Tri-modal Recurrent Attention Networks for Emotion Recognition

Jiyoung Lee, *Student Member, IEEE,* Sunok Kim, *Member, IEEE,* Seungryong Kim, *Member, IEEE,* and Kwanghoon Sohn, *Senior Member, IEEE*

***Abstract*—Recent deep networks based methods have achieved state-of-the-art performance on a variety of emotion recognition tasks. Despite such progress, previous researches on affective computing to estimate human-centric emotion have mainly focused on analyzing color recording videos only. However, complex emotion having dynamic facial expression, lighting conditions and various skin colors can be fully-understood by integrating information from multiple modality videos. We present a novel method that estimates dimensional emotion states taking color, depth, and thermal recording videos as inputs which could be complementary to each other. Our networks, termed as tri-modal recurrent attention networks (TRAN), learn spatiotemporal attention cubes to robustly recognize the emotion based on attention-boosted feature cubes. We leverage the depth and thermal sequences as guidance priors and transfer the guidance attention cubes from guidance stream to color stream for selectively focusing on emotional discriminative regions within facial videos. We also introduce a novel benchmark for tri-modal emotion recognition, called TAVER, which consists of color, depth, and thermal recording videos with continuous arousal-valence score. The experimental results show that our method can achieve the state-of-the-art results in dimensional emotion recognition on exsiting color recording datasets including RECOLA, SEWA, and our TAVER datasets.**

***Index Terms*—Tri-modal Emotion Recognition, Dimensional (Continuous) Emotion Recognition, Attention Mechanism**

I. Introduction

UNDERSTANDING human emotions from visual contents has attracted significant attention in numerous affective computing and computer vision applications such as health [1], personal assistance robots [2], and many other human-computer interaction systems [3].

There are two major emotion recognition models according to theories in psychology research [4]: categorical models and dimensional models. Most efforts in emotion recognition [5]–[9] have focused on categorical emotion description, where emotions are grouped into discrete categories such as surprise, fear, etc. In the last few years, several methods have tried to recognize the six basic emotions [5]–[11]. Although the state-of-the-art methods have shown satisfactory performance in categorical emotion recognition, those six basic emotions do not cover the full range of possible emotions, which hinders the application of emotion recognition methods to practical systems.

Jiyoung Lee, Sunok Kim and Kwanghoon Sohn are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea. E-mail: {easy00, kso428, khsohn}@yonsei.ac.kr

Seungryong Kim is with École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. E-mail: seungryong.kim@epfl.ch

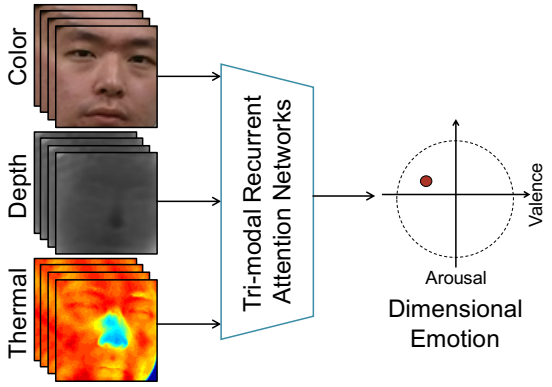


Fig. 1. Illustration of TRAN: using tri-modal facial videos including color, depth, and thermal videos, TRAN recognizes dimensional emotion.

To alleviate these limitations, dimensional emotion descriptions [8], [12]–[15] have attracted much attention, where emotions are described in a continuous space, consisting of two representative domains, called Arousal and Valence. Arousal represents how engaged or apathetic a subject appears while valence represents how positive or negative a subject appears. Those models can represent more complex and subtle emotions with the higher-dimensional descriptions compared to the categorical emotion description. Some researches proposed to estimate dimensional emotion state based on still frames [13], [14]. These methods effectively extract spatial information but fail to model the variability of emotion expression in temporal factors. Thus, researchers have tried to capture the dynamic variation from consecutive frames based on hand-crafted [16] or learned features [8], [15].

On the other hand, there have been attempts to recongnize human emotion through vaious signals such as face, voice, and biological signal to improve the estimation accuracy of emotional state [16]–[18]. In this paper, we focus on recognizing human emotion through fusion of multiple modalities with heterogeneous characteristics such as color, depth, and thermal information. Researchers in affective computing have used depth sensors and 3D models to improve facial feature tracking and expression recognition [19], [20]. Although there are several facial expression databases that include 3D data [19], [21], they are all based on 3D model, not including thermal information. Moreover, thermal sensor has also been used for these tasks [22], [23] due to the reason of its sensitivity to human's skin temperature and relative insensitivity to lighting conditions and skin color. Nevertheless, previous works have mainly utilized the temperature information as a single modal-

ity. Most recently, Zheng *et al.* [21] has built multi-modal 3D dynamic benchmark including the thermal video. However, it was annotated for categorical emotion labels, not dimensional domains.

In this paper, we present a novel framework, called tri-modal recurrent attention networks (TRAN), to estimate dimensional human emotion by exploiting not only a color recording video but also depth and thermal recording videos in a joint and boosting manner, as illustrated in Fig. 1. The key ingredient of this approach is to seek the spatio-temporal attention part by leveraging complementary tri-modal information and use this for dimensional emotion recognition. The proposed network consists of four sub-networks; spatial encoder networks, temporal decoder networks, attention inference networks, and emotion recognition networks. First of all, we extract facial features by spatial encoder networks with spatial associations. In temporal decoder networks, we adaptively fuse the tri-modal input features to learn 'where' and 'what' to attend guided by information from other modalities. To this end, we propose a guided attention long-short term momory (GA-LSTM) module that learns emotionally attentive faical parts both spatially and temporally with tri-modal facial videos. Temporally-stacked feature cubes are multiplied with estimated attention cubes to make attention-boosted feature cubes in attention inference networks. Lastly, the emotion recognition networks are formulated using successive 3D convolutional neural networks (3D-CNNs) to deal with the sequential data for recognizing dimensional emotion scores.

In addition, we build a new tri-modal database including color, depth and thermal videos for the dimensional emotion recognition task, termed as tri-modal arousal-valence emotion recognition database (TAVER). To the best of our knowledge, it is the first publicly available dataset for dimensional emotion recognition based on tri-modalities facial videos. By focusing on discriminative parts of facial with tri-modal videos, the proposed emotion recognition technique achieves the state-of-the-art performance on the tri-modal benchmark such as the TAVER and various uni-modal benchmarks such as RECOLA [24] and SEWA [25].

This manuscript extends the conference version of this work [26] through (1) a tri-modal extension of URAN, called TRAN; (2) an introduction of novel database, called TAVER and (3) an extensive comparative study with state-of-the-art CNN-based methods using various datasets.

The rest of this paper is organized as follows. We discuss the related work in Sec. II and describe our emotion recognition algorithm in Sec. III. We introduce the novel TAVER benchmark in Sec. IV. Sec. V presents the details of our experiments and Sec. VI concludes this paper.

II. Related Works

A. Emotion Recognition Methods

There have been numerous approaches to recognize humans emotion which can be described into two ways: *categorical* models and *dimensional* models. A large portion of the previous research has focused on recognizing categorical emotions, where emotions are grouped into discrete categories [11], [27], [28]. Although these approaches can recognize humans emotion with high accuracy, the categorical emotion model has limitation that does not cover the full range of humans emotion. As an alternative way to model emotions, dimensional approaches have been proposed [8], [12]–[15], where humans emotion can be described using a low-dimensional signal.

Recently, deep convolutional neural networks have been shown to substantially outperform previous approaches in various applications such as face recognition [29], facial point detection [30], and face verification [31] with extraction of more discriminative features. Meanwhile, the dynamic emotion recognition models have also highly benefited with the advent of deep CNNs leveraging various types of deep CNNs such as Time-Delay, recurrent neural networks (RNN), and LSTM networks [32]–[34]. In the field of categorical emotion recognition, Ebrahimi *et al.* [35] combined CNNs and RNNs to recognize categorical emotions in videos. The networks were first trained to classify emotion from static images, then the extracted features from the CNNs were used to train RNNs to estimate emotion for the whole video. Dynamic and dimensional models have also been considered in [36], [37]. He *et al.* [37] used deep bidirectional LSTM architecture for the fusion of multimodal features to dimensional affect prediction. While [38] used static methods to make the initial affect predictions at each time step, it used particle filters to make the final prediction. Similarly, Khorrami *et al.* [36] showed the possibility that combination of CNNs and RNNs could improve the performance of dimensional emotion recognition. However, most works for dynamic and dimensional emotion recognition have been relied on color recording videos, and thus they have limited ability to exploit complementary tri-modal information.

B. Emotion Recognition Benchmarks

Most databases that deal with emotion recognition [39]–[42] were recorded by color cameras, although they posed a challenge to affective computing due to various characteristics from age, gender and skin color of people. It may yield limited performance to apply the system in practical applications. In this regards, capturing spontaneous expression has become a trend in the affective computing community. For example, recording the responses of participants' faces while watching a stimulti (*e.g.*, DISFA [43] and AM-FED [44]) or performing laboratory-based emotion inducing tasks (*e.g.*, Belfast [45]). These databases often capture multiple attributes such as voice, biological signals, etc. Sequences of frames are usually captured that enable researchers to work on temporal and dynamic aspects of expressions.

