.72
Former-DFER [73]	63.94	63.66	47.42	50.92
STT [47]	66.45	67.03	49.11	54.23
NR-DFERNet [34]	-	-	48.37	53.54
DPCNet [61]	57.47	64.06	47.86	51.67
AEN	**67.49**	**67.38**	**50.88**	**54.64**

Table 4. Comparison with state-of-the-art methods on AFEW.

3.01%, respectively.

Figure 4 shows confusion matrices for fine-grained and coarse labels. The left and right figures represent the confusion matrices of DFER-Former and AEN, respectively. In DFER-Former on the DFEW dataset, a phenomenon occurs where the prediction is concentrated in the neutral class. Our AEN seems to solve the above problem somewhat. Especially, confusion matrices for the coarse emotion classes show AEN has fewer hierarchy violation cases compared with Former-DFER.

As shown in Table 4, we evaluate our AEN on the AFEW dataset. STT [47] is the state-of-the-art method with the UAR and the WAR for 3 and 7 emotions. Our proposed AEN achieves the best results both in UAR and WAR for the 3 and 7 emotions class. The proposed method outperforms STT by 1.77% and 0.41% with respect to the UAR and WAR for 7 emotions, respectively. AEN also produces better results in terms of the UAR and the WAR of 3 emotions compared with STT by 1.04% and 0.35%, respectively.

As shown in Table 5, we conduct a further evaluation of FERV39K. GCA+IAL is the state-of-the-art method with the WAR for 7 emotions. Although GCA+IAL has a slightly higher WAR than ours in 7 emotion classes, the UAR performance of this model is significantly lower than our baseline as well as the proposed model. This is a result of emotion group learning, which leads to balanced training. In 3 emotion classes, our AEN outperforms our baseline by 2%/3.44% of UAR/WAR. AEN's confusion matrix for the FERV39K dataset can be seen in Figure 5.

4.6. Visualization

To demonstrate that our proposed AEN and loss function work as intended, we visualize the learned feature map. As shown in Figure 6, we visualize activation maps generated by Grad-CAM [51] for the proposed AEN. We extract activation maps for the temporal transformer at each semantic level and the left of the figure indicates the semantic level s. Since the input of global classifier V_2 should be trained as a discriminative feature for both coarse and fine-grained labels, the activation map at $s = 2$ pays attention general

Method	3 Emotions (%)		7 Emotions (%)	
	UAR	WAR	UAR	WAR
C3D [54]	-	-	22.68	31.69
I3D-RGB [2]	-	-	30.17	38.78
R(2+1)D [55]	-	-	31.55	41.28
3D ResNet18 [14]	-	-	26.67	37.57
ResNet18+LSTM [15, 16]	-	-	30.92	42.95
Former-DFER [73]	59.33	61.35	36.94	46.13
NR-DFERNet [34]	-	-	33.99	45.97
GCA+IAL [33]	-	-	35.82	**48.54**
AEN	**61.33**	**64.79**	**38.18**	47.88

Table 5. Comparison with state-of-the-art methods on FERV39K.

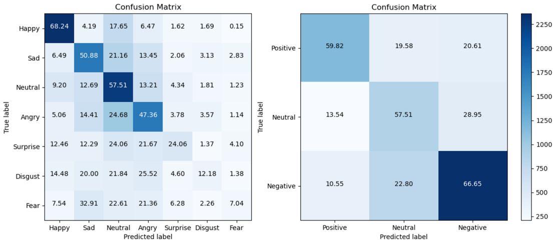

Figure 5. The confusion matrices of our proposed AEN were evaluated on FERV39K dataset.

face area. To classify only coarse labels at local classifier, F_1 is combined with the general face area and some local areas of the face if necessary. To classify fine-grained labels, local classifier at $a = 2$ focuses on the specific regions of the face (mouth and wrinkles) while utilizing the general face area. Hence, it can be seen that AEN learns according to an easy-to-difficult strategy inspired by human cognitive mode.

