An End-to-End Visual-Audio Attention Network for Emotion Recognition in User-Generated Videos

Sicheng Zhao,[1#] **Yunsheng Ma,**[2,3#] **Yang Gu,**[2] **Jufeng Yang,**[4*]
Tengfei Xing,[2] **Pengfei Xu,**[2] **Runbo Hu,**[2] **Hua Chai,**[2] **Kurt Keutzer**[1]
[1]University of California, Berkeley, USA [2]Didi Chuxing, China
[3]Harbin Institute of Technology, Weihai, China [4]Nankai University, China
{schzhao, yunsheng.ma98}@gmail.com, yangjufeng@nankai.edu.cn
{guyangdavid, xingtengfei, xupengfeipf, hurunbo, chaihua}@didiglobal.com, keutzer@berkeley.edu

Abstract

Emotion recognition in user-generated videos plays an important role in human-centered computing. Existing methods mainly employ traditional two-stage shallow pipeline, *i.e.* extracting visual and/or audio features and training classifiers. In this paper, we propose to recognize video emotions in an end-to-end manner based on convolutional neural networks (CNNs). Specifically, we develop a deep Visual-Audio Attention Network (VAANet), a novel architecture that integrates spatial, channel-wise, and temporal attentions into a visual 3D CNN and temporal attentions into an audio 2D CNN. Further, we design a special classification loss, *i.e.* polarity-consistent cross-entropy loss, based on the polarity-emotion hierarchy constraint to guide the attention generation. Extensive experiments conducted on the challenging VideoEmotion-8 and Ekman-6 datasets demonstrate that the proposed VAANet outperforms the state-of-the-art approaches for video emotion recognition. Our source code is released at: https://github.com/maysonma/VAANet.

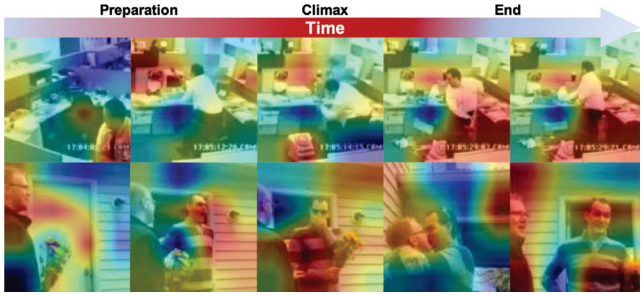

Figure 1: Illustration of the keyframes and discriminative regions for emotion recognition in user-generated videos. Although the story in a video may contain different stages, the emotion is mainly evoked by some keyframes (as shown by the temporal attentions in the color bar) and corresponding discriminative regions (as illustrated by the spatial attentions in the heat map).

Introduction

The convenience of mobile devices and social networks has enabled users to generate videos and upload to Internet in daily life to share their experiences and express personal opinions. As a result, an explosive growing volume of videos are being created, which results in urgent demand for the analysis and management of these videos. Besides the objective content recognition, such as objects and actions (Zhu et al. 2018; Choutas et al. 2018), understanding the emotional impact of the videos plays an important role in human-centered computing. On the one hand, the videos can, to a large extent, reflect the psychological states of the video generators. We can predict the generators' possible extreme behaviors, such as depression and suicide, and take corresponding preventive actions. On the other hand, the videos that evoke strong emotions can easily resonate with viewers and bring them immersive watching experiences. Appropriate emotional resonation is crucial in intelligent advertising and video recommendation. Further, emotion recognition in user-generated videos (UGVs) can help companies analyze how customers evaluate their products and assist governments to manage the Internet.

Although with the advent of deep learning, remarkable progress has been made on text sentiment classification (Zhang, Wang, and Liu 2018), image emotion analysis (Zhao et al. 2018a; 2018b; Yang et al. 2018a), and video semantic understanding (Zhu et al. 2018; Choutas et al. 2018). Emotion recognition in UGVs still remains an unsolved problem, due to the following challenges. (1) *Large intra-class variation.* Videos captured in quite different scenarios may evoke similar emotions. For example, visiting an amusement park, taking part in sport competition, and playing video games may all make viewers feel "excited". This results in obvious "affective gap" between low-level features and high-level emotions. (2) *Low structured consistency.* Unlike professional and commercial videos, such as movies (Wang and Cheong 2006) and GIFs (Jou, Bhattacharya, and Chang 2014; Yang, Zhang, and Luo 2019), UGVs are usually taken with diverse structures, *e.g.* various resolutions and image blurring noises. (3) *Sparse keyframe expression.* Only limited keyframes directly convey and de-

*Corresponding Author. # Equal Contribution.

termine emotions, as shown in Figure 1, while the rest are used to introduce the background and context.