Many researchers have developed databases for the dimensional model in the continuous domain from controlled setting [24], [45]–[47] to wild setting [25], [42]. The Belfast database [45] contains recordings of mild to moderate emotional responses of 60 participants to a series of laboratory-based emotion inducing tasks (*e.g.*, surprise response by setting off a loud noise when the participant is asked to find something in a black box.) The recordings were labeled by information on self-report of emotion, the gender of the participant/experimenter, and the valence in the continuous

domain. The arousal dimension was not annotated in Belfast database. The participants reported their arousal and valence through the self-assessment manikin (SAM) [46] questionnaire before and after the tasks. Audio-Visual Emotion recognition Challenge (AVEC) series of competitions [16], [18], [48] have provided benchmarks of automatic audio, video and audio-visual emotion analysis in dimensional emoition recognition, where SEMAINE, RECOLA, and SEWA benchmarks were included. Various dimensions of emotion recognition were explored in each challenge year such as valence, arousal, expectation, power, and dominance, where the prediction of valence and arousal are studied in all challenges.

Although there are several emotion recognition databases based on 3D data [19], [20], [49], they are all based on posed behavior and typically include few subjects, little diversity, limited ground-truth labels, and limited metadata. Recently, Zheng *et al.* [21] has built multi-modal 3D dynamic benchmark including thermal video. However, it was annotated for categorical emotion labels and AUs, not dimensional domains.

C. Multi-spectral Fusion

It is desirable to leverage multi-modal information to overcome the limitations of color recording visual contents in various computer vision applications such as image dehazing, image denoising, pedestrian detection, and human body segmentation, providing complementary information [50]–[54]. For example, Feng *et al.* [50] proposed an image dehazing method by modeling a dissimilarity between color and NIR images. The NIR image was used as a guidance image in image denoising applications [52] and as a supplementary data in pedestrian detection systems [51], [53]. Kim *et al.* [52] leveraged CNNs to enhance a noisy RGB image using a aligned NIR image via alternating minimization. Park *et al.* [53] simultaneously fused each distinctive color and NIR features to get optimal performance. Incorporating visible images and other spectral images into a high-level framework provides complementary information and improves the performance. Palmero *et al.* [54] proposed human body segmentation dataset including color, depth, and thermal modalities in indoor scenarios to segment human subjects automatically in tri-modal video sequences based on learning-based fusion strategies.

D. Attention Inference

Attention is widely known as playing an important role in human perception system [55]. One important property of a human visual system is that one does not attempt to process a whole scene at once. Instead, humans exploit a sequence of partical glimpses and selectively focus on salient parts in order to capture the better visual structure [56].

Recently, there have been several attempts to incorporate attention processing to improve the performance of CNNs in image classification and object detection tasks [57]–[60]. Wang *et al.* [57] proposed residual attention networks for generation of attention-aware feature maps. By leveraging global average pooling layer, Zhou *et al.* [58] built class activation maps in CNNs. Hu *et al.* [59] introduced a squeeze-and-excitation module to exploit the inter-channel relationship. Woo *et al.* [60] simultaneously estimated spatial and channel attention with convolutional block attention module (CBAM).

Previous attention-based techniques using recurrent modules have estimated the attention by stack of LSTM modules [61], [62]. For example, Jia *et al.* [62] has proposed the extension of LSTM model, called gLSTM, for image caption generation. Although they employ temporal information, they cannot take a spatial correlation into consideration. To alleviate this limitation, Li *et al.* [63] have employed ConvLSTM to predict the spatiotemporal attention, but they fail to predict a pixel-level attention due to the lack of mechanism to deconvolutional ConvLSTM modules. Moreover, there exists no attempt to fuse tri-modal information within ConvLSTM modules. For incorporating attention mechanism to dimensional emotion recognition, we consider spatiotemporal facial attention that selectively focuses on emotionally sailent parts by aggregating tri-modal facial videos.

III. Proposed Method

A. Problem Formulation and Overview

Formally, given a tri-modal facial video clip composed of three sequences, i.e., color recording sequences I, depth recording sequences D, and thermal recording sequences F, the objective of our approach is to recognize a dimensional emotion score (e.g., arousal or valence) $y \in [-1, 1]$ for each input $\{I, D, F\}$.

Concretely, to estimate human emotion for the tri-modal facial video clip, we adopt a strategy that frame-wise attention maps are first extracted. Attention-boosted features are then used for emotion recognition. We present a novel learnable network that implicitly estimates tri-modal recurrent attentions for the video. We formulate encoder-decoder module in the tri-modal recurrent attention network, where an encoder module consists of convolution layers to extract the features with spatial associations of each frame and a decoder module consists of GA-LSTM layers followed by sequential upsampling layers to estimate spatiotemporal attention cubes. To fuse tri-modal information within the recurrent network, the hidden states on each module are connected each other. We further build an emotion recognition network to estimate continuous emotion scores by leveraging 3D-CNNs to encode both spatial and temporal information simultaneously with tri-modal recurrent attention. The configuration of the overall framework is depicted in Fig. 2.

B. Network Architecture

In this section, we describe the details of tri-modal recurrent attention networks (TRAN) and its unimodal version. Since there is lack of supervision for spatio-temporal attentions of facial videos, we design TRAN in an end-to-end manner where the attention can be learned implicitly during learning the emotion recognition module with the supervision of a continuous emotion label only.

1) Spatial Encoder Networks: To extract features from each frame, we build the spatial encoder networks consisting of 2D convolutional layers and max-pooling layers. Since tri-modal input such as color, depth, and thermal have heterogeneous

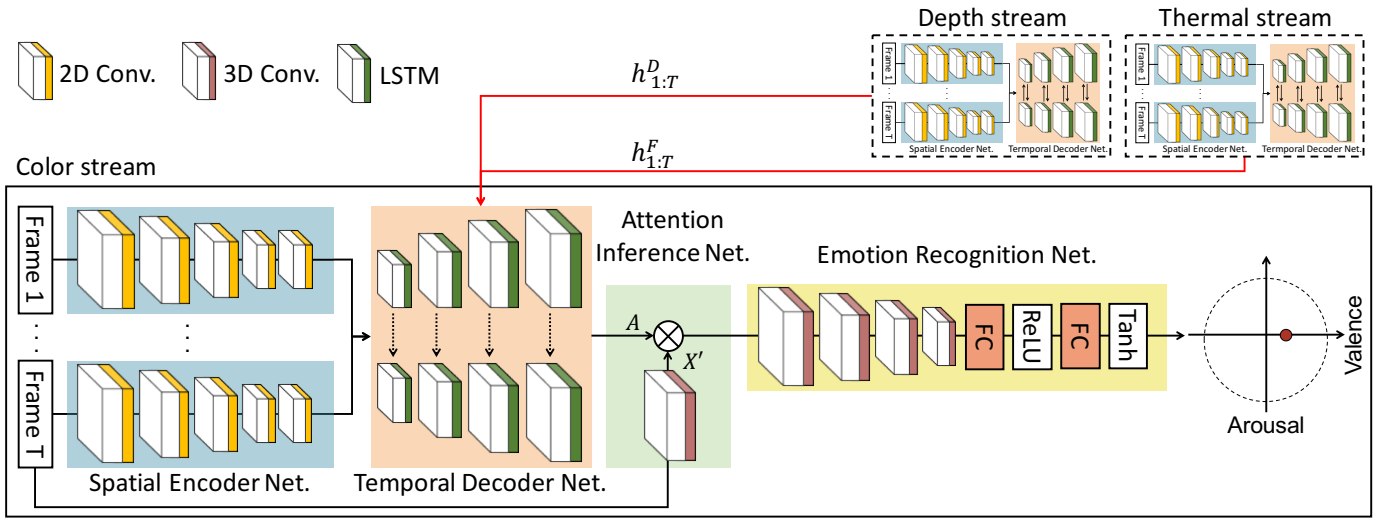


Fig. 2. The network configuration of TRAN which consists of four sub-networks, including spatial encoder networks, temporal decoder networks, attention inference networks, and emotion recognition networks. Given color, depth, and thermal facial videos, TRAN estimates the attention cubes and then recognizes output valence scores. Especially, red lines indicate temporal connection between GA-LSTM in the color stream and A-LSTM in depth and thermal streams. The detail of GA-LSTM module is illustrated in Fig. 3.

properties, we formulate the tri-modal encoder networks and temporally share the parameters of each network to extract common discriminative properties. Formally, we extract convolutional feature maps x^I, x^D, and x^F corresponding to color I, depth D, and thermal F in input sequences within the siamese network [64], where the weights and biases of each kernel are shared (i.e., replicated across all frames from same streams and updated together during training phase but not shared across spectral streams), enabling us to reduce the number of parameters and prevent the over-fitting problem.

The spatial encoder networks consist of successive 3×3 convolution layers and rectified linear unit (ReLU) layers, followed by max-pooling layers with stride 2×2.

2) Temporal Decoder Networks: We design the temporal decoder networks with stacked guided attention ConvLSTM (GA-LSTM) layers followed by upsampling layers that make the resolution of attention map same to the input features. Specifically, for input features x^I, x^D, and x^F from the spatial encoder networks, the temporal decoder networks predict a spatiotemporal attention cube corresponding the feature activation of color X^I to focus more relevant parts.