To verify that the model is properly trained with emotion grouping, moreover, we utilize t-SNE [56] to visualize feature distributions on DFEW fold 1 as shown in Figrue 7. We extracted V_0 for the input frames and implemented t-SNE. In our proposed AEN, it is clearly observed that the feature distribution of fine-grained emotion labels is clustered into coarse labels (Positive, Neutral, and Negative) compared to the baseline Former-DFER. Moreover, the boundaries of the feature distributions between different classes generated from AEN are more obvious, whereas the feature distributions generated from the baseline seem relatively vague. It was confirmed that our model generates discriminative features for each class, and our proposed method satisfies the emotion group learning strategy.

5. Conclusion

The videos of the "in-the-wild" environment are challenging because these datasets generally contain unconstrained dynamic facial expressions with inconsistent emotion, which include some expressions that do not match the target label. Existing methods directly learn to extract discriminative features without any guide. In this paper, we proposed the AEN with emotion group learning and frame-level emotion-guided loss functions. A bottom-up branch in AEN extracts feature representation at the different semantic levels and a top-down branch learns discriminative representations at each affective level by combining feature vectors of each semantic level using S2AC. The frame-level emotion-guided loss functions allow temporal transformer to prevent the loss of target emotional information and to understand what emotions are included in the video clip. The evaluation results have shown that our proposed method exceeded state-of-the-art methods on challenging dynamic FER datasets. Although we have not completely solved the data imbalance problem with the frame-level emotion-guided loss functions, we have seen the potential, so it is necessary to solve the data imbalance problem using single image datasets in future studies.

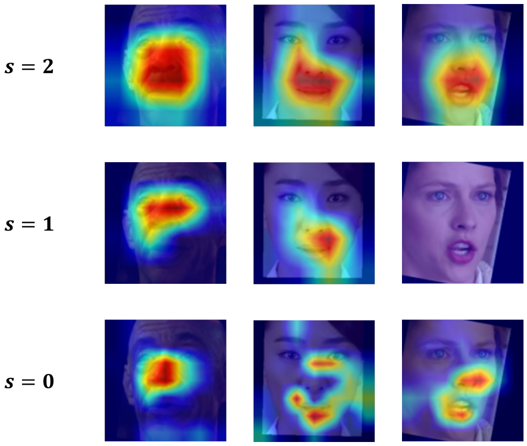

Figure 6. Visualization of the activation maps generated by Grad-CAM for the proposed AEN.

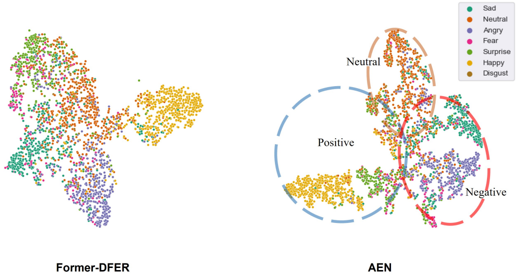

Figure 7. t-SNE visualization results of feature distributions about Former-DFER and our proposed AEN on DFEW fold1. 'Positive', 'Neutral' and 'Negative' denote coarse labels.

Acknowledgment The authors of Korea University were supported by DMLab.

References

[1] Amal Azazi, Syaheerah Lebai Lutfi, Ibrahim Venkat, and Fernando Fernández-Martínez. Towards a robust affect recognition: Automatic facial expression recognition in 3d faces. *Expert Systems with Applications*, 42(6):3056–3066, 2015. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 7, 8

[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 7

[4] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020. 6

[5] Abhinav Dhall. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *2019 International Conference on Multimodal Interaction*, pages 546–550, 2019. 2, 7

[6] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516, 2013. 1, 5

[7] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 467–474, 2015. 2

[8] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992. 6

[9] FriesenW V FacialActionCodingSystem EkmanP. A tech—niqueforthem easurem entoffacialm ovement. *Palo Alto. CA: ConsultingPsychologistsPress*, 12(1):271, 1978. 3

[10] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. Video-based emotion recognition using deeply-supervised neural networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 584–588, 2018. 2

[11] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 445–450, 2016. 2

[12] 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. 1

[13] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 6

[14] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 7, 8

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 7, 8

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 7, 8

[17] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2881–2889, 2020. 1, 5

[18] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 5

[19] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. *arXiv preprint arXiv:2207.01138*, 2022. 1

[20] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 1

[21] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 1

[22] 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. 1

[23] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 1

[24] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn 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*, pages 1–23, 2019. 1

[25] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1

[26] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 1

[27] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 1

[28] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. In *2020 15th*

IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pages 600–605. IEEE, 2020. 2