Most existing approaches on emotion recognition in UGVs focus on the first challenge, *i.e.* employing advanced image representations to bridge the affective gap, such as (1) mid-level attribute features (Jiang, Xu, and Xue 2014; Tu et al. 2019) like ObjectBank (Li et al. 2010) and SentiBank (Borth et al. 2013), (2) high-level semantic features (Chen, Wu, and Jiang 2016) like detected events (Jiang et al. 2017; Caba Heilbron et al. 2015), objects (Deng et al. 2009), and scenes (Zhou et al. 2014), and (3) deep convolutional neural network (CNN) features (Xu et al. 2018; Zhang and Xu 2018). Zhang and Xu (2018) transformed frame-level spatial features to another kernelized feature space via discrete Fourier transform, which partially addresses the second challenge. For the third challenge, the videos are either downsampled averagely to a fixed number of frames (Zhang and Xu 2018), or represented by continuous frames from only one segment (Tu et al. 2019).

The above methods have contributed to the development of emotion recognition in UGVs, but they still have some problems. (1) They mainly employ a two-stage shallow pipeline, *i.e.* extracting visual and/or audio features and training classifiers. (2) The visual CNN features of each frame are separately extracted, which ignore the temporal correlation of adjacent frames. (3) The fact that emotions may be determined by keyframes from several discrete segments is neglected. (4) Some methods require auxiliary data, which is not always available in real applications. For example, the extracted event, object, and scene features in (Chen, Wu, and Jiang 2016) are trained on FCVID (Jiang et al. 2017) and ActivityNet (Caba Heilbron et al. 2015), ImageNet (Deng et al. 2009), and Places205 (Zhou et al. 2014) datasets, respectively. (5) They do not consider the correlations of different emotions, such as the polarity-emotion hierarchy constraint, *i.e.* the relation of two different emotions belonging to the same polarity is closer than those from opposite polarities.

In this paper, we propose an end-to-end Visual-Audio Attention Network, termed VAANet, to address the above problems for recognizing the emotions in UGVs, without requiring any auxiliary data except the data for pre-training. First, we spit each video into an equal number of segments. Second, for each segment, we randomly select some successive frames and feed them into a 3D CNN (Hara, Kataoka, and Satoh 2018) with both spatial and channel-wise attentions to extract visual features. Meanwhile, we transform the corresponding audio waves into spectrograms and feed them into a 2D CNN (He et al. 2016) to extract audio features. Finally, the visual and audio features of different segments are weighted by temporal attentions to obtain the whole video's feature representation, which is followed by a fully connected layer to obtain emotion predictions. Considering the polarity-emotion hierarchy constraint, we design a novel classification loss, *i.e.* polarity-consistent cross-entropy (PCCE) loss, to guide the attention generation.

In summary, the contributions of this paper are threefold:

1. We are the first to study the emotion recognition task in user-generated videos in an end-to-end manner.
2. We develop a novel network architecture, *i.e.* VAANet, that integrates spatial, channel-wise, and temporal attentions into a visual 3D CNN and temporal attentions into an audio 2D CNN for video emotion recognition. We propose a novel PCCE loss, which enables VAANet to generate polarity preserved attention map.
3. We conduct extensive experiments on the VideoEmotion-8 (Jiang, Xu, and Xue 2014) and Ekman-6 (Xu et al. 2018) datasets, and the results demonstrate the superiority of the proposed VAANet method, as compared to the state-of-the-art approaches.

Related Work

Video Emotion Recognition: Psychologists usually employ two kinds of models to represent emotions: categorical emotion states (CES) and dimensional emotions space (DES). CES classify emotions into several basic categories, such as Ekman's 6 basic categories (Ekman 1992) and Plutchik's wheel of emotions (Plutchik and Kellerman 1980). DES usually employ a Cartesian space to represent emotions, such as valence-arousal-dominance (Schlosberg 1954). Since CES are easy for users to understand and label, here we adopt CES to represent emotions in videos.

Early research on video emotion recognition mainly focused on movies, which are well structured. Kang (2003) employed a Hidden Markov Model to detect affective event based on low-level features, including color, motion, and shot cut rate. Joint combination of visual and audio features with support vector machine (Wang and Cheong 2006) and conditional random fields (Xu et al. 2013) achieves promising result. Some recent methods work on Animated GIFs (Jou, Bhattacharya, and Chang 2014; Chen and Picard 2016; Yang, Zhang, and Luo 2019). Jou, Bhattacharya, and Chang (2014) firstly proposed to recognize GIF emotions by using features of different types. Chen and Picard (2016) improved the performance by adopting 3D ConvNets to extract spatiotemporal features. Human-centered GIF emotion recognition is conducted by considering human related information and visual attention (Yang, Zhang, and Luo 2019).