In depth and theraml streams, we build the temporal decoder networks with basic ConvLSTM modules [65], termed as attention LSTM (A-LSTM). Two guided streams have convolutional structures in both input-to-state and state-to-state transitions to maintain a spatial locality in the cell state while encoding the temporal correlation. Given input features x_{t-1} at time step $(t-1)$ from each stream, the ConvLSTM module updates as follows

$$\begin{aligned} i_t &= \sigma(w_{xi} * x_t + w_{hi} * h_{t-1} + w_{ci} * c_{t-1} + b_i), \\ f_t &= \sigma(w_{xf} * x_t + w_{hf} * h_{t-1} + w_{cf} * c_{t-1} + b_f), \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(w_{xc} * x_t + w_{hc} * h_{t-1} + b_c), \\ o_t &= \sigma(w_{xo} * x_t + w_{ho} * h_{t-1} + w_{co} \odot c_t + b_o), \\ h_t &= o_t \odot \tanh(c_t), \end{aligned} \tag{1}$$

where i_t, f_t, o_t, c_t and h_t represent the input gate, forget gate, output gate, cell activation, and cell output at time t, respectively. They are composed of 3D convolutional activations. $*$ denotes the convolution operator and $\odot$ denotes the Hadamard product. w is the filter connecting different gates, and b is the corresponding bias vector. The recurrent connections only operate over the temporal dimension, and use local convolutions to capture spatial context. However, original ConvLSTM module does not fully exploit the reciprocal information contained in the tri-modal videos.

Due to heterogeneous characteristics of tri-modal input, the direct fusion of these tri-modal input does not provide the optimal performance [54]. To fuse outputs of ConvLSTM modules across tri-modal streams with learnable modules, we thus extend existing ConvLSTM module in a way that the hidden states of each spectral stream are connected to guide the attention estimation in a boosting fashion. Specifically, given tri-modal features x^I_t, x^D_t, and x^F_t, the guided attention ConvLSTM (GA-LSTM) module updates at time step t as follows:

$$\begin{aligned} i_t &= \sigma(w_{xi} * x^I_t + \sum\nolimits_g w^g_{hi} * h^g_{t-1} + w_{ci} * c_{t-1} + b_i), \\ f_t &= \sigma(w_{xf} * x^I_t + \sum\nolimits_g w^g_{hf} * h^g_{t-1} + w_{cf} * c_{t-1} + b_f), \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(w_{xc} * x^I_t + \sum\nolimits_g w^g_{hc} * h^g_{t-1} + b_c), \\ o_t &= \sigma(w_{xo} * x^I_t + \sum\nolimits_g w^g_{ho} * h^g_{t-1} + w_{co} \odot c_t + b_o), \\ h^I_t &= o_t \odot \tanh(c_t), \end{aligned} \tag{2}$$

where $g \in \{I, D, F\}$ and h^g are hidden features from tri-modal streams, e.g., color, depth, and thermal streams. Namely, h^D and h^F are hidden features from A-LSTM modules and h^I is a hidden feature from GA-LSTM module. The key challenge in tri-modal recurrent attention networks is how to borrow comlementary information from each other. As hidden states in conventional LSTM modules are represented by taking the previous hidden states and current inputs, it can be leveraged as frame-level output [63]. Likewise, we use hidden states of depth and thermal streams as guidance

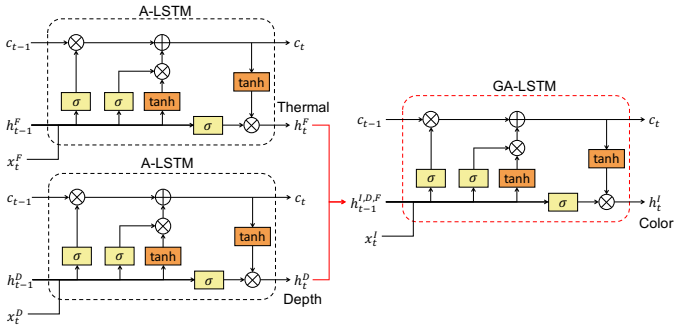


Fig. 3. Illustration of the proposed GA-LSTM module. Red lines indicate temporally additional connections. By considering the hidden activations in A-LSTM as extra inputs in GA-LSTM, we encourage TRAN to exploit attention map following temporal tri-modal guidances.

information for color stream to predict attention maps. To exploit the inter relationships between hidden representations from tri-modal input, we adopt weighted aggregating hidden states of other A-LSTM outputs to combine the information from each modality as shown in Fig. 3. Thus, the hidden representations from one time step of depth and thermal domain will be fed into target domain, *i.e.*color, at the next time step for refinement of attention.

The temporal decoder networks consist of successive 3×3 kernel in both A-LSTM and GA-LSTM modules and tanh [65]. Furthermore, we progressively enlarge the spatial resolution of stacked feature activations through sequential deconvolutions similar to [64]. We build the sequence of deconvolution with a factor of 2 after each LSTM module. Note that unlike other deconvolution layers as in [64], we utilize the proposed recurrent modules that encode the temporal correlation across inter-frames while preserving the spatial structure over sequences.

3) Attention Inference Networks: The tri-modal recurrent attentions are combined with input features as a soft attention in a manner that the attention is multiplied to feature activations of color frames. The attention maps obtained by the GA-LSTM module are normalized using spatial softmax function as follows:

$$A_{t,i} = \frac{\exp(H_{t,i})}{\sum_j \exp(H_{t,j})}, \tag{3}$$

where $H_{t,i}$ is the hidden state and $A_{t,i}$ is an attention cube for each location $i \in \{1, \cdots, H \times W\}$ and time step $t \in \{1, \cdots, T\}$.

4) Emotion Recognition Networks: We design emotion recognition networks to recognize final dimensional emotion states from attention-boosted features. Unlike exsting emotion recognition methods that consider the facial expression in a static image only [8], [36], we aim to simultaneously encode spatial and temporal cues. Specifically, by leveraging the multi-spectral recurrent attention $A_{1:T}$, our method produces attention boosted features for target modality, i.e., color. While the 2D-CNNs [36] can be used to predict the emotion for the facial video, it processes multiple input frames as different input channels independently, thus providing limited performances. To overcome this limitation, we employ the 3D-CNNs to deal with temporal information, which simultaneously consider spatial and temporal correlations across the input frames and directly regress the emotion.

TABLE I
NETWORK CONFIGURATION OF TRAN.

Spatial Encoder Networks				
Layer	Kernel	Ch I/O	Input	Output
conv1	3×3	3/32	I	conv1
pool1	2×2	32/32	conv1	pool1
conv2	3×3	32/64	pool1	conv2
pool2	2×2	64/64	conv2	pool2
conv3	3×3	64/128	pool2	conv3
pool3	2×2	128/128	conv3	pool3
Temporal Decoder Networks				
Layer	Kernel	Ch I/O	Input	Output
lstm1	3×3	128/64	pool3	lstm1
up1	2×2	64/64	lstm1	up1
lstm2	3×3	64/32	up1	lstm2
up2	2×2	32/32	lstm2	up2
lstm3	3×3	32/1	up2	A
Attention Inference Networks				
Layer	Kernel	Ch I/O	Input	Output
conv1	$3 \times 3 \times 3$	3/32	I	X'
Emotion Recognition Networks				
Layer	Kernel	Ch I/O	Input	Output
conv1	$3 \times 3 \times 3$	32/32	X''	conv1
pool1	$2 \times 2 \times 2$	32/32	conv1	pool1
conv2	$3 \times 3 \times 3$	32/64	pool1	conv2
pool2	$2 \times 2 \times 2$	64/64	conv2	pool2
conv3	$3 \times 3 \times 3$	64/128	pool2	conv3
pool3	$2 \times 2 \times 2$	128/128	conv3	pool3
conv4	$3 \times 3 \times 3$	256/256	pool3	conv4
pool4	$2 \times 2 \times 2$	256/256	conv4	pool4
fc1	–	9216/1024	pool4	fc1
fc2	–	1024/1	fc2	y

To elegantly incorporate the spatiotemporal attention to emotion recognition through 3D-CNNs, we extract convolutional feature activation X' using 3D convolutional layers for the color video I as an input. Then, we multiply spatiotemporal attention A to across the feature X' to estimate the attention-boosted feature activations as follows:

$$X'' = A \odot X', \tag{4}$$

where $\odot$ denotes the Hadamard product and X'' is a final refined feature map. Note that the pipeline for emotion recognition with the 3D-CNNs is inspired by the recognition networks in action recognition [66], because 3D-CNNs is well suited for spatiotemporal feature learning [66] owing to 3D convolution and 3D pooling layers.

By leveraging the attention-boosted feature cubes X'', our method then estimates a dimensional emotion scores y with 3D-CNNs [66] to simultaneously encode spatial and temporal information. Temporally stacked attentive feature cube pass the three 3D convolutional layers and 3D pooling layers which have $3 \times 3 \times 3$ kernels and $2 \times 2 \times 2$ kernels, respectively. Table I summarizes the overall network configuration of TRAN. The last fully-connected layer has a single output channels as f and we use a linear regression layer to estimate the output valence. We use tanh activation function followed by the last fully-connected layer that limits the range of output estimator to $[-1, 1]$.

5) Uni-modality Model: TRAN described so far can be simplified in a uni-modal framework, called Uni-modal Recurrent Attention Networks (URAN), to recognize human emotion from color recording videos only. In the networks,

all GA-LSTM modules are replaced with A-LSTM modules. Because our two types of models can be applied to various type of visual signals capturing facial expression, TRAN and URAN can be utilized for various environments.

