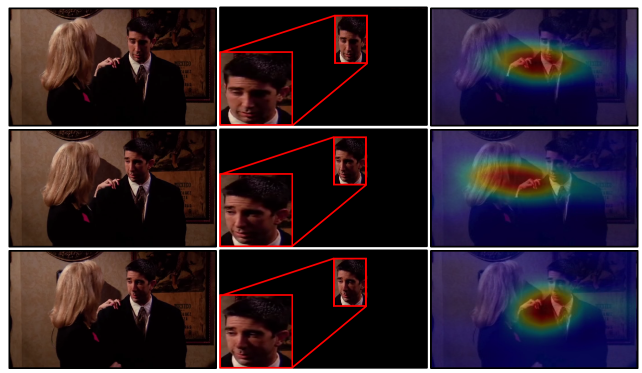
# Context-Aware Emotion Recognition Networks

Jiyoung Lee[1], Seungryong Kim[2], Sunok Kim[1], JungIn Park[1], Kwanghoon Sohn[1,*]
[1]Yonsei University, [2]École Polytechnique Fdrale de Lausanne (EPFL)
{easy00,kso428,newrun,khsohn}@yonsei.ac.kr, seungryong.kim@epfl.ch

## Abstract

*Traditional techniques for emotion recognition have focused on the facial expression analysis only, thus providing limited ability to encode context that comprehensively represents the emotional responses. We present a deep network for context-aware emotion recognition, called CAER-Net, that exploits not only human facial expression but also context information in a joint and boosting manner. The key idea is to hide human faces and seek other contexts based on an attention mechanism. Our networks consist of two sub-networks, including two-stream encoding networks to seperately extract the features of face and context regions, and adaptive fusion networks to fuse such features in an adaptive fashion. We also introduce a novel benchmark for context-aware emotion recognition, called CAER, that is more appropriate than existing benchmarks both qualitatively and quantitatively. On several benchmarks, CAER-Net proves the effect of context for emotion recognition.*



(a) (b) (c)

Figure 1. Intuition of CAER-Net: for untrimmed videos as in (a), conventional methods for emotion recognition that focus on the facial regions only as in (b) cannot provide high accuracy. Unlike these methods, CAER-Net simultaneously focuses on face and attentive context regions as in (c), thus providing improved accuracy.

## 1. Introduction

Recognizing human emotions from visual contents has attracted significant attention in numerous computer vision applications such as health care and human-computer interaction systems [1, 2, 3].

Previous researches for the emotion recognition based on handcrafted features [4] or deep convolutional neural networks (CNNs) [5] have mainly focused on the perception of the facial expression, based on the assumption that facial images are one of the most discriminative features of emotional responses. In this regard, the most widely used datasets, such as the AFEW [6] and FER2013 [7], only provide the cropped and aligned facial images. However, those conventional methods [4, 5] with the facial image data [6, 7] frequently fail to provide satisfactory performance when the emotional signals in the faces are indistinguishable and ambiguous. Otherwise, human recognizes the emotion of others from not only their faces but also their surrounding contexts, such as action, interaction with others, and place [8]. Given untrimmed videos in Fig. 1(a), can we catch how the man feels solely from his facial expression as in Fig. 1(b)? It may be ambiguous to estimate his emotion only with cropped facial videos. However, we can easily recognize his emotion as "sad" with contexts that another woman puts her hand on his shoulder and comforting him as shown in Fig. 1(c). Nevertheless, such contexts have been rarely considered in most existing emotion recognition methods and benchmarks.

Some methods [9, 8] have shown that the emotion recognition performance can be significantly boosted by considering context information such as gesture and place [9, 8]. Meanwhile, in visual sentimental analysis [10, 11] to recognize the sentiment of an image, similar to emotion recognition but not tailored to humans, the holistic visual appearance was used to encode such contexts. However, these approaches are not practical for extracting the salient context information from visual contents. Moreover, large-scale emotion recognition datasets, including various context in-

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science and ICT (NRF-2017M3C4A7069370).

*Corresponding author

formation close in real environments, are absence.

To overcome these limitations, we present a novel framework, called Context-Aware Emotion Recogntion Networks (CAER-Net), to recognize human emotion from images and videos by exploiting not only human facial expression but also scene contexts in a joint and boosting manner, instead of focusing on the facial regions only as in most existing methods [4, 5]. The networks are designed in a two-stream architecture, including two feature encoding stream; face encoding and context encoding streams. Our key ingradient is to seek other relevant contexts by hiding human faces based on an attention mechanism, which enables the networks to reduce an ambiguity in emotion recognition. The face and context features are then fused to predict the emotion class in an adaptive fusion network by inferring an optimal fusion weight among the two-stream features.

In addition, we build a novel database, called Context-Aware Emotion Recognition (CAER) benchmark, by collecting a large amount of video clips from TV shows and annotating the ground-truth emotion. Experimental results show that CAER-Net outperforms some baseline networks for context-aware emotion recognition on several benchmarks, including AFEW [6] and our CAER dataset.

## 2. Related Works

**Emotion recognition approaches.** Most approaches to recognize the human emotion have focused on facial expression analysis [4, 12, 5]. Some methods are based on the facial action coding system [13, 14], where a set of localized movements of the face is used to encode facial expression. Compared to conventional methods that have relied on handcrafted features and shallow classifiers [4, 12], recent deep CNNs based approaches have made significant progress [5]. Various techniques to capture temporal dynamics in videos also have been proposed making connections across the time using recurrent neural networks (RNNs) or deep 3D-CNNs [15, 16]. However, most works have been relied on human face analysis, and thus they have limited ability to exploit context information for emotion recognition in the wild.