Because of the content diversity and low quality, UGVs are more challenging to recognize emotions. Jiang, Xu, and Xue (2014) investigated a large set of low-level visual-audio features and mid-level attributes, *e.g.* ObjectBank (Li et al. 2010) and SentiBank (Borth et al. 2013). Chen, Wu, and Jiang (2016) extracted various high-level semantic features based on existing detectors. Compared with hand-crafted features, deep features are more widely used (Xu et al. 2018; Zhang and Xu 2018). By combining low-level visual-audio-textual features, Pang, Zhu, and Ngo (2015) showed that learned joint representations are complementary to hand-crafted features. Different from these methods, which employ a two-stage shallow pipeline, we propose the first end-to-end method to recognize emotions in UGVs by extracting attended visual and audio CNN features.

Please note that emotion recognition has also been widely studied in other modalities, such as text (Zhang, Wang, and

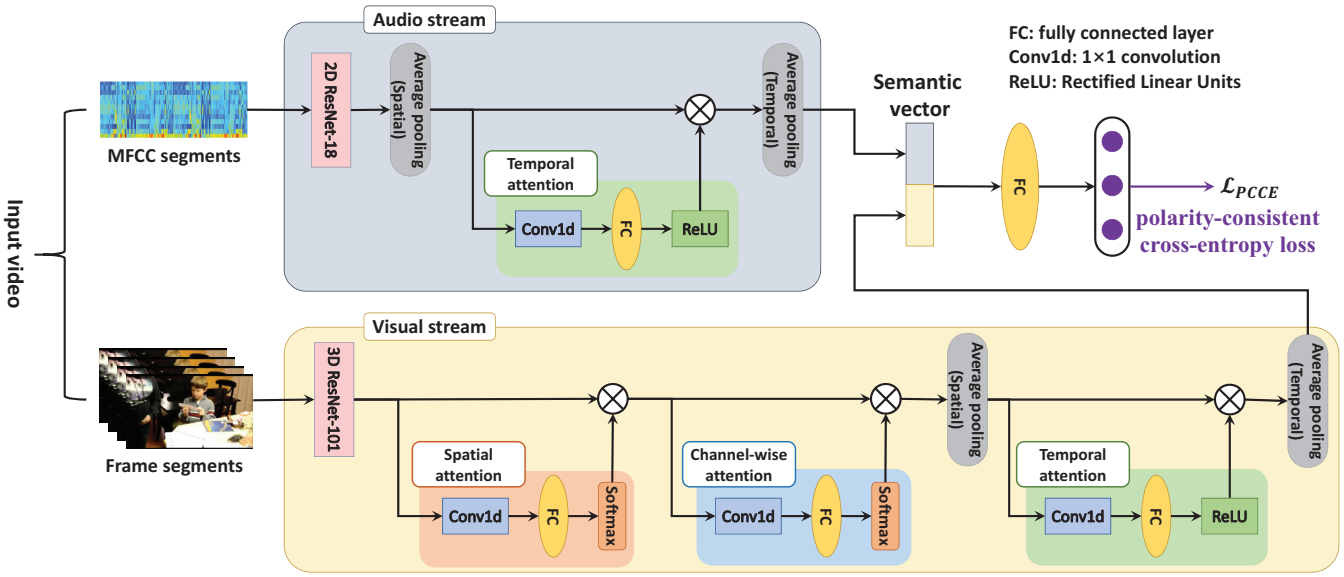

Figure 2: The framework of the proposed Visual and Audio Attention Network (VAANet). First, the MFCC descriptor from the soundtrack and the visual information are both divided into segments and fed into 2D ResNet-18 and 3D ResNet-101 respectively to extract audio and visual representation. The response feature maps of the visual stream are then fed into the stacked spatial attention, channel-wise attention, and temporal attention sub-networks, and the response feature map of the audio stream are fed into a temporal attention module. Finally, the attended semantic vectors that carry visual and audio information are concatenated. Meanwhile, a novel polarity-consistent cross-entropy loss is optimized to guide the attention generation for video emotion recognition.

Liu 2018), images (Zhao et al. 2017; Yang et al. 2018b; Zhao et al. 2018a; Yang et al. 2018a; Zhao et al. 2018c; 2019c; 2019b; Yao et al. 2019; Zhan et al. 2019), speech (El Ayadi, Kamel, and Karray 2011), physiological signals (Alarcao and Fonseca 2017; Zhao et al. 2019a), and multi-modal data (Soleymani et al. 2017; Zhao et al. 2019d).

Attention-Based Models: Since attention can be considered as a dynamic feature extraction mechanism that combines contextual fixations over time (Mnih et al. 2014; Chen et al. 2017), it has been seamlessly incorporated into deep learning architectures and achieved outstanding performances in many vision-related tasks, such as image classification (Woo et al. 2018), image captioning (You et al. 2016; Chen et al. 2017; 2018), and action recognition (Song et al. 2017). These attention methods can be roughly divided into four categories: spatial attention (Song et al. 2017; Woo et al. 2018), semantic attention (You et al. 2016), channel-wise attention (Chen et al. 2017; Woo et al. 2018), and temporal attention (Song et al. 2017).