C. Loss Function

During training, we minimize a mean squared error between estimated labels and given ground-truth labels. Given a collection of mini-batch M training sequences, a mean squared error criterion is adoped, defined as follows:

$$\mathcal{L} = \frac{1}{M}\sum_{m=1}^{M} \|\hat{y}_m - y_m\|_2, \tag{5}$$

where $\hat{y}_m$ and y_m are ground-truth valence label and prediction of the proposed method, respectively. TRAN is learned only with a ground-truth valence label as a supervision. Note that our method does not need explicit pre-defined AUs [67] and salient facial regions (e.g., facial landmarks). All parameters in TRAN can be implicitly learned using a stochastic gradient descent scheme.

IV. TAVER Benchmark

Most existing emotion recognition datasets [24], [25], [42], [45]–[47] have focused on the color image analysis, and thus they cannot be used for tri-modal emotion recogntion. In this section, we introduce a new benchmark for dimensional emotion recognition from tri-modal input such as color, depth, and thermal.

A. Data Acquisition

1) Recording System Setup and Synchronization: The data capture system included Microsoft Kinect v2 with time-of-flight sensor [1] and FLIR A65 thermal imaging temperature sensor as shown in Fig. 4. We used Microsoft Kinect v2 to obtain RGB and depth videos. It has been known that the Kinect v2 provides sharper and more detailed depth with high-quality color streams compared to the Kinect v1 with structured light method. The resolution of each color and depth streams were 1920×1080 and 512×424 pixels, respectively. We set a threshold of 0.5-7 meters due to inaccuracies in depth measurements at near and far ranges. The field of view (FoV) is $70° \times 60°$. Thus, we set the Kinect v2 camera far from 1 meters to subjects on a straight line as shown in Fig. 4. The thermal camera that we used is FLIR A65 thermal imaging temperature sensor. This camera captures thermal videos in resolution of 640×512 per frame with temperature range of -25 and $135°$C. The spectral range is $7.5 - 13\mu m$ and FoV is $45° \times 37°$. In order to better synchronize all sensors in our system, we set the capture rate of the thermal sensor to 10 fps. The thermal sensor stands next to the interviewer in a fixed position as shown in Fig. 4.

Note that the system synchronization is critical for data collection from various modality sensors. Since each sensor has its own machine to control, we developed a program to trigger the recording from the start to the end across all three sensors simultaneously. It is realized through the control of a master machine by sending a trigger signal to three sensors concurrently.

[1]https://developer.microsoft.com/en-us/windows/kinect.

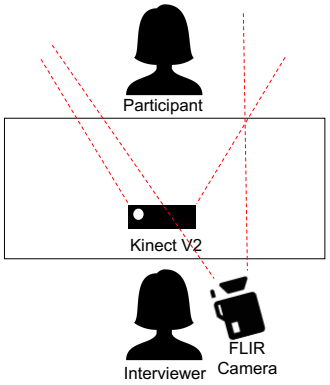


Fig. 4. Recording system setting used in TAVER to collect tri-modal data. We set up the Kinect v2 camera to record color and depth sequence data and FLIR camera to record thermal sequence data for TAVER benchmark.

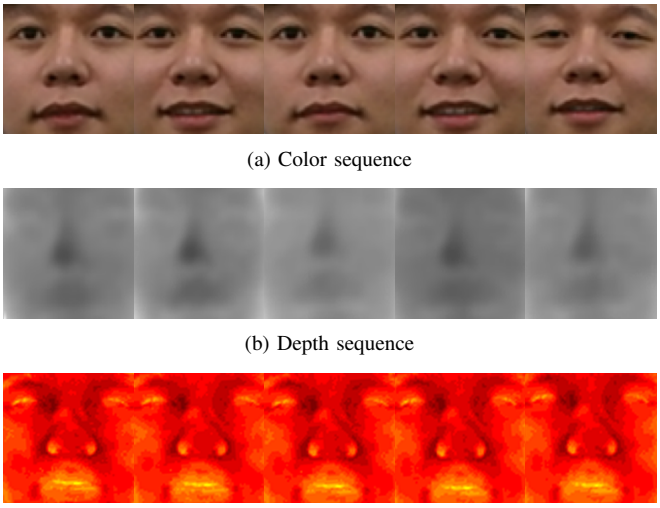
2) Participants: 100 subjects have been recruited to participate in data collection from which 46 subjects were recorded with a fully multi-spectral setting, and 17 subjects agreed to share their data. There are 7 males and 10 females, with ages ranging from 21 to 38 years old. All subjects have same mother languages as Korean. Following the IRB approved protocol, the informed consent form was signed by each subject before starting of data collection.

3) Emotion Elicitation: During data construction, we first showed relaxed videos in 10 minutes to subjects that make feel comfortable. We then composed an unannounced short interview in 5 minutes with subjects to interviewees and interviewers who use another language (English). When interviewers ask questions, people are embarrassed and stressed to answer the questions due to the inconvenience and burden of other languages. In their self-reports, subjects also said feel uncomfortable for the interviews with another language.

B. Data Pre-processing

1) Calibration: To calibrate color and depth streams, we used iai Kinect v2 libary [68]. We acquired several pieces of color, IR, and raw depth images containing a checkerboard pattern of 5×7. The distance between corners was set to 0.03m. Using the calibration toolbox provided in the iai kinect v2 libary [68], we estimated the shift parameter between IR and depth images and the projection matrix between IR and color images. The warping process consists of two parts. The depth map was warped into IR image coordinate using a shift parameter. The depth map on IR camera coordinate was finally projected into the RGB image coordinate. The library performed warping into the cropped RGB image coordinates. After projecting depth values to the RGB camera coordinate, we discarded the region exceeding the field of view of depth camera. Thus, the cropped color and depth images have 1408×792 resolutions.

2) Face Detection: To recognize the emotion from color recording videos, we first detected the human face in each color video frame using face and landmark detector in Dlib-ml [69], and then cropped the detected face region. We



(a) Color sequence

(b) Depth sequence

(c) Thermal sequencee

Fig. 5. Example of sample data sequence from a participant including color, depth, and thermal frames.

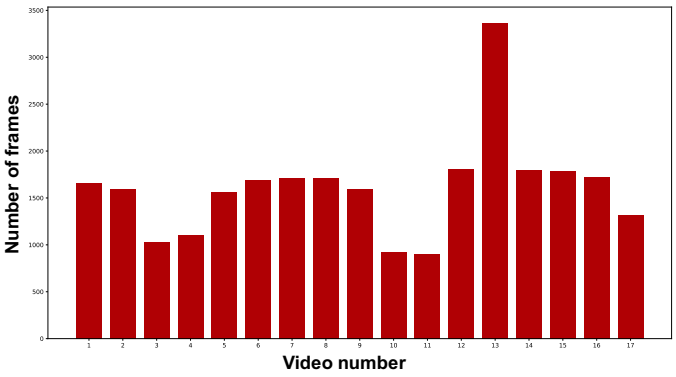


Fig. 6. Amount of frames in each video in TAVER benchmark.

then map the detected landmark points to pre-defined pixel locations in order to normalize the eye and nose coordinates between adjacent frames. For the depth stream, we warp the detected landmark points from color image to depth with calibration parameters. On the other hand, we use FLIR's ResearchIR software [2] to detect FLIR face region. Fig. 5 illustrates sample data sequences of tri-modal from a subject.

C. Annotation

We modified a web-based annotation interface [41] to annotate affective dimensions [3]. The definition of valence and arousal dimensions was adapted from [16]. We hired 6 annotators aged between 20 and 25. 3 annotators were assigned to each video sequence for more accurate annotations. The annotators were instructed to simultaneously and time-continuously consider the intensity of valence and arousal during the annotation. The two affective dimensions (arousal and valence) were annotated using a slider with values ranging from -10 to 10 and a step of 1. Each annotator was instructed orally and received instructions with a 3 pages document explaining in details the procedure and including some examples of annotations to follow for the annotation task. In order to

[2]https://www.flir.com/discover/rd-science/matlab/

[3]https://github.com/JeanKossaifi/valence_arousal_annotator/

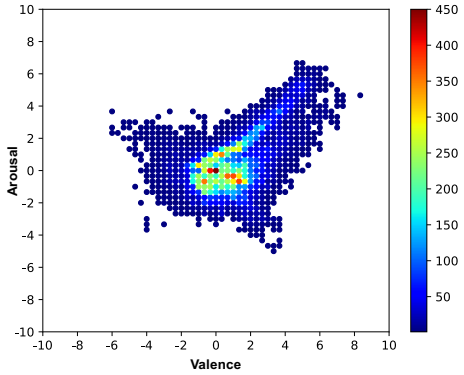


Fig. 7. Distribution of arousal and valence scores in TAVER benchmark.

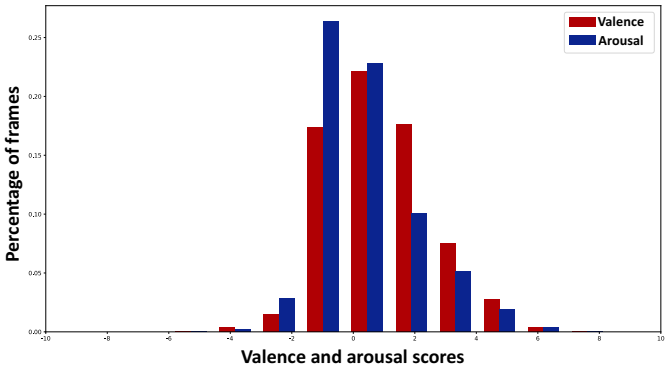


Fig. 8. Histogram of arousal and valence scores in TAVER benchmark.

deal with issue of missing values in the annotations, data were interpolated using 1D bilinear interpolation. Finally, the ground truth of a sequence was estimated by mean filtering the annotations provided by all the 3 annotators.