To solve these limitations, some approaches using other visual clues apart from the facial appearance have been proposed [17, 18, 9, 8]. While Nicolaou *et al*. [17] used the location of shoulders, Schindler *et al*. [18] used the body pose to recognize six emotion categories under controlled conditions. Chen *et al*. [9] detected events, objects, and scenes using pre-learned CNNs and fused each score with context fusion. In [8], manually annotated body bounding boxes and holistic images were leveraged. However, there are lack of practical solutions to extract the sailent context information and exploit it to context-aware emotion recognition. Moreover, the aforementioned methods have estimated the emotion using pre-traiend networks [9] or on the static images only [8], and thus they have a limited ability to encode dynamic signals captured in a spontaneous uncontrolled manners.

**Emotion recognition datasets.** Conventional datasets, such as CK+ [19] and MMI [20], have been taken under controlled illumination and pose conditions in lab-controlled environments. Recently, datasets with the wild settings [6, 21] have attracted much attention. AFEW benchmark [6] of the EMOTIW challenge [22] provides video frames extracted from movies and TV shows. SFEW database [23] has been built as a static subset of the AFEW. FER-Wild [21] database contains 24,000 images that are obtained by querying emotion-related terms from search engines. MS-COCO database has been recently annotated with object attributes, including some emotion categories for human, but the attributes in the MS-COCO are not intended to be exhaustive for emotion recognition, and not all the people in the database are annotated with emotion attributes. Some studies [24, 25] built the database consisting of a spontaneous subset acquired under a restrictive setting to establish the relationship between emotion and body posture. EMOTIC database [8] has been introduced providing the manually annotated body regions which might contain emotional state. Although these datasets investigate a different aspect of emotion recognition with contexts, large-scale datasets for context-aware emotion recognition are absence that contain various context information.

**Attention inference.** Since deep CNNs have achieved a great success in many computer vision areas [26, 27, 28], numerous attention inference models [29, 30] have been investigated to identify discriminative regions in which the networks attend, by mining discriminative regions [31], implicitly analyzing the higher-layer activation maps [29, 30], and designing different architecture of attention modules [32, 33]. Although the attention map produced by these conventional methods can be used as a prior for various tasks, it only covers small and narrow discriminative regions of the object of interest, and thus frequently fails to capture other discriminative parts that can help performance improvement.

Most related to our work is a method that discovers attentive areas for visual sentiment recognition [11, 34]. Although it produces an emotion sentiment map using deep CNNs, it only focuses on image-level sentiment analysis, not human-centric emotion.

## 3. Proposed Method

### 3.1. Motivation and Overview

In this section, we describe a simple yet effective framework for context-aware emotion recognition in images and videos that exploits the facial expression and context information in a boosting and synergistic manner. A simple solu-

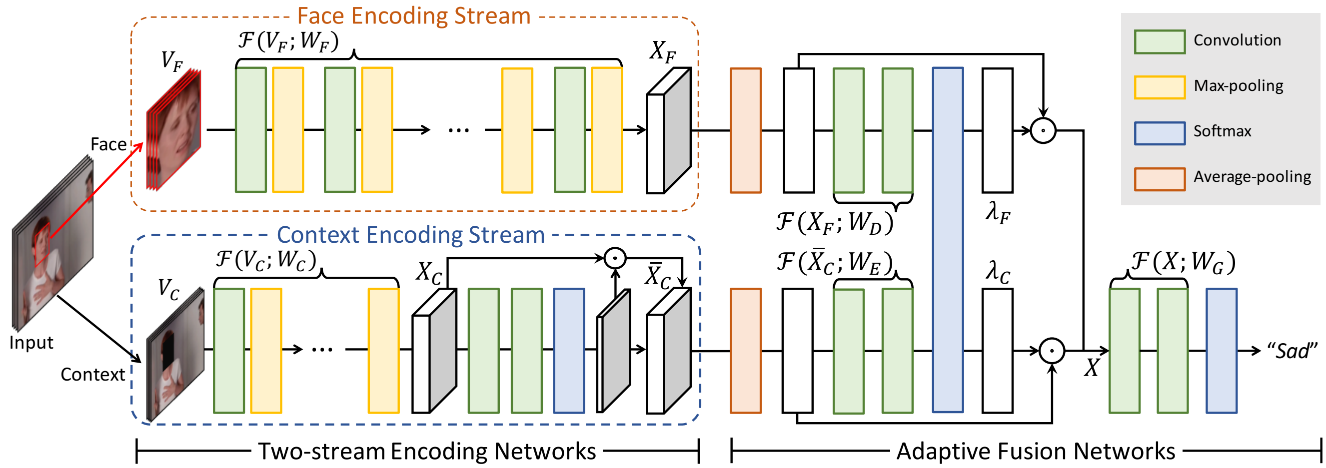



Figure 2. Network configuration of CAER-Net, consisting of two-stream encoding networks and adaptive fusion networks.

tion is to use the holistic visual appearance similar to [8, 9], but such a model cannot encode salient contextual regions well. Based on the intuition that emotions can be recognized by understanding the context components of scene, as well as facial expression together, we present an attention inference module that estimates the context information in images and videos, except for the facial information. By hiding the facial regions in inputs and seeking the attention regions, our networks localize more discriminative context regions that are used to improve emotion recognition accuracy in a context-aware manner.