[29] Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. A multitask learning approach for fake news detection: novelty, emotion, and sentiment lend a helping hand. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 1, 2

[30] Joosung Lee. The emotion is not one-hot encoding: Learning with grayscale label for emotion recognition in conversation. *arXiv preprint arXiv:2206.07359*, 2022. 2

[31] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152, 2019. 1

[32] Min Kyu Lee, Dong Yoon Choi, Dae Ha Kim, and Byung Cheol Song. Visual scene-aware hybrid neural network architecture for video-based facial expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 2

[33] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2208.10335*, 2022. 1, 2, 6, 7, 8

[34] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, et al. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*, 2022. 1, 2, 7, 8

[35] Hangyu Li, Nannan Wang, Xinpeng Ding, Xi Yang, and Xinbo Gao. Adaptively learning facial expression representation via cf labels and distillation. *IEEE Transactions on Image Processing*, 30:2016–2028, 2021. 1, 2, 3

[36] 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. 5

[37] Zongxi Li, Haoran Xie, Gary Cheng, and Qing Li. Word-level emotion distribution with two schemas for short text emotion classification. *Knowledge-Based Systems*, 227:107163, 2021. 2

[38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[39] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. Multi-feature based emotion recognition for video clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 630–634, 2018. 2

[40] Daizong Liu, Xi Ouyang, Shuangjie Xu, Pan Zhou, Kun He, and Shiping Wen. Saanet: Siamese action-units attention network for improving dynamic facial expression recognition. *Neurocomputing*, 413:145–157, 2020. 1

[41] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10631–10642, 2021.

[42] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[43] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 1

[44] Cheng Lu, Wenming Zheng, Chaolong Li, Chuangao Tang, Suyuan Liu, Simeng Yan, and Yuan Zong. Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 646–652, 2018. 2

[45] Cheng Lu, Wenming Zheng, Chaolong Li, Chuangao Tang, Suyuan Liu, Simeng Yan, and Yuan Zong. Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 646–652, 2018. 2

[46] 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 *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010. 1

[47] Fuyan Ma, Bin Sun, and Shutao Li. Spatio-temporal transformer for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2205.04749*, 2022. 1, 2, 6, 7

[48] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. Affect2mm: Affective analysis of multimedia content using emotion causality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5661–5671, 2021. 1

[49] Xi Ouyang, Shigenori Kawaai, Ester Gue Hua Goh, Shengmei Shen, Wan Ding, Huaiping Ming, and Dong-Yan Huang. Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 577–582, 2017. 2

[50] W Gerrod Parrott. *Emotions in social psychology: Essential readings*. psychology press, 2001. 3

[51] 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. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[52] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6248–6257, 2021. 1

[53] Suramya Tomar. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10, 2006. 5

[54] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 7, 8

[55] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 7, 8

[56] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[57] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 569–576, 2017. 2

[58] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020. 1

[59] 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*, 29:4057–4069, 2020. 1

[60] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20922–20931, 2022. 1, 5

[61] Yan Wang, Yixuan Sun, Wei Song, Shuyong Gao, Yiwen Huang, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 101–110, 2022. 1, 2, 6, 7

[62] Yanan Wang, Jianming Wu, and Keiichiro Hoashi. Multi-attention fusion network for video-based emotion recognition. In *2019 International Conference on Multimodal Interaction*, pages 595–601, 2019. 2

[63] Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. Mdan: Multi-level dependent attention network for visual emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9479–9488, 2022. 2, 3, 4, 5

[64] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2412–2418, 2022. 3, 5

[65] Zhenbo Yu, Guangcan Liu, Qingshan Liu, and Jiankang Deng. Spatio-temporal convolutional features with nested lstm for facial expression recognition. *Neurocomputing*, 317:50–57, 2018. 1

[66] 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. 1

[67] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018. 1

[68] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021. 1

[69] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. *arXiv preprint arXiv:2207.10299*, 2022. 1

[70] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, and Shiguang Shan. M 3 f: Multi-modal continuous valence-arousal estimation in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 632–636. IEEE, 2020. 2

[71] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and vision computing*, 29(9):607–619, 2011. 1

[72] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *European conference on computer vision*, pages 425–442. Springer, 2016. 1

[73] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021. 2, 5, 6, 7, 8