D. Analysis

In total, we annotated more than 27K frames with frame levels valence and arousal intensities in the range of -10 to 10. The number of frames at each video is shown in Fig. 6. In Fig. 7, we show the distribution of the values of arousal and valence in the TAVER benchmark as well as histogram of arousal and valence in Fig. 8. We compared TAVER benchmark with other public datasets for dimensional emotion recognition such as RECOLA [24], SEWA [25], AFEW-VA [41], BP4D [19], and BP4D+ [21]. Even if the datasets for multi-spectral facial expression analysis including BP4D and BP4D+ collected large-scale 3D facial models with Di3D dynamic imaging system in Table II, they annotated AUs and discrete emotion categories not including arousal and valence scores. Compared with those datasets, we collect the videos from participants without artificial acting for more natural emotion elicitation.

For the analysis of the annotation of the affective behaviors, we computed the MSE, the mean correlation coefficient and the Cronbach's α [24] in Table III. The Cronbach's α is an estimate of the internal consistency between annotations; $\alpha > 0.7$ is considered as an acceptable internal consistency and $\alpha > 0.8$ is considered as a good consistency. Results from the raw data show that their internal consistency is acceptable for valence and arousal after zero-mean normalization.

TABLE II
COMPARISON OF TAVER WITH EXISTING EMOTION RECOGNITION DATASETS. ALTHOUGH TAVER HAS LITTLE SUBJECTS, IT CONTAINS COLOR, DEPTH, AND THERMAL VIDEOS FOR TRI-MODAL EMOTION RECOGNITION.

Database	Subjects	Annotation type	Amount of data	Ellicitation method	Environment	Illumination	Data type
RECOLA [24]	27 participants	Dimensional	27 videos of 5 mn.	Online interactions	Controlled	Controlled	Color
SEWA [25]	84 participants	Dimensional	300 videos of 6s to 4mn.	Human-computer interaction	webcam	Indoor-In-the-Wild	Color
AFEW-VA [41]	240 subjects	Dimensional	600 video clips.	Movie actors	Indoor-In-the-Wild	Indoor-In-the-Wild	Color
BP4D [19]	41 subjects	Categorical	328 videos of 1-4 mn.	Actors	Controlled	Controlled	Color + Depth
BP4D+ [21]	140 subjects	Categorical	1400 videos of 1-2 mn.	Actors	Controlled	Controlled	Color + Depth + Thermal
TAVER	17 participants	Dimensional	17 videos of 1-4 mn.	Human-human interaction	Controlled	Controlled	Color + Depth + Thermal

TABLE III
STATISTICS OF THE CONTINUOUS EMOTION AFTER APPLYING ZERO MEAN NORMALIZATION. % POS. MEANS PERCENTAGE OF POSITIVE FRAMES.

Dimension	% pos.	Corr.	α
Arousal	45.4	0.424	0.72
Valence	46.9	0.479	0.79

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present a detailed analysis and evaluation of our approach on dimensional emotion recognition. Specifically, we firstly evaluated the influence of our proposed method with the TAVER benchmark with ablation evaluations, with respect to 1) various combination of different modalities such as color, depth, and thermal, 2) the proposed sub-networks, and 3) length of the clips. Note that we reformulate tri-modal recurrent attention network (TRAN) to the uni-modal recurrent attention network (URAN) to compare our method with the state-of-the-art methods on two publicly available benchmark datasets performing in the color information only.

A. Implementation Details

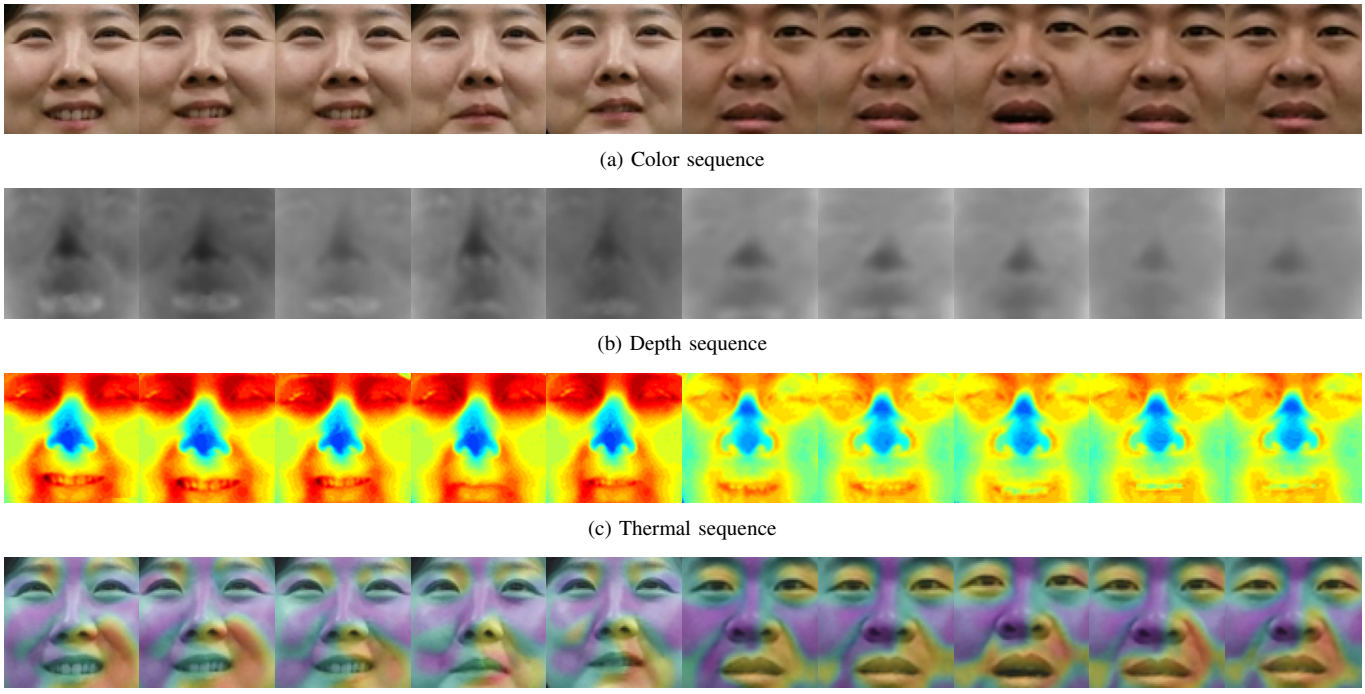
We implemented our network using the PyTorch library [70]. To reduce the effect of overfitting, we employed the dropout scheme with the ratio of 0.5 between fully-connected layers, and data augmentation schemes such as flips, contrast, and color changes. The videos in the training set were split into non-overlapped 16-frame clips, and thus the input of model has a frame rate of 4 fps. For optimization, we choose Adam [71] due to its faster convergence than standard stochastic gradient descent with momentum. For tri-modal emotion recognition, we trained TRAN from scratch using mini-batches of 4 clips, with initial learning rate as $\lambda = 1e-4$. Meanwhile, we also trained URAN from scratch with mini-batches of 8 clips and initial learning rate as $\lambda = 1e-4$ for the comparison with subsets of RECOLA [16] and SEWA [25] benchmarks. The filter weights of each layer were initialized by Xavier distribution, which was proposed by Glorot and Bengio [72], due to its properly scaled uniform distribution for initialization. In the SEWA and RECOLA datasets, we detected the face in each video frame using face and landmark detector in Dlib-ml [69], and then cropped the detected face region for all database. We then mapped the detected landmark points to pre-defined pixel locations in order to normalize the eye and nose coordinates between adjacent frames to recognize the emotion from a facial video.

B. Experimental Settings

For baseline models, we reported the results of the VGG-16 [73] and ResNet-50 [74] networks pre-trained on the ImageNet dataset [75]. We also considered the VGG-Face network pre-trained on VGG-Face dataset [76]. In order to consider the temporal information between the frames, we extended the VGG-Face network to the CNN-LSTM model, which consists of one fully-connected layer and two hidden layers with 128 units similar to [42]. In the following, we evaluated the proposed method in comparison to the baseline approach such as AV+EC challenge baseline methods [16], [18]. Several deep CNNs-based approaches were also compared, such as Chen *et al.* (LGBP-TOP + LSTM) [33], He *et al.* (LGBP-TOP + Bi-Dir. LSTM) [37], Chao *et al.* (LGBP-TOP + LSTM + ϵ-loss and CNN + LSTM + ϵ-loss) [34], and Khorrami *et al.* (CNN+RNN) [36] with RECOLA dataset. We reimplemented methods of [36] and evaluated on the SEWA and TAVER datasets. Moreover, we reimplemented AffWild-Net [42] to compare in TAVER dataset. For all the investigated methods, we interpolated the valence scores from adjacent frames related to dropped frames that the face detector missed. In addition, following the AV+EC's post-processing procedure of predictions [16], [48], we applied the same chain of post-processing on the obtained predictions; smoothing, centering and scaling except time-shifting.

1) Datasets: In experiments, we used the proposed TAVER dataset splitted into 12 training and 5 test videos. Furthermore, we also used RECOLA dataset [24] and SEWA dataset [25] used in AV+EC 2015 [16] and AV+EC 2017 [18] challenges, respectively.