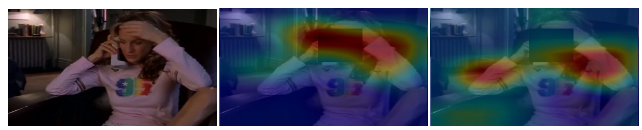
Concretely, let us denote a video that consists of a sequence of $T$ images as $V = \{I_1, \dots, I_T\}$, respectively. Our objective is to infer the discrete emotion label $y$ among $K$ emotion labels $\{y_1, \dots, y_K\}$ of the image $I$ or video clip $V$ with deep CNNs. To solve this problem, we present a network architecture consisting of two sub-networks, including a *two-stream encoding network* and an *adaptive fusion network*, as illustrated in Fig. 2. The two-stream encoding networks consist of *face stream* and *context stream* in which facial expression and context information are encoded in the seperate networks. By combining two features in the adaptive fusion network, an optimal performance for context-aware emotion recognition can be attained.

## 3.2. Network Architectures

### 3.2.1 Two-stream Encoding Networks

In this section, we first present a dynamic model of the two-stream encoding networks for analyzing videos, and then present a static model for analyzing images.

**Face encoding stream.** As in existing facial expression analysis approaches [5, 16, 35], our networks also have the facial expression encoding module. We first detect and crop the facial regions to build input of face stream $V_F$ using the off-the-shelf face detectors [36]. The facial expression encoding module is designed to extract the facial expression features denoted as $X_F$ from temporally stacked face-cropped inputs $V_F$ by feed-foward process such that



(a) input (b) static model (c) dynamic model

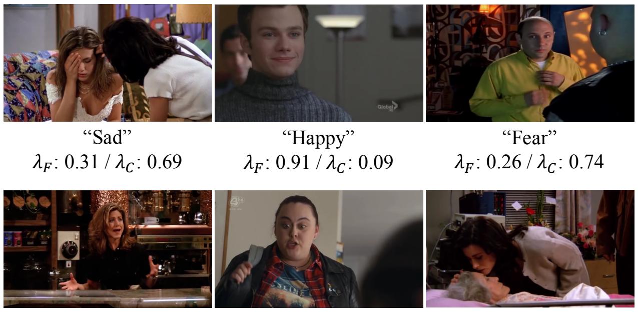
Figure 3. Visualization of the attention maps of (b) static and (c) dynamic context encoding models in CAER-Net.

$$X_F = \mathcal{F}(V_F; W_F), \tag{1}$$

with face stream parameters $W_F$. The facial expression encoding module is designed based on the basic operations of 3D-CNNs which are well-suited for spatiotemporal feature representation. Compared to 2D-CNNs, 3D-CNNs have the ability to model temporal information better for videos due to 3D convolution and 3D pooling operations.

Specifically, the face encoding module consist of 5 convolutional layers with $3 \times 3 \times 3$ kernels followed by batch normalization (BN), rectified linear unit (ReLU) layers and max-pooling layers with stride $2 \times 2 \times 2$ except for the first layer. The first pooling layer has a kernel size $1 \times 2 \times 2$ with the intention of not to merge the temporal signal too early. The number of kernels for five convolution layers are 32, 64, 128, 256 and 256, respectively. The final feature $X_F$ is spatially averaged in the average-pooling layer.

**Context encoding stream.** In comparison to the face encoding stream, the context encoding stream includes a context encoding module and an attention inference module. To extract the context information except the facial expression, we present a novel strategy that hides the faces and seeks contexts based on the attention mechanisms. Specifically, the context encoding module is designed to extract the context features denoted as $X_C$ from temporally stacked face-



"Sad" $\lambda_F$: 0.31 / $\lambda_C$: 0.69 | "Happy" $\lambda_F$: 0.91 / $\lambda_C$: 0.09 | "Fear" $\lambda_F$: 0.26 / $\lambda_C$: 0.74

"Surprise" $\lambda_F$: 0.62 / $\lambda_C$: 0.38 | "Anger" $\lambda_F$: 0.75 / $\lambda_C$: 0.25 | "Sad" $\lambda_F$: 0.13 / $\lambda_C$: 0.87

Figure 4. Some examples of the attention weights, i.e., $\lambda_F$ and $\lambda_C$, in CAER-Net.

hidden inputs $V_C$ by feed-foward process:

$$X_C = \mathcal{F}(V_C; W_C), \tag{2}$$

with context stream parameters $W_C$.

In addition, an attention inference module is learned to extract attention regions of input, enabling the context encoding stream to focus on the sailent contexts. Concretely, the attention inference module takes an intermediate feature $X_C$ as input to infer the attention $A \in \mathbb{R}^{H \times W}$, where $H \times W$ is the spatial resolution of the $X_C$. To make an attention-boosted feature $\hat{X}_C$, we normalize the attention $A$ spatially by using the spatial softmax as follows [37]:

$$\hat{A}_i = \frac{\exp(A_i)}{\sum_j \exp(A_j)}, \tag{3}$$

where $\hat{A}$ is the attention for each pixel $i, j \in \{1, \cdots, H \times W\}$. Since we temporally aggregate the features using 3D-CNNs, we only normalize the attention weight across spatial axises not temporal axis. Note that the attention is implicitly learned in an unsupervised manner.

Normalized attention $\hat{A}$ is then applied to the feature $X_C$ to make the attention-boosted feature $\hat{X}_C$ as follows:

$$\bar{X}_C = \hat{A} \odot X_C, \tag{4}$$

where $\odot$ is an element-wise multiplication operator.

Specifically, we use five convolution layers to extract intermediate feature volumes $X_C$ followed by BN, ReLU layers, and max-pooling layers. All max-pooling layers except for the first layer have $2 \times 2 \times 2$ kernel with stride 2. The first pooling layer has kernel size $1 \times 2 \times 2$ similar to facial expression encoding stream. The number of filters for five convolution layers are 32, 64, 128, and 256, respectively. In the attention inference module, we use two convolution layers with $3 \times 3 \times 3$ kernels producing 128 and 1 feature channels, followed by BN and ReLU layers. The final feature $\bar{X}_C$ is spatially averaged in the average-pooling layer.