There are also several methods that employ attention for emotion recognition in images (You, Jin, and Luo 2017; Yang et al. 2018a; Zhao et al. 2019b) and speech (Mirsamadi, Barsoum, and Zhang 2017). The former methods mainly consider spatial attention except PDANet (Zhao et al. 2019b) which also employs channel-wise attention, while the latter one only uses temporal attention. To the best of our knowledge, attention has not been studied on emotion recognition in user-generated videos. In this paper, we systematically investigate the influence of different attentions in video emotion recognition, including the importance of local spatial context by spatial attention, the interdependency between different channels by channel-wise attention, and the importance of different segments by temporal attention.

Visual-Audio Attention Network

We propose a novel CNN architecture with spatial, channel-wise, and temporal attention mechanisms for emotion recognition in user generated videos. Figure 2 shows the overall framework of the proposed VAANet. Specifically, VAANet has two streams to respectively exploit the visual and audio information. The visual stream consists of three attention modules and the audio stream contains a temporal attention module. The spatial attention and the channel-wise attention sub-networks in the visual stream are designed to automatically focus on the regions and channels that carry discriminative information within each feature map. The temporal attention sub-networks in both the visual and audio streams are designed to assign weights to different segments of a video. The training of VAANet is performed by minimizing the newly designed polarity-consistent cross-entropy loss in an end-to-end manner.

Visual Representation Extraction

To extract visual representations from a long-term video, following (Wang et al. 2016), the visual stream of our model works on short snippets sparsely sampled from the entire video. Specifically, we divide each video into t segments

with equal duration, and then randomly sample a short snippet of k successive frames from each segment. We use 3D ResNet-101 (Hara, Kataoka, and Satoh 2018) as backbone of the visual stream. It takes the t snippets (each has k successive frames) as input and independently processes them up to the last spatiotemporal convolutional layer conv5 into a super-frame. Suppose we are given N training samples $\{(\mathbf{x}_l^V, \mathbf{y}_l)\}_{l=1}^N$, where $\mathbf{x}_l^V$ is the visual information of video l, and $\mathbf{y}_l$ is the corresponding emotion label. For sample $\mathbf{x}_l^V$, suppose the feature map of the conv5 in 3D ResNet-101 is $\mathbf{F}_l^V \in \mathbb{R}^{t\times h\times w\times n}$ (we omit l for simplicity in the following), where h and w are the spatial size (height and width) of the feature map, n is the number of channels, and t is the number of snippets. We reshape $\mathbf{F}^V$ as

$$\mathbf{F}^V = \begin{bmatrix} \mathbf{f}_{11}^V & \mathbf{f}_{12}^V & \cdots & \mathbf{f}_{1m}^V \\ \mathbf{f}_{21}^V & \mathbf{f}_{22}^V & \cdots & \mathbf{f}_{2m}^V \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{f}_{t1}^V & \mathbf{f}_{t2}^V & \cdots & \mathbf{f}_{tm}^V \end{bmatrix} \in \mathbb{R}^{t\times m\times n}, \quad (1)$$

by flattening the height and width of the original $\mathbf{F}^V$, where $\mathbf{f}_{ij}^V \in \mathbb{R}^n$ and $m = h \times w$. Here we can consider $\mathbf{f}_{ij}^V$ as the visual feature of the j-th location in the i-th super-frame. In the following, we omit the superscript V for simplicity.

Visual Spatial Attention Estimation

We employ a spatial attention module to automatically explore the different contributions of the regions in super-frames to predict the emotions. Following (Chen et al. 2017), we employ a two-layer neural network, *i.e.* a 1×1 convolutional layer followed by a fully-connected layer with a softmax function to generate the spatial attention distributions over all the super-frame regions. That is, for each $\mathbf{F}_i \in \mathbb{R}^{m\times n}(i = 1, 2, \cdots, t)$

$$\begin{aligned} \mathbf{H}_i^S &= \mathbf{W}^{S_1}(\mathbf{W}^{S_2}\mathbf{F}_i^\top)^\top, \\ \mathbf{A}_i^S &= \text{Softmax}(\mathbf{H}_i^S), \end{aligned} \quad (2)$$

where $\mathbf{W}^{S_1} \in \mathbb{R}^{m\times m}$ and $\mathbf{W}^{S_2} \in \mathbb{R}^{1\times n}$ are two learnable parameter matrices, $\top$ is the transpose of a matrix, and $\mathbf{A}_i^S \in \mathbb{R}^{m\times 1}$.