The AV+EC 2015 challenge used the subset of RECOLA dataset [24], which was recorded for 27 French-speaking subjects. The dataset contains two types of continuous labels, arousal and valence, which were manually annotated by six annotators. Each continuous emotion label ranges from $[-1, 1]$. Raw interview video frame has 1080×1920 resolution and



(a) Color sequence

(b) Depth sequence

(c) Thermal sequence

(d) Attention score

Fig. 9. Visualization of spatiotemporal attention maps learned by TRAN for two subjects in TAVER benchmark: Attention score is normalized by the spatial softmax. Red indicates higher weight of the frame and blue indicates lower weight. Specifically, the areas around eyes and mouth are considered to be important to estimate emotion.

16 fps. Since the test set labels were not readily available, we evaluate all of our experiments on the development set.

Compared to RECOLA dataset, the subset of SEWA dataset [25] [4] used in the AV+EC 2017 challenge was acquired in various places such as home and work place with diverse personal equipments such as webcams and microphones. The dataset contains three types of continuous labels such as arousal, valence and liking, which were manually annotated by six annotators. Thus, it is more challenging and tailors to real-life applications of affective computing technologies than RECOLA dataset.

2) Metrics: For quantitative evaluation, we computed three metrics: (i) Root Mean Square Error (RMSE), (ii) Pearson Correlation Coefficient (CC), and (iii) Concordance Correlation Coefficient (CCC) as used in [36]. First of all, RMSE is the most common evaluation metric in a continuous domain which is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}^i - y^i)^2} \tag{6}$$

where $\hat{y}^i$ and y^i are the ground truth and prediction of the i th sample, and n is the number of samples in the evaluation set. Note that RMSE-based evaluation can heavily weigh the outliers [18], and thus it is not able to provide the covariance of prediction and ground-truth to show how they change with respect to each other. Pearson correlation coefficient (CC) is therefore proposed in [18] to overcome this limitation:

$$\rho = \frac{COV(\hat{y}, y)}{\sigma_{\hat{y}}\sigma_y} = \frac{E[(\hat{y} - \mu_{\hat{y}})(y - \mu_y)]}{\sigma_{\hat{y}}\sigma_y}, \tag{7}$$

[4] http://sewaproject.eu

TABLE IV
EFFECTIVENESS OF TRI-MODAL INPUT FOR DIMENSIONAL EMOTION RECOGNITION. TRAN IS TRAINED ON THE TRAINING AND VALIDATION SETS AND EVALUATED WITH THE TEST SET ON THE TAVER BENCHMARK.

Color	Depth	FIR	RMSE	CC	CCC
✓			0.120	0.481	0.446
✓	✓		0.117	0.501	0.482
✓		✓	0.114	0.546	0.499
✓	✓	✓	0.112	0.563	0.521

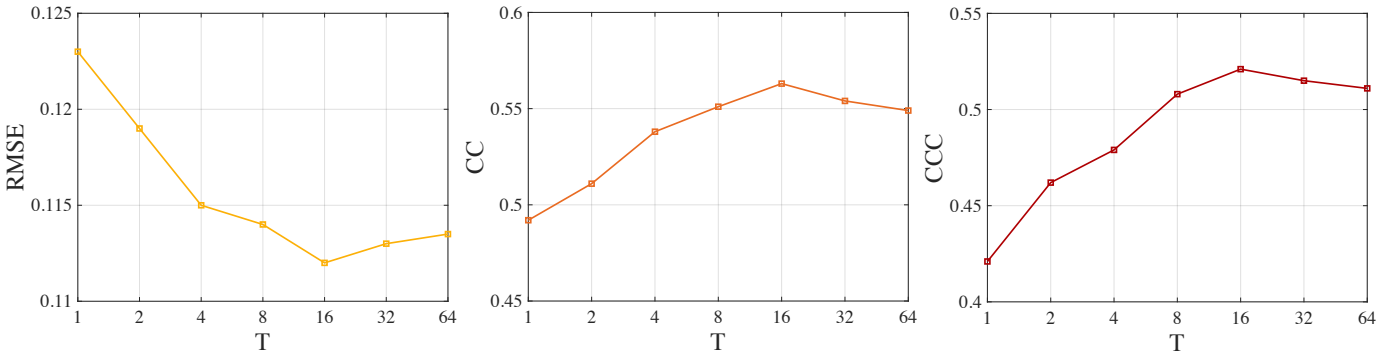
where ρ indicates the Pearson correlation coefficient, σ_x^2 and σ_y^2 are the variances of the predicted and ground truth values, and μ_x and μ_y are their means, respectively. Especially, the CCC tries to measure the agreement between two variables using the following expression:

$$\rho_c = \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \tag{8}$$

where ρ_c indicates the concordance correlation coefficient. Unlike CC, the predictions that are well correlated with the ground-truth but shifted in value are penalized in proportion to the deviation in CCC. The highest CC and CCC values thus represent the best recognition performance.

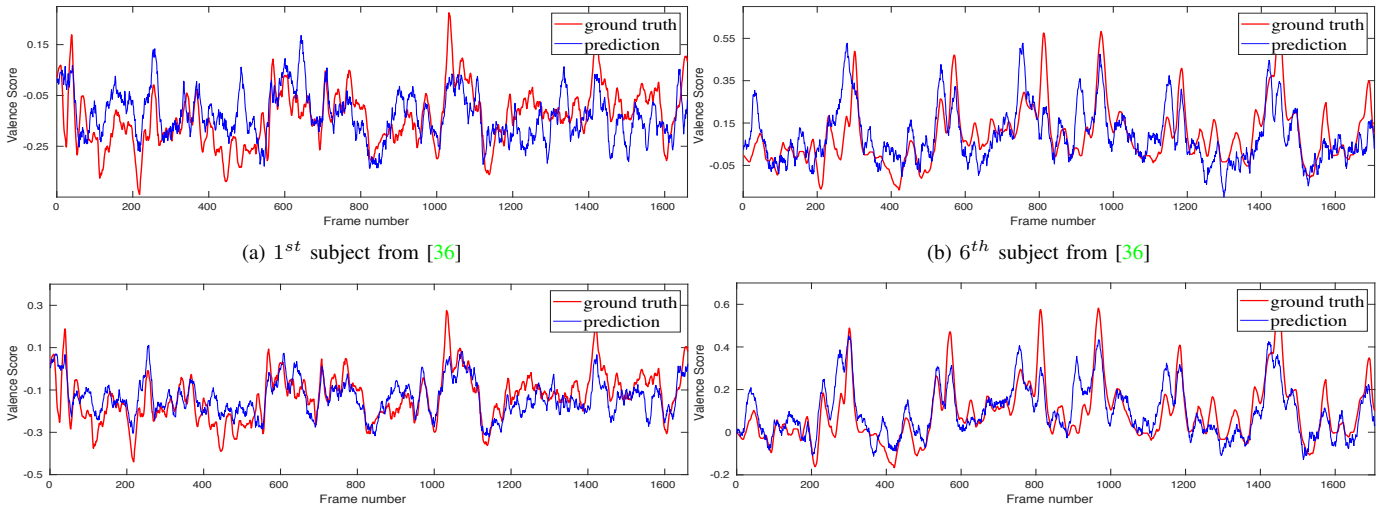
C. Results on TAVER Benchmark

1) Analysis on Tri-modal Input: To verify the effects of the tri-modal input to estimate dimensional emotion, we analyzed the performance of each modality in Table IV. We set up the performance using only color videos as baseline performance. By leveraging depth videos, the estimation performances improve 0.02 and 0.036 for CC and CCC scores compared to the baseline. In respect to thermal videos, the estimation



(a) RMSE (b) CC (c) CCC

Fig. 10. Ablation study of TRAN for various length of clips on TAVER benchmark. When we set T as 16, it shows best performance. In the remaining experiments, we use $T = 16$ frames as length of input sequence.



(a) 1^{st} subject from [36] (b) 6^{th} subject from [36]

(c) 1^{st} subject from URAN (d) 6^{th} subject from URAN

Fig. 11. Estimated valence graph of two subjects in the development set on TAVER benchmark with [36] and URAN. The x-axis is the number of frames detected in face detector, and the y-axis is the valence score. Red line is ground truth labels and blue line is estimated scores. Note that these graphs are not interpolated at the dropped frames.

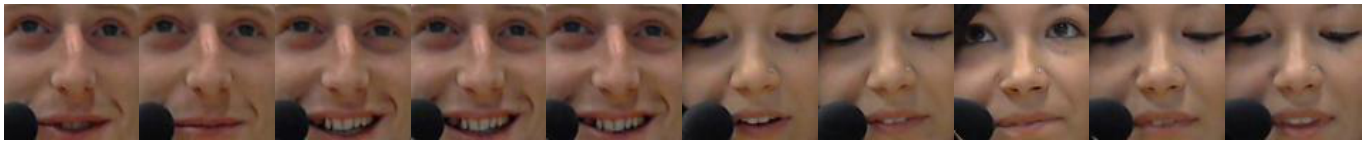
TABLE V
ABLATION STUDY OF TRAN ON TAVER BENCHMARK. WITH ATTENTION NETWORKS, WE ACHIEVED THE BEST RESULT IN ALL MEASUREMENTS.