**Static model.** Dynamic model described in above can be simplified for emotion recognition in images. Unlike this, a static model takes both a single frame face-cropped image $I_F$ and face-hidden image $I_C$ as input. In networks, all 3D convolution layers and 3D max-pooling layers are replaced with 2D convolution layers and 2D max-pooling layers, respectively. Note that our two types of models have general property regardless of the data type, thus CAER-Net can be utilized for various environments.

Fig. 3 visualizes the attention maps of static and dynamic models. As expected, our networks both with static and dynamic models localize the context information well, except for the face expression. By exploiting the temporal connectivity, the dynamic model can localize more sailent regions compared to the static model.

#### 3.2.2 Adaptive Fusion Networks

To recognize the emotion by using the face and context information in a joint manner, the features extracted from two modules should be combined. However, a direct concatenation of different features often fails to provide optimal performance. To alleviate this limitation, we build the adaptive fusion networks with an attention model for inferring an optimal fusion weight for each feature $X_F$ and $\bar{X}_C$. The attention for them is learned such that $\lambda_F = \mathcal{F}(X_F; W_D)$ and $\lambda_C = \mathcal{F}(\bar{X}_C; W_E)$ with network parameters $W_D$ and $W_E$, respectively. Softmax function make the sum of these attentions to be 1, *i.e.*, $\lambda_F + \lambda_C = 1$.

Fig. 4 shows some examples of the attention weights, i.e., $\lambda_F$ and $\lambda_C$, in CAER-Net. According to contents, the attention weights are adaptively determined to yield an optimal solution.

Unlike methods using the simple concatenation [8], the learned attentions are applied to inputs as

$$X_A = \Pi(X_F \odot \lambda_F, \bar{X}_C \odot \lambda_C), \tag{5}$$

where $\Pi$ is a concatenation operator. We then estimate the final output $y_k$ for emotion category $k$ by classifier:

$$y_k = \mathcal{F}(X_A; W_G), \tag{6}$$

where $W_G$ represents the remainder parameters of the adaptive fusion networks.

Specifically, the fusion networks consist of 6 convolution layers with $1 \times 1$ kernels. The four layers use to produce fusion attention $\lambda_{C,F}$. While the intermediate two layers that receive each stream feature as input produce 128 channel feature, the remaining two layers produce 1 channel attention for facial and contextual features. For the two layers that act as final classifiers, the first convolution layer produces 128 channel feature followed by ReLU and dropout layers to prevent the problem of the network overfitting, and the second convolution layer produces $K$ channel feature to estimated the emotional category.

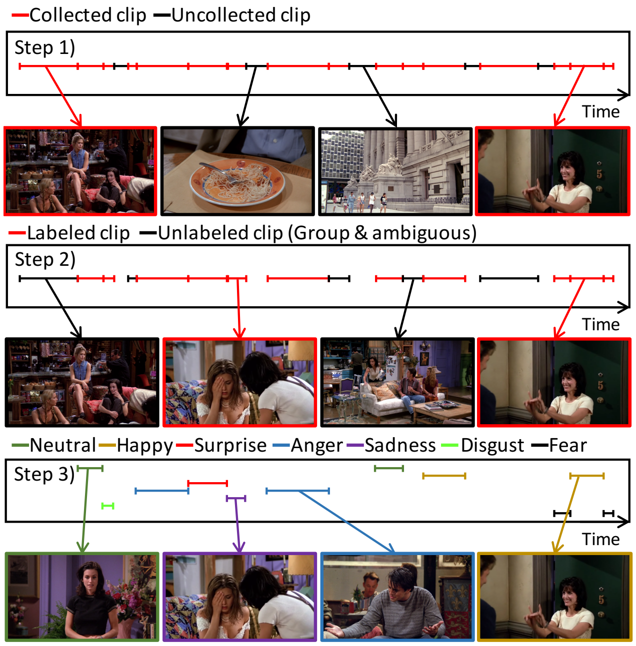



Figure 5. Procedure for building CAER benchmark: we divide the video clips to the shot with shot boundary detection method, and remove face-undetected shots, group-level and ambiguous shots to estimate the emotion. Finally, we annotate the emotion category.

## 4. The CAER Benchmark

Most existing datasets [7, 38] have focused on the human facial analysis, and thus they are inappropriate for context-aware emotion recogntion. In this section, we introduce a benchmark by collecting video clips from TV shows and annotating them for context-aware emotion recogntion. The dataset is available at https://caer-dataset.github.io.

### 4.1. Annotation

We first collected the video clips from 79 TV shows and then refined them using the shot boundary detector, face detector/tracking and feature clustering [1]. Each video clip was manually annotated with six emotion categories, including “anger”, “disgust”, “fear”, “happy”, “sad”, and “surprise“, as well as “neutral”. Six annotators were recruited to assign the emotion category to the 20,484 clips in the initial collection. Since all the video clips have audio and visual tracks, the annotators labeled the database while listening to the audio tracks for more accurate annotations. Each clip was evaluated by three different annotators. The annotation was performed blindly and independently, *i.e*. the annotators were not aware of the other annotator’s response. Importantly, in comparison of existing datasets [6, 8], confidence scores were annotated as well as emotion category, which can be thought as the probability of the annotation reliability. If two more annotators assigned the same emotion categories, the clip was remained in the database. We also remove the clips which have lower confidence average under the 0.5. Finally, 13,201 clips and about 1.1M frames were available. The videos range from short (around 30 frames) to longer clips (more than 120 frames). The average of sequence length is 90 frames. In addition, we extracted about 70K static images from CAER to create a static image subset, called CAER-S. The dataset is randomly split into training (70%), validation (10%), and testing (20%) sets. Overall stage of data acquisition and annotation is illustrated in Fig. 5. Table 1 summarizes the number of clips per each cateogry in the CAER benchmark.