And then we can obtain a weighted feature map based on spatial attention as follows

$$\mathbf{F}_i^S = \mathbf{A}_i^S \otimes \mathbf{F}_i, \quad (3)$$

where $\otimes$ is the multiplication of a matrix and a vector, which is performed by multiplying each value in the vector to each column of the matrix.

Visual Channel-Wise Attention Estimation

Assuming that each channel of a feature map in a CNN is a response activation of the corresponding convolutional layer, channel-wise attention can be viewed as a process of selecting semantic attributes (Chen et al. 2017). To generate the channel-wise attention, we first transpose $\mathbf{F}^V$ to $\mathbf{G}$

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}_{11} & \mathbf{g}_{12} & \cdots & \mathbf{g}_{1n} \\ \mathbf{g}_{21} & \mathbf{g}_{22} & \cdots & \mathbf{g}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_{t1} & \mathbf{g}_{t2} & \cdots & \mathbf{g}_{tn} \end{bmatrix} \in \mathbb{R}^{t\times n\times m}, \quad (4)$$

where $\mathbf{g}_{ij} \in \mathbb{R}^m$ represents the j-th channel in the i-th super-frame of the feature map $\mathbf{G}$. The channel-wise attention for $\mathbf{G}_i \in \mathbb{R}^{n\times m}(i = 1, 2, \cdots, t)$ is defined as

$$\begin{aligned} \mathbf{H}_i^C &= \mathbf{W}^{C_1}(\mathbf{W}^{C_2}\mathbf{G}_i^\top)^\top, \\ \mathbf{A}_i^C &= \text{Softmax}(\mathbf{H}_i^C), \end{aligned} \quad (5)$$

where $\mathbf{W}^{C_1} \in \mathbb{R}^{n\times n}$ and $\mathbf{W}^{C_2} \in \mathbb{R}^{1\times m}$ are two learnable parameter matrices, and $\mathbf{A}_i^C \in \mathbb{R}^{n\times 1}$.

And then a weighted feature map based on channel-wise attention is computed as follows

$$\mathbf{G}_i^C = \mathbf{A}_i^C \otimes \mathbf{G}_i, \quad (6)$$

where $\otimes$ is the multiplication of a matrix and a vector.

Visual Temporal Attention Estimation

For a video, the discriminability of each frame to recognize emotions is obviously different. Only some keyframes contain discriminative information, while the others only provide the background and context information (Song et al. 2017). Based on such observations, we design a temporal attention sub-network to automatically focus on the important segments that contain keyframes. To generate the temporal attention, we first apply spatial average pooling to $\mathbf{G}^C$ and reshape it to $\mathbf{P}$

$$\mathbf{P} = [\ \mathbf{p}_1, \cdots, \mathbf{p}_t\ ] \in \mathbb{R}^{t\times n}, \quad (7)$$

where $\mathbf{p}_j \in \mathbb{R}^n(j = 1, 2, \cdots, t)$. Here we can consider $\mathbf{p}_j$ as the visual feature of the j-th super-frame. The temporal attention is then defined as

$$\begin{aligned} \mathbf{H}^T &= \mathbf{W}^{T_1}(\mathbf{W}^{T_2}\mathbf{P}^\top)^\top, \\ \mathbf{A}^T &= \text{ReLU}(\mathbf{H}^T), \end{aligned} \quad (8)$$

where $\mathbf{W}^{T_1} \in \mathbb{R}^{t\times t}$ and $\mathbf{W}^{T_2} \in \mathbb{R}^{1\times n}$ are two learnable parameter matrices, and $\mathbf{A}^T \in \mathbb{R}^{t\times 1}$. Following (Song et al. 2017), we use ReLU (Rectified Linear Units) as the activation function here for its better convergence performance. The final visual embedding is the weighted sum of all the super-frames

$$\mathbf{E}^V = \sum_{j=1}^{t} \mathbf{p}_j \cdot \mathbf{A}_j^T \in \mathbb{R}^n. \quad (9)$$

Audio Representation Extraction

Audio features are complementary to visual features, because they contain information of another modality. In our problem, we choose to use the most well-known audio representation: the mel-frequency cepstral coefficients (MFCC). Suppose we are given N audio training samples $\{(\mathbf{x}_l^A, \mathbf{y}_l)\}_{l=1}^N$, where $\mathbf{x}_l^A$ is a descriptor from the entire

soundtrack of the video V_l and $\mathbf{y}_l$ is the corresponding emotion label. We center-crop $\mathbf{x}_l^A$ to a fixed length of q to get $\mathbf{x}_l^{A'}$, and pad itself when it is necessary. Similar to the method we take in extracting visual representation, we divide each descriptor into t segments and use 2D ResNet-18 (He et al. 2016) as backbone of the audio stream of our model which processes descriptor segments independently. For descriptor $\mathbf{x}_l^A$, suppose the feature map of the conv5 in 2D ResNet-18 is $\mathbf{F}_l^A \in \mathbb{R}^{t\times h'\times w'\times n'}$ (we omit l for simplicity in the following), where h' and w' are the height and width of the feature map, n' is the number of channels, and t is the number of segments. We apply spatial average pooling to $\mathbf{F}^A$ and obtain $\mathbf{F}^{A'} \in \mathbb{R}^{t\times n'}$.