Method	RMSE	CC	CCC
TRAN (w.o./Attention)	0.116	0.504	0.473
TRAN	0.112	0.563	0.521

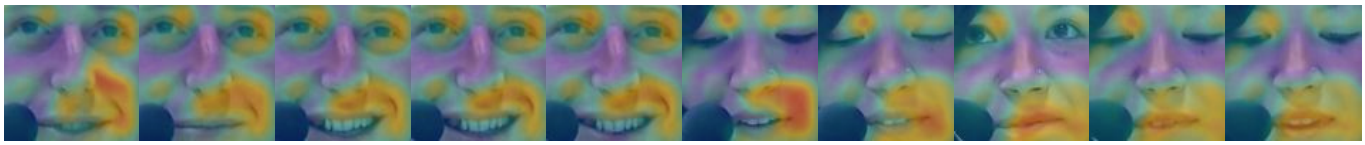
performances were 0.065 and 0.053 higher for CC and CCC scores than the baseline. When we used all the color, depth, and thermal videos for learning the networks, the estimation performances were 0.082 and 0.075 higher for CC and CCC scores than the baseline which shows the robustness of using tri-modal input to estimate the dimensional emotion. The usage of tri-modal input also showed the lowest value 0.112 for RMSE score. Although the bi-modal input improves the recognition ability, the full usage of tri-modal input including color, depth, and thermal shows the best performance. Note that the A-LSTM module was used for input of single modality instead of the GA-LSTM module.

2) The Effects of Attention Inference: In Table V, we evaluated the effects of attention inference modules. For this experiment, we removed the attention networks in the proposed TRAN, and fed the 3D convolutional feature activations into emotion recognition networks. To verify the effectiveness of the attention to estimate dimensional emotions, we visualized the normalized attention maps where model focused on parts of the face, while improving the emotion recognition performance. As shown in Fig. 9, the proposed model effectively learns the important parts in consecutive frames in same subjects, especially eyes and mouth. At different frames, the proposed model captures different parts, since GA-LSTM deals with spatiotemporal correspondence. As a result, the proposed attention cube highlights salient parts of emotion recognition and implicitly learn to detect specific AUs in facial images.

3) The Effects of the Number of Frames: In Fig. 10, we estimated RMSE, CC and CCC scores for TRAN on the TAVER with respect to various length of clip. Overall, CC and CCC scores increase with the number of frames until 16 frames. However, CC and CCC scores decrease after 16 frames. In addition, RMSE was also decreased after 16 frames, which means that the overlength of clip decreases the performance. Thus, we used the 16 frames as length of clip for other experiments.



(a) Color sequence



(b) Attention score

Fig. 12. Visualization of spatiotemporal attention maps learned by the proposed network for two subjects in RECOLA benchmark [24]: Attention score is normalized by the spatial softmax. Red indicates higher weight of the frame and blue indicates lower weight. Specifically, the areas around eyes and mouth are considered to be important to estimate emotion.

TABLE VI
QUANTITATIVE EVALUATION OF URAN AND TRAN FOR THE PREDICTED VALENCE ON TAVER BENCHMARK. NOTE THAT WE USED COLOR VIDEOS IN TAVER BENCHMARK FOR COMPARISON TO BASELINE METHODS.

Architectures	Method	RMSE	CC	CCC
CNN	CNN [36]	0.132	0.421	0.387
	VGG-16 [73]	0.131	0.426	0.395
	ResNet-50 [74]	0.128	0.443	0.409
	VGG-Face [76]	0.125	0.478	0.451
CNN+RNN	CNN + RNN ($\approx$ 4 sec.) [36]	0.127	0.458	0.413
	VGG-Face-LSTM [76]	0.124	0.482	0.457
	AffWildNet [42]	0.123	0.499	0.468
	URAN	0.120	0.481	0.446

4) Comparison to Other Methods: Table VI summarizes the RMSE, CC and CCC values obtained when applying all the developed CNN-based architectures including VGG-16, ResNet-50 and VGG-Face, and CNN-RNN architectures including VGG-Face-LSTM, Khorrami *et al.* [36] and AffWildNet [42]. Our proposed method provides state-of-the-art performance with the same length of clip, which means that attention mechanism in URAN improves the performance. We trained all the methods with color videos in TAVER dataset, then compared to color stream with TRAN. Note that we reimplemented the methods of AffWildNet [42] and Khorrami *et al.* [36] with PyTorch library to compare with our method. In Fig. 11, we compared the estimated valence graph from [36] and TRAN with color, which show that TRAN outperforms [36].

D. Results on Other Benchmarks

In the following, we evaluated the proposed network through comparisons to state-of-the-art CNNs-based approaches [33], [34], [36], [37] on the RECOLA dataset [24], which has been adopted for the AudioVisual Emotion recognition Challenges (AV+EC) in 2015 [16] and 2016 [48]. We also compared the proposed method to the state-of-the-art on the subset of SEWA dataset [25] used in AV+EC in 2017 [18]. Because all the RECOLA and SEWA benchmarks are composed of only color recording facial videos, we reformulated tri-modal recurrent attention network (TRAN) to the uni-modal recurrent attention network (URAN), which replaced the proposed GA-LSTM modules to simple A-LSTM modules for this comparison.

TABLE VII
QUANTITATIVE EVALUATION OF URAN FOR THE PREDICTED VALENCE ON THE RECOLA DATASET [24]. WE DENOTE A RESULT OF BASELINE METHOD FROM AVEC'15 CHALLENGE [16].

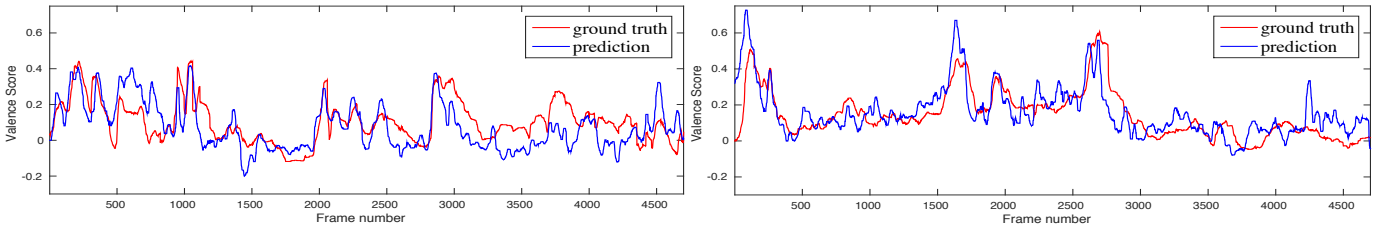
Method	RMSE	CC	CCC
Baseline [16]	0.117	0.358	0.273
CNN [36]	0.113	0.426	0.326
CNN + RNN ($\approx$ 1 sec.) [36]	0.111	0.501	0.474
CNN + RNN ($\approx$ 4 sec.) [36]	0.108	0.544	0.506
LGBP-TOP + LSTM [33]	0.114	0.430	0.354
LGBP-TOP + Bi-Dir. LSTM [37]	0.105	0.501	0.346
LGBP-TOP + LSTM + ϵ-loss [34]	0.121	0.488	0.463
CNN + LSTM + ϵ-loss [34]	0.116	0.561	0.538
URAN	0.102	0.572	0.546

TABLE VIII
QUANTITATIVE EVALUATION OF URAN FOR THE PREDICTED VALENCE ON THE SEWA BENCHMARK [18]. WE DENOTE A RESULT OF BASELINE METHOD FROM AVEC'17 CHALLENGE [18].

Method	RMSE	CC	CCC
Baseline [18]	-	-	0.400
CNN [36]	0.114	0.564	0.528
CNN + RNN ($\approx$ 4 sec.) [36]	0.104	0.616	0.588
URAN	0.099	0.638	0.612

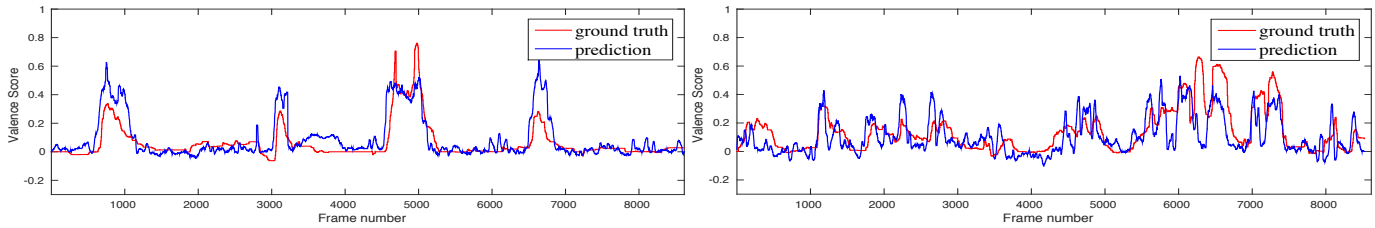
We compared URAN with the state-of-the-art methods such as CNN-based approaches [36] and LSTM-based approaches [34] on the subset of RECOLA dataset [24] in Table VII. The results showed that the proposed method exhibits a better recognition performance than conventional methods [33], [34], [36], [37]. We also visualize the spatiotemporal attention maps obtained by URAN in the RECOLA dataset in Fig. 12. Although we trained URAN without guidance of depth and thermal recording videos, our attention network found discriminative parts well in face owing to spatial and temporal encoder-decoder architecture.

In Table VIII, we also compared our method with the RNN-based approach [36] on the subset of SEWA dataset [25], which includes 34 training and 14 development videos. The results have also shown that the proposed method exhibits a better recognition performance compared to the conventional methods. We also visualize the valence scores predicted by the proposed method for two subjects of RECOLA and SEWA datasets in Fig. 13 and Fig. 14, respectively. The proposed models can detect the valence score especially on the peak points by demonstrating the effects of URAN.