[1] https://github.com/pyannote/pyannote-video

| Category | # of clips | # of frames | % |
|---|---|---|---|
| Anger | 1,628 | 139,681 | 12.33 |
| Disgust | 719 | 59,630 | 5.44 |
| Fear | 514 | 46,441 | 3.89 |
| Happy | 2,726 | 219,377 | 20.64 |
| Neutral | 4,579 | 377,276 | 34.69 |
| Sad | 1,473 | 138,599 | 11.16 |
| Surprise | 1,562 | 126,873 | 11.83 |
| Total | 13,201 | 1,107,877 | 100 |

Table 1. Amount of video clips in each category on the CAER.

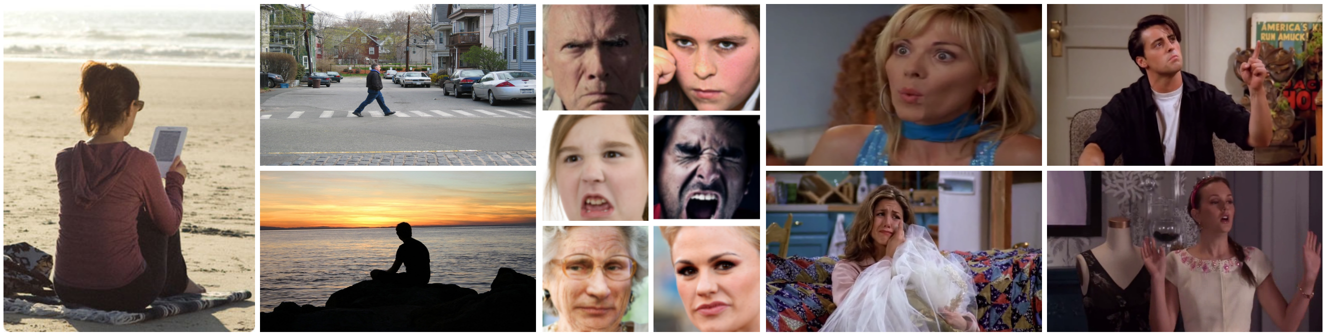
### 4.2. Analysis

We compare the CAER dataset with other widely used datasets, such as EMOTIC [8], AffectNet [38], AFEW [39], and Video Emotion datasets [40], as shown in Table 2. According to the data type, the datasets are grouped into the static and dynamic. Even if static databases for facial expression analysis such as AffectNet [38] and FER-Wild [21] collect a large amount of facial expression images from the web, they have only face-cropped images not including surrounding context. In addition, EMOTIC [8] do not contain human facial images, as exampled in Fig. 6, thus causing subjective and ambiguous labelling from observers. On the other hand, commonly used video emotion recognition datasets had insufficient amount of data than image-based datasets [40, 41]. Compared to these datasets, the CAER dataset provides the large-scale video clips which are sufficient amount to learn the machine learning algorithms for context-aware emotion recognition.

## 5. Experiments

### 5.1. Implementation Details

The CAER-Net was implemented with PyTorch library [42]. We trained CAER-Net from scratch with learning rate initialized as $5 \times 10^{-3}$ and dropped by a factor of 10 every 4 epochs. CAER-Net was learned with the cross-entropy loss function [43] with ground-truth emotion labels with batch size to 32. As CAER dataset has various length of videos, we randomly extracted single non-overlapped consecutive 16 frame clips from every training video which sampled at 10 frames per second. While the clips of fa-




(a) EMOTIC [8] (b) AffectNet [38] (c) CAER

Figure 6. Examples in the EMOTIC [8], AffectNet [38] and CAER. While EMOTIC includes face-unvisible images to yeild ambiguous emotion recognition, AffectNet includes face-cropped images which have limited to use of context.

| Data type | Dataset | Amount of data | Setting | Annotation type | Context |
|---|---|---|---|---|---|
| Static (Images) | EMOTIC [8] | 18,316 images | Web | 26 Categories | ✓ |
| | AffectNet [38] | 450,000 images | Web | 8 Categories | ✗ |
| | CAER-S | 70,000 images | TV show | 7 Categories | ✓ |
| Dynamic (Videos) | AFEW [39] | 1,809 clips | Movie | 7 Categories | ✗ |
| | CAER | 13,201 clips | TV show | 7 Categories | ✓ |

Table 2. Comparison of the CAER with existing emotion recognition datasets such as EMOTIC [8], AffectNet [38], AFEW [39], and Video Emotion [40] datasets. Compared to existing datasets, CAER contains large amount of video clips for context-aware emotion recognition.

cial $V_F$ are resized to have the frame size of $96 \times 96$, the clips of contextual parts $V_C$ are resized to have the frame size of $128 \times 171$ and randomly cropped into $112 \times 112$ at training stage. We also trained static model of CAER-Net-S with CAER-S dataset with the input size of $224 \times 224$. To reduce the effects of the network overfitting, we employed the dropout scheme with the ratio of 0.5 between $1 \times 1$ convolution layers, and data augmentation schemes such as flips, contrast, and color changes. At testing phase, we used a single center crop per contextual parts clips. For video predictions, we split a video into 16 frame clips with a 8 frame overlap between two consecutive clips then average clip predictions of all clips.