Audio Temporal Attention Estimation

With similar motivation to integrate temporal attention subnetwork into the visual stream, we introduce a temporal attention sub-network to explore the influence of audio information in different segments for recognizing emotions as

$$\begin{aligned}\mathbf{H}^A &= \mathbf{W}^{A1}(\mathbf{W}^{A2}(\mathbf{F}^{A'})^\top)^\top,\\ \mathbf{A}^A &= \mathrm{ReLU}(\mathbf{H}^A),\end{aligned} \tag{10}$$

where $\mathbf{W}^{A1} \in \mathbb{R}^{t\times t}$ and $\mathbf{W}^{A2} \in \mathbb{R}^{1\times n'}$ are two learnable parameter matrices, and $\mathbf{A}^A \in \mathbb{R}^{t\times 1}$. The final audio embedding is the weighted sum of all the segments

$$\mathbf{E}^A = \sum_{j=1}^{t} \mathbf{F}_j^{A'} \cdot \mathbf{A}_j^A \in \mathbb{R}^{n'}. \tag{11}$$

Polarity-Consistent Cross-Entropy Loss

We concatenate $\mathbf{E}^V$ and $\mathbf{E}^A$ to obtain an aggregated semantic vector $\mathbf{E} = [\mathbf{E}^V, \mathbf{E}^A]$, which can be viewed as the final representation of a video and is fed into a fully connected layer to predict the emotion labels. The traditional cross-entropy loss is defined as

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} \mathbb{1}_{[c=y_i]} \log p_{i,c}, \tag{12}$$

where C is the number of emotion classes ($C = 8$ for VideoEmotion-8 and $C = 6$ for Ekman-6 in this paper), $\mathbb{1}_{[c=y_i]}$ is a binary indicator, and $p_{i,c}$ is the predicted probability that video i belongs to class c.

Directly optimizing the cross-entropy loss in Eq. (12) can lead some videos to be incorrectly classified into categories that have opposite polarity. In this paper, we design a novel polarity-consistent cross-entorpy (PCCE) loss to guide the attention generation. That is, the penalty of the predictions that have opposite polarity to the ground truth is increased. The PCCE loss is defined as

$$\mathcal{L}_{PCCE} = -\frac{1}{N}\sum_{i=1}^{N}(1+\lambda(g(\hat{y}_i, y_i)))\sum_{c=1}^{C} \mathbb{1}_{[c=y_i]} \log p_{i,c}, \tag{13}$$

where λ is a penalty coefficient that controls the penalty extent. Similar to the indicator function, g(.,.) represents whether to add the penalty or not and is defined as

$$g(\hat{y}, y) = \begin{cases} 1, & \text{if polarity}(\hat{y}) \neq \text{polarity}(y),\\ 0, & \text{otherwise,}\end{cases} \tag{14}$$

where polarity(.) is a function that maps an emotion category to its polarity (positive or negative). Since the derivatives with respect to all parameters can be computed, we can train the proposed VAANet effectively in an end-to-end manner using off-the-shelf optimizer to minimize the loss function in Eq. (13).

Experiments

In this section, we evaluate the proposed VAANet model on emotion recognition in user-generated videos. We first introduce the employed benchmarks, compared baselines, and implementation details. And then we report and analyze the major results together with some empirical analysis.

Experimental Settings

Benchmarks We evaluate the performances of the proposed method on two publicly available datasets that contain emotion labels in user-generated videos: VideoEmotion-8 (Jiang, Xu, and Xue 2014) and Ekman-6 (Xu et al. 2018).

VideoEmotion-8 (Jiang, Xu, and Xue 2014) (VE-8) consists of 1,101 videos collected from Youtube and Flickr with average duration 107 seconds. The videos are labeled into one of the Plutchik's eight basic categories (Plutchik and Kellerman 1980): negative *anger*, *disgust*, *fear*, *sadness* and positive *anticipation*, *joy*, *surprise*, *trust*. In each category, there are at least 100 videos. Ekman-6 (Xu et al. 2018) (E-6) contains 1,637 videos also collected from Youtube and Flickr. The average duration is 112 seconds. The videos are labeled with Ekman's six emotion categories (Ekman 1992), *i.e.* negative *anger*, *disgust*, *fear*, *sadness* and positive *joy*, *surprise*.