(a) 5^{th} subject (b) 8^{th} subject

Fig. 13. Estimated valence graph of subjects in the development set on RECOLA benchmark [16] The x-axis is the number of frames detected in face detector, and the y-axis is the valence score. Red line is ground truth labels and blue line is estimated scores. Note that these graphs are not interpolated at the dropped frames.



(a) 4^{th} subject (b) 10^{th} subject

Fig. 14. Estimated valence graph of subjects in the development set on SEWA benchmark [25]. The x-axis is the number of frames detected in face detector, and the y-axis is the valence score. Red line is ground truth labels and blue line is estimated scores. Note that these graphs are not interpolated at the dropped frames.

VI. Conclusion

In this paper, we presented TRAN for dimensional emotion recognition by jointly utilizing tri-modal color, depth, and thermal recording videos. The key idea of this approach is to combine heterogeneous modality within unified deep networks, where discriminative and salient parts of faces were implicitly detected to boost the recognition accuracy. TRAN estimated the attentive region of temporally varying human face and the continuous emotion score effectively by leveraging 3D-CNNs. Moreover, our unified framework was implicitly learned to estimate the attention in face videos without any pixel-level annotations. We also introduced TAVER benchmark that is more robust in a variety of environments such as illumination or skin color. An extensive experimental analysis showed the benefits of TRAN for tri-modal dimensional emotion recognition on TAVER benchmark and URAN achieves state-of-the-art emotion recognition performances on both RECOLA and SEWA benchmarks. We believe that the results of this study will facilitate further advances in tri-modal emotion recognition and its related tasks.

References

[1] C. Lisetti, F. Nasoz, C. LeRouge, O. Ozyer, and K. Alvarez, "Developing multimodal intelligent affective interfaces for tele-home health care," *IJHCS*, 2003. 1

[2] S. D'Mello, R. Picard, and A. Graesser, "Toward an affect-sensitive autotutor," *Int. Systems*, 2007. 1

[3] G. Yannakakis and J. Togelius, "Experience-driven procedural content generation," *IEEE Trans. AC*, 2011. 1

[4] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *IVC*, vol. 31, no. 2, pp. 120–136, 2013. 1

[5] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," *In: CVPR*, 2005. 1

[6] ——, "Fully automatic facial action recognition in spontaneous behavior," *In: FG*, 2006. 1

[7] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. PAMI*, vol. 29, no. 6, pp. 915–928, 2007. 1

[8] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" *In: ICCV Workshop*, 2015. 1, 2, 5

[9] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, and R. C. Ferrari, "Combining modality specific deep neural networks for emotion recognition in video," *In: ICMI*, 2013. 1

[10] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. IP*, vol. 26, no. 9, pp. 4193–4203, 2017. 1

[11] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. IP*, vol. 28, no. 1, pp. 356–370, 2018. 1, 2

[12] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Research Personal.*, vol. 11, no. 3, pp. 273–294, 1977. 1, 2

[13] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Trans. AC*, vol. 7, no. 1, pp. 17–28, 2016. 1, 2

[14] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," *In: CVPR*, 2017. 1, 2

[15] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *PRL*, vol. 66, pp. 22–30, 2015. 1, 2

[16] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "Avec 2015: The 5th international audio/visual emotion challenge and workshop," *In: Multimedia*, 2015. 1, 3, 7, 8, 11, 12

[17] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," *In: ICMI*, 2012. 1

[18] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmi, and M. Pantic, "Avec 2017–real-life depression, and a ect recognition workshop and challenge," 2017. 1, 3, 8, 9, 11

[19] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *IVC*, vol. 32, no. 10, pp. 692–706, 2014. 1, 3, 7, 8

[20] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. V. Gool, "A 3-d audio-visual corpus of affective communication," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 591–598, 2010. 1, 3

[21] Z. Zhang, J. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," *In: CVPR*, 2016. 1, 2, 3, 7, 8

[22] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 682–691, 2010. 1

[23] P. Liu and L. Yin, "Spontaneous facial expression analysis based on temperature changes and head motions," *In: FG*, 2015. 1

[24] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," *In: FG*, 2013. 2, 6, 7, 8, 11

[25] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *arXiv preprint arXiv:1901.02839*, 2019. 2, 6, 7, 8, 9, 11, 12

[26] J. Lee, S. Kim, S. Kim, and K. Sohn, "Spatiotemporal attention based deep neural networks for emotion recognition," *In: ICASSP*, 2018. 2

[27] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, and P. E. Ricci-Bitti, "Universals and cultural differences in the judgments of facial expressions of emotion." *J. Personal. Social Psychology*, vol. 53, no. 4, p. 712, 1987. 2

[28] P. Ekman, "Strong evidence for universals in facial expressions: a reply to russell's mistaken critique." 1994. 2

[29] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *In: CVPR*, 2015. 2

[30] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," *In: CVPR*, 2013. 2

[31] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," *In: CVPR*, 2014. 2

[32] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, "Time-delay neural network for continuous emotional dimension prediction from facial expression sequences," *IEEE Trans. on Cyb.*, vol. 46, no. 4, pp. 916–929, 2016. 2

[33] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," *In: AVEC*, 2015. 2, 8, 11

[34] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," *In: AVEC*, 2015. 2, 8, 11

[35] K. S. Ebrahimi, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," *In: ICMI*, 2015. 2

[36] P. Khorrami, T. L. Paine, K. Brady, C. Dagli, and T. S. Huang, "How deep neural networks can improve emotion recognition on video data," *In: ICIP*, 2016. 2, 5, 8, 9, 10, 11

[37] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," *In: AVEC*, 2015. 2, 8, 11

[38] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," *In: ICMI*, 2012. 2

[39] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," *In: FG*, 1998. 2

[40] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," *In: CVPR Workshop*, 2010. 2

[41] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," *IVC*, vol. 65, pp. 23–36, 2017. 2, 7, 8

[42] D. Kollias, P. Tzirakis, M. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *IJCV*, pp. 1–23, 2019. 2, 6, 8, 11

[43] S. Mavadati, M. Mahoor, K. Bartlett, P. Trinh, and J. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Trans. AC*, vol. 4, no. 2, pp. 151–160, 2013. 2

[44] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected," *In: CVPR Workshops*, 2013. 2

[45] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The belfast induced natural emotion database," *IEEE Trans. AC*, vol. 3, no. 1, pp. 32–41, 2012. 2, 6

[46] M. Bradley and P. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *JBTEP*, vol. 25, no. 1, pp. 49–59, 1994. 2, 3, 6

[47] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. AC*, vol. 3, no. 1, pp. 5–17, 2012. 2, 6

[48] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," *In: AVEC*, 2016. 3, 8, 11

[49] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3d facial expression database for facial behavior research," *In: FG*, 2006. 3

[50] C. Feng, S. Zhuo, X. Zhang, L. Shen, and S. Süsstrunk, "Near-infrared guided color image dehazing," *In: ICIP*, 2013. 3

[51] H. Choi, S. Kim, K. Park, and K. Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks," *In: ICPR*, 2016. 3

[52] Y. Kim, H. Jung, D. Min, and K. Sohn, "Deeply aggregated alternating minimization for image restoration," *In: CVPR*, 2017. 3

[53] K. Park, S. Kim, and K. Sohn, "Unified multi-spectral pedestrian detection based on probabilistic fusion networks," *Pattern Recognition*, vol. 80, pp. 143–155, 2018. 3

[54] C. Palmero, A. Clapés, C. Bahnsen, A. Møgelmose, T. Moeslund, and S. Escalera, "Multi-modal rgb–depth–thermal human body segmentation," *IJCV*, vol. 118, no. 2, pp. 217–239, 2016. 3, 4

[55] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998. 3

[56] H. Larochelle and G. Hinton, "Learning to combine foveal glimpses with a third-order boltzmann machine," *In: NeurIPS*, 2010. 3

[57] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," *In: CVPR*, 2017. 3

[58] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *In: CVPR*, 2016. 3

[59] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *In: CVPR*, 2018. 3

[60] S. Woo, J. Park, J. Lee, and S. In, "Cbam: Convolutional block attention module," *In: ECCV*, 2018. 3

[61] S. Sharma, R. Kiros, and r. Salakhutdinov, "Action recognition using visual attention," *arXiv:1511.04119*, 2015. 3

[62] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," *In: ICCV*, 2015. 3

[63] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. Snoek, "Videolstm convolves, attends and flows for action recognition," *CVIU*, vol. 166, pp. 41–50, 2018. 3, 4

[64] J. Lee, H. Jung, Y. Kim, and K. Sohn, "Automatic 2d-to-3d conversion using multi-scale deep neural network," *In: ICIP*, 2017. 4, 5

[65] S. Xingjian, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *In: NeurIPS*, 2015. 4, 5

[66] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. PAMI*, vol. 35, no. 1, pp. 221–231, 2013. 5

[67] P. Ekman and W. V. Friesen, "Facial action coding system," 1977. 6

[68] T.Wiedemeyer, "IAI Kinect2," https://github.com/code-iai/iai_kinect2, Institute for Artificial Intelligence, 2015. 6

[69] D. E. King, "Dlib-ml: A machine learning toolkit," *JMLR*, vol. 10, no. Jul, pp. 1755–1758, 2009. 6, 8

[70] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017. 8

[71] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014. 8

[72] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *In: AIStats*, 2010. 8

[73] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 8, 11

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *In: CVPR*, 2016. 8, 11

[75] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *In: CVPR*, 2009. 8

[76] O. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition," *In: BMVC*, 2015. 8, 11