### 5.2. Experimental Settings

We evaluated CAER-Net on the CAER dataset and AFEW dataset [6]. For evaluation of the proposed networks quantitatively, we measured the emotion recognition performance by classification accuracy as used in [22]. For the basic CNN models, we report the results of using four classical deep learning methods pre-trained on ImageNet [44] and Sports-1M [45], and fine-tuned on the CAER dataset: AlexNet [26], VGGNet [27], ResNet [28], and C3D [46]. For pre-trained models, we used ImageNet [44] and Sports-1M [45] pre-trained models. Otherwise we trained all parameters of learning rate $10^{-4}$ for fine-tuned models.

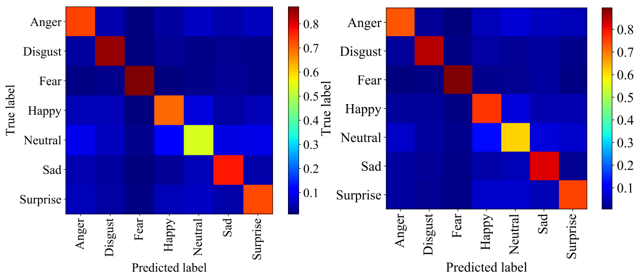
### 5.3. Results on the CAER dataset

**Ablation study.** We analyzed CAER-Net-S and CAER-Net with ablation studies as varying the combination of different inputs such as cropped face and context, and attention modules such as context and fusion attention modules.

| Methods | w/F | w/C | w/cA | w/fA | Acc. (%) |
|---|---|---|---|---|---|
| CAER-Net-S | ✓ | | | | 70.09 |
| | | ✓ | ✓ | | 65.65 |
| | ✓ | ✓ | ✓ | ✓ | 73.51 |
| CAER-Net | ✓ | | | | 74.13 |
| | | ✓ | ✓ | | 71.94 |
| | ✓ | ✓ | | | 74.36 |
| | ✓ | ✓ | ✓ | | 74.94 |
| | ✓ | ✓ | | ✓ | 75.57 |
| | ✓ | ✓ | ✓ | ✓ | 77.04 |

Table 3. Ablation study of CAER-Net-S and CAER-Net on the CAER-S and CAER datasets, respectively. 'F', 'C', 'cA', and 'fA' denote face encoding stream, context encoding stream, context attention module, fusion attention module, respectively.

For all those experiments, CAER-Net-S and CAER-Net were trained and tested on the CAER-S and CAER datasets, respectively. Table 3 shows the quantitative results for ablation study, measured the average accuracy the estimator and ground-truth labels. The results show that the best result can be obtained when both the face and context are used as inputs. As our baseline, CAER-Net w/F that considers facial expression only for emotion recognition provides the accuracy 74.13 %. Compared to this, our CAER-Net that fully makes use of both face and context shows the best performance. When we compared the static and dynamic models, CAER-Net shows 3.53 % improvement




(a) CAER-Net w/F (b) CAER-Net

Figure 7. Confusion matrix of CAER-Net with face stream only and with face and context streams on the CAER benchmark.

(a) (b) (c) (d) (e)

Figure 8. Visualization of the attention: (a) inputs, (b) CAER-Net-S and (c) CAER-Net without hiding the face, (d) CAER-Net-S and (e) CAER-Net with hiding the face during training.

than CAER-Net-S, which shows the importance to consider the temporal dynamic inputs for context-aware emotion recognition. Fig. 7 demonstrates the confusion matrix of CAER-Net w/F and CAER-Net, which also verify that compared to the model that only focuses on facial stream only, a joint model that considers facial stream and context stream simultaneously can highly boost the emotion recognition performance. Happy and neutral accuracies were increased by 7.48% and 5.65%, respectively, which clearly shows that context information helps distinguishing these two categories rather than only using facial expression. Finally, we conducted an ablation study for the context attention module. First of all, when we trained CAER-Net-S and CAER-Net without hiding the face, they tended to focus on the most discriminative parts only (*i.e.*, faces) as depicted in top Fig. 8, respectively. Secondly, we conducted another experiment on *actionless* frames. As shown in bottom Fig. 8, both CAER-Net-S and CAER-Net focused on the salient scenes that can be an emotion signals.

To summarize, through our experiments, we validated the effectiveness of the joint use of facial expression and contextual information. The context in the visual contents was helpful even when facial information was lacking.

**Comparison to baseline methods.** In Fig. 9 and Table 4, we evaluated CAER-Net-S with baseline 2D CNNs based approaches. The standard networks including AlexNet [26], VGGNet [27], and ResNet [28] trained with ImageNet were examined, which perform well for general image classification tasks. In addition, we also fine-tuned these networks on the CAER-S dataset. Compared to these baseline methods, our CAER-Net-S improves the classification

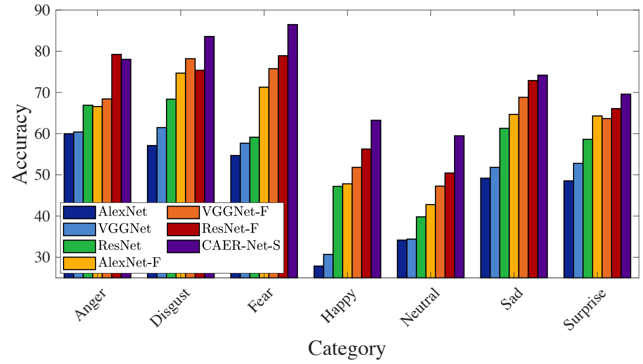



Figure 9. Quantitative evaluation of CAER-Net-S in comparison to baseline methods on each category in the CAER-S benchmark.