Baselines To compare VAANet with the state-of-the-art approaches for video emotion recognition, we select the following methods as baselines: (1) SentiBank (Borth et al. 2013), (2) Enhanced Multimodal Deep Bolzmann Machine (E-MDBM) (Pang, Zhu, and Ngo 2015), (3) Image Transfer Encoding (ITE) (Xu et al. 2018), (4) Visual+Audio+Attribute (V.+Au.+At.) (Jiang, Xu, and Xue 2014), (5) Context Fusion Net (CFN) (Chen, Wu, and Jiang 2016), (6) V.+Au.+At.+E-MDBM (Pang, Zhu, and Ngo 2015), (7) Kernelized and Kernelized+SentiBank (Zhang and Xu 2018).

Implementation Details Following (Jiang, Xu, and Xue 2014; Zhang and Xu 2018), the experiments on VE-8 are conducted 10 runs. In each run, we randomly select 2/3 of the data from each category for training and the rest for testing. We report the average classification accuracy of the 10 runs. For E-6, we employ the split provided by the dataset, *i.e.* 819 videos for training and 818 for testing. The classification accuracy on the test set is evaluated.

Table 1: Comparison between the proposed VAANet and several state-of-the-art methods on the VE-8 dataset, where 'Visual', 'Audio', and 'Attribute' indicate whether corresponding features are used, 'Auxiliary' means whether no auxiliary data is used except the commonly used ImageNet (Deng et al. 2009) and Kinetics (Kay et al. 2017) for pre-training, and 'End-to-end' indicates whether the corresponding algorithm is trained in an end-to-end manner. The best method is emphasized in bold. Our method achieves the best results, outperforming the state-of-the-art approaches.

Method	Visual	Audio	Attribute	Auxiliary	End-to-end	Accuracy
SentiBank (Borth et al. 2013)	✓		✓			35.5
E-MDBM (Pang, Zhu, and Ngo 2015)	✓	✓		✓		40.4
ITE (Xu et al. 2018)	✓	✓	✓			44.7
V.+Au.+At. (Jiang, Xu, and Xue 2014)	✓	✓	✓			46.1
CFN (Chen, Wu, and Jiang 2016)	✓	✓	✓			50.4
V.+Au.+At.+E-MDBM (Pang, Zhu, and Ngo 2015)	✓	✓	✓			51.1
Kernelized (Zhang and Xu 2018)	✓			✓		49.7
Kernelized+SentiBank (Zhang and Xu 2018)	✓		✓			52.5
VAANet (Ours)	✓	✓		✓	✓	**54.5**

Table 2: Comparison between the proposed VAANet and several state-of-the-art methods on the E-6 dataset. The best method is emphasized in bold. Our method performs better than the state-of-the-art approaches.

Method	Accuracy
ITE (Xu et al. 2018)	51.2
CFN (Chen, Wu, and Jiang 2016)	51.8
Kernelized (Zhang and Xu 2018)	54.4
VAANet (Ours)	**55.3**

Our model is based on two state-of-the-art CNN architectures: 2D ResNet-18 (He et al. 2016) and 3D ResNet-101 (Hara, Kataoka, and Satoh 2018), which are initialized with the weights pre-trained on ImageNet (Deng et al. 2009) and Kinetics (Carreira and Zisserman 2017), respectively. In addition, for the visual stream, we divide the input video into 10 segments and sample 16 successive frames from each of them. We resize each frame of the visual sample and make the short side length of the sample equal to 112 pixels, and then apply random horizontal flips and crop a random 112 x 112 patch as data augmentation to reduce overfitting. In our training, Adam (Kingma and Ba 2014) is adopted to automatically adjust the learning rate during optimization, with the initial learning rate set to 0.0002 and the model is trained with batch-size 32 for 150 epochs. Our model is implemented using PyTorch.

Comparison with the State-of-the-art

The extracted features, training strategies, and average performance comparisons between the proposed VAANet and the state-of-the-art approaches are shown in Tables 1 and 2 on VE-8 and E-6 datsets, respectively. From the results, we have the following observations:

(1) All these methods consider visual features, which is reasonable since the visual content in videos is the most direct way to evoke emotions. Further, existing methods all employ the traditional shallow learning pipeline, which indicates that the corresponding algorithms are trained step by step instead of end-to-end.

(2) Most previous methods extract attribute features. It is demonstrated that attributes indeed contribute to the emotion recognition task (Chen, Wu, and Jiang 2016). However, this requires some auxiliary data to train attribute classifiers. For example, though highly related to emotions, the adjective noun pairs obtained by SentiBank are trained on the VSO dataset (Borth et al. 2013). Besides high computation cost, the auxiliary data to train such attribute classifiers are often not available in real applications.

(3) Without extracting attribute features or requiring auxiliary data, the proposed VAANet is the only end-to-end model and achieves the best emotion recognition accuracy. Compared with the reported state-of-the-art results, *i.e.* Kernelized+SentiBank (Zhang and Xu 2018) on VE-8 and Kernelized (Zhang and Xu 2018) on E-6, VAANet can respectively obtain 2% and 0.9% performance gains. The performance improvements benefit from the advantages of VAANet. First, the various attentions enable the network to focus on discriminative key segments, spatial context, and channel interdependency. Second, the novel PCCE loss considers the polarity-emotion hierarchy constraint, *i.e.* the emotion correlations, which can guide the detailed learning process. Third, the visual features extracted by 3D ResNet-101 can model the temporal correlation of the adjacent frames in a given segment.

Ablation Study

The proposed VAANet model contains two major components: a novel attention mechanism and a novel cross-entropy loss. We conduct ablation study to further verify their effectiveness by changing one component and fixing the other. First, using polarity-consistent cross-entropy loss, we investigate the influence of different attentions, including visual spatial (VS), visual channel-wise (VCW), visual temporal (VT), and audio temporal (AT) ones. The emotion recognition accuracy of each emotion category and the average accuracy on VE-8 and E-6 datasets are shown in Table 3 and 4, respectively. From the results, we can observe that: (1) visual attentions even only using spatial at-

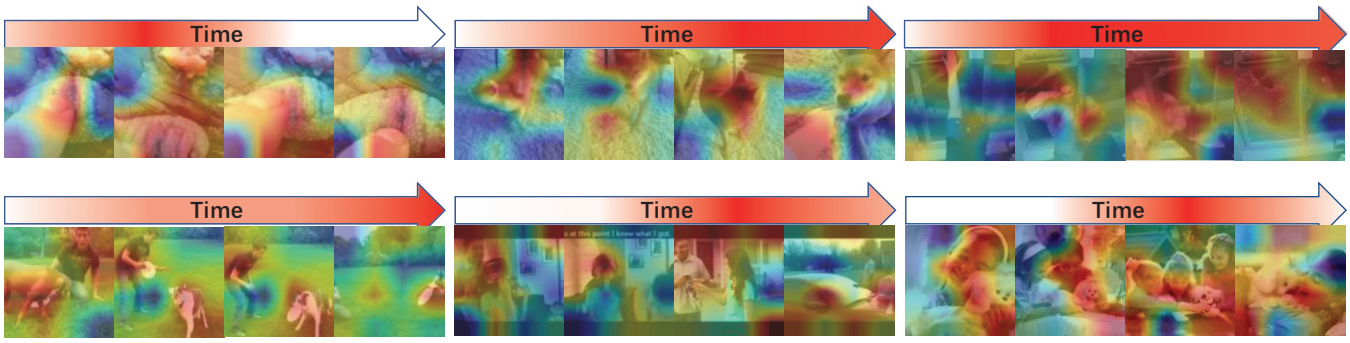

Figure 3: Visualization of the learned visual spatial attention and visual temporal attention. In both learned color bar and attention maps, red regions indicate more attention. The proposed VAANet can focus on the salient and discriminative frames and regions for emotion recognition in user-generated videos. Note that all the shown examples are drawn from the test set of VE-8.

Table 3: Ablation study of different attentions in the proposed VAANet for video emotion recognition on the VE-8 dataset, where 'VS', 'VCW', 'VT', and 'AT' are short for visual spatial, visual channel-wise, visual temporal, and audio temporal attentions, respectively. All the attentions contribute to the emotion regression task.

Attentions	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	**Average**
AT	11.3	3.0	19.1	46.1	46.1	42.4	74.7	10.7	**41.4**
VS	45.4	27.9	44.8	59.5	49.1	51.0	65.1	35.6	**51.7**
VS+VCW	46.2	25.9	42.8	63.2	50.9	40.2	67.5	45.2	**52.6**
VS+VCW+VT	55.6	30.8	37.5	60.4	57.7	50.0	65.2	34.6	**53.6**
VS+VCW+VT+AT	48.2	24.1	33.3	55.9	55.9	52.5	77.1	35.6	**54.5**

Table 4: Ablation study of different attentions in the proposed VAANet for video emotion recognition on the E-6 dataset.

Attentions	Anger	Disgust	Fear	Joy	Sadness	Surprise	**Average**
AT	32.4	14.1	38.5	28.1	63.9	46.5	**37.2**
VS	59.9	44.8	49.7	46.2	35.3	62.5	**50.8**
VS+VCW	58.4	52.6	49.2	46.3	44.0	65.1	**53.4**
VS+VCW+VT	57.1	53.1	48.0	57.0	38.7	65.0	**54.5**
VS+VCW+VT+AT	55.1	50.6	45.7	53.7	50.3	68.9	**55.3**

Table 5: Performance comparison between traditional cross-entropy loss (CE) and our polarity-consistent cross-entropy loss (PCCE) measured by average accuracy.

Attentions	Loss	VE-8	E-6
VS+VCW+VT	CE	51