| Methods | Acc. (%) |
|---|---|
| ImageNet-AlexNet [26] | 47.36 |
| ImageNet-VGGNet [27] | 49.89 |
| ImageNet-ResNet [28] | 57.33 |
| Fine-tuned AlexNet [26] | 61.73 |
| Fine-tuned VGGNet [27] | 64.85 |
| Fine-tuned ResNet [28] | 68.46 |
| CAER-Net-S | 73.51 |

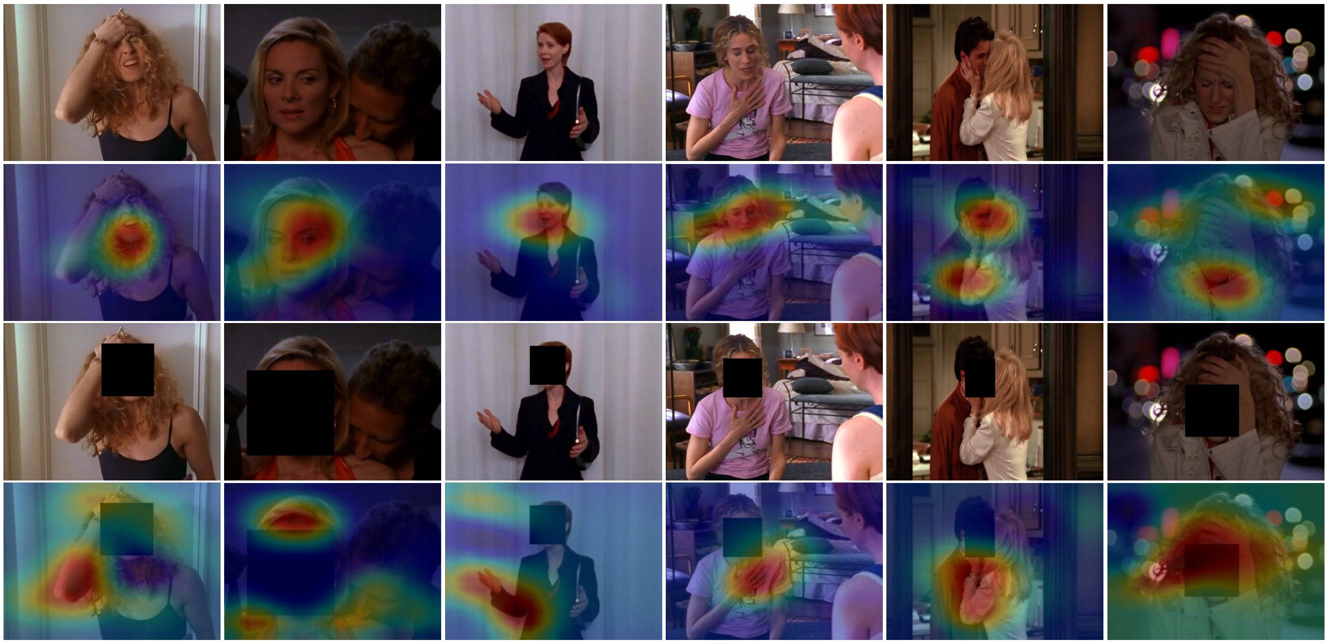
Table 4. Quantitative evaluation of CAER-Net-S in comparison to baseline methods on the CAER-S benchmark .

performance than fine-tuned ResNet by 5.05%. Moreover, CAER-Net-S consistently performs favorably against baseline deep networks on each category in the CAER-S benchmark, which illustrates that CAER-Net can learn more discriminative representation for this task. In addition, we evaluated CAER-Net with a baseline 3D CNNs based approach in Table 5. Compared to C3D [46], our CAER-Net has shown the state-of-the-art performance on the CAER benchmark.

Finally, Fig. 10 shows the qualitative results with learned attention maps obtained by CAM [29] with fine-tuned VGGNet and in context encoding stream of CAER-Net-S. Note that images in Fig. 10 were correctly classified to ground-truth emotion categories both with fine-tuned VGGNet and CAER-Net-S. Unlike CAM [29] that only considers facial expressions, the attention mechanism in CAER-Net-S localizes context information well that can boost the emotion recognition performance in a context-aware manner.

### 5.4. Results on the AFEW dataset

We conducted an additional experiment to verify the effectiveness of the CAER dataset compared to the AFEW dataset [6]. For this experiment, we learned CAER-Net on the CAER dataset and tested on the AFEW dataset. In the results, the networks trained on CAER dataset show the better performance, which means that the CAER dataset is more appropriate then the AFEW dataset in terms of both quantity and quality. When we learned CAER-Net on the



(a) "Disgust" (b) "Fear" (c) "Surprise" (d) "Sad" (e) "Happy" (f) "Fear"

Figure 10. Visualization of learned attention maps in CAER-Net-S: (from top to bottom) inputs, attention maps of CAM [29], inputs of context encoding stream, attention maps in context encoding stream. Note that red color indicates attentive regions and blue color indicates suppressed regions. Best viewed in color.

| Methods | Acc. (%) |
|---|---|
| Sports-1M-C3D [46] | 66.38 |
| Fine-tuned C3D [46] | 71.02 |
| CAER-Net | 77.04 |

Table 5. Quantitative evaluation of CAER-Net in comparison to C3D [46] on the CAER benchmark .

| Methods | Training data | Acc. (%) |
|---|---|---|
| VielZeuf *et al*. [47] w/F | FER+AFEW | 48.60 |
| Fan *et al*. [15] w/F | FER+AFEW | 48.30 |
| Hu *et al*. [48] w/F | AFEW | 42.55 |
| Fan *et al*. [35] w/F | FER+AFEW | 57.43 |
| CAER-Net w/F | AFEW | 41.86 |
| CAER-Net | CAER | 38.65 |
| CAER-Net | AFEW | 43.12 |
| CAER-Net | CAER+AFEW | 51.68 |

Table 6. Quantitative evaluation of CAER-Net on the AFEW [6] benchmark, as varying training datasets.

combination of CAER and AFEW datasets, the highly improvement was attained. It should be noted that the state-of-the-art methods such as Fan *et al*. [35] have shown the better performance, they are formulated the networks with the ensemble of various networks and learning techniques on the facial images only, without considering both face and context. Unlike this, we focused on investigating how context information, excluding the face, helps to boost the emotion recognition performance. For this purpose, we choice shallow architecture rather than Fan *et al*. [35]. If the face encoding stream adopt more complicated networks such Fan *et al*. [35], the performance of CAER-Net also will be highly boosted. We reserve this as further works.

## 6. Conclusion

In this paper, we presented CAER-Net for context-aware emotion recognition from visual contents that exploits not only human facial expression but also contexts in a joint and boosting manner. The key idea of this approach is to seek sailent context information by hiding the facial regions with an attention mechanism, and utilize this to estimate the emotion from contexts, as well as the facial information together. We proposed the simple yet effective deep architecture consisting of two-stream encoding networks and adaptive fusion networks. The two-stream network jointly recognizes facial expression and context. The face and context features are fused in the adaptive fusion networks. We also introduced the CAER benchmark that is more appropriate for context-aware emotion recognition than existing benchmarks both qualitatively and quantitatively. Experimental results show that facial and context information can serve as complementary one to the other in emotion recognition. We believe that the results of this study will facilitate further advances in context-aware emotion recognition and its related tasks.

## References


[1] Sidney D'Mello, Rosalind W Picard, and Arthur Graesser. Toward an affect-sensitive autotutor. *IEEE Int. Systems*, 2007.

[2] Christina Lisetti, Fatma Nasoz, Cynthia LeRouge, Onur Ozyer, and Kaye Alvarez. Developing multimodal intelligent affective interfaces for tele-home health care. *Int. Jou. of Hum.-Comp. Stud.*, 2003.

[3] Georgios N Yannakakis and Julian Togelius. Experience-driven procedural content generation. *IEEE Trans. AC*, 2011.

[4] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vis. Comput.*, 2009.

[5] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. *In: CVPR*, 2016.

[6] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Acted facial expressions in the wild database. *Technical Report TR-CS-11*, 2011.

[7] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. *In: ICONIP*, 2013.

[8] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. *In: CVPR*, 2017.

[9] Chen Chen, Zuxuan Wu, and Yu-Gang Jiang. Emotion in context: Deep semantic feature fusion for video emotion recognition. *In: MM*, 2016.

[10] Bing Li, Weihua Xiong, Weiming Hu, and Xinmiao Ding. Context-aware affective images classification based on bilayer sparse representation. *In: MM*, 2012.

[11] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. *In: CVPR*, 2018.

[12] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. Learning active facial patches for expression analysis. *In: CVPR*, 2012.

[13] E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 1978.

[14] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE Trans. IP*, 2015.

[15] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. *In: ICMI*, 2016.

[16] Jiyoung Lee, Sunok Kim, Seungryong Kim, and Kwanghoon Sohn. Spatiotemporal attention based deep neural networks for emotion recognition. *In: ICASSP*, 2018.

[17] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. AC*, 2011.

[18] Konrad Schindler, Luc Van Gool, and Beatrice de Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neur. Net.*, 2008.

[19] 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: CVPR Work.*, 2010.

[20] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. *In: ICME*, 2005.

[21] Ali Mollahosseini, Behzad Hasani, Michelle J Salvador, Hojjat Abdollahi, David Chan, and Mohammad H Mahoor. Facial expression recognition from world wild web. *In: CVPR Work.*, 2016.

[22] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. *In: ICMI*, 2016.

[23] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. *In: ICCV Work.*, 2011.

[24] Andrea Kleinsmith and Nadia Bianchi-Berthouze. Recognizing affective dimensions from body posture. *In: ACII*, 2007.

[25] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic recognition of non-acted affective postures. *IEEE Trans. Systems*, 2011.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *In: NeurIPS*, 2012.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *In: CVPR*, 2016.

[29] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *In: CVPR*, 2016.

[30] 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: ICCV*, 2017.

[31] Krishna Kumar Singh, Fanyi Xiao, and Yong Jae Lee. Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection. *In: CVPR*, 2016.

[32] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. *In: ECCV*, 2018.

[33] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *In: CVPR*, 2018.

[34] Quanzeng You, Hailin Jin, and Jiebo Luo. Visual sentiment analysis by attending on local image regions. *In: AAAI*, 2017.

[35] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. Video-based emotion recognition using deeply-supervised neural networks. *In: ICMI*, 2018.

[36] Davis E King. Dlib-ml: A machine learning toolkit. *Joul. of Mach. Learn. Res.*, 2009.

[37] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv:1511.04119*, 2015.

[38] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. AC*.

[39] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multi.*, 2012.

[40] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. Predicting emotions in user-generated videos. *In: AAAI*, 2014.

[41] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vis. Comput.*, 2017.

[42] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[43] Sunok Kim, Dongbo Min, Seungryong Kim, and Kwanghoon Sohn. Unified confidence estimation networks for robust stereo matching. *IEEE Trans. IP*, 2018.

[44] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *In: CVPR*, 2009.

[45] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. *In: CVPR*, 2014.

[46] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. *In: ICCV*, 2015.

[47] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. Temporal multimodal fusion for video emotion classification in the wild. *In: ICMI*, 2017.

[48] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. Learning supervised scoring ensemble for emotion recognition in the wild. *In: ICMI*, 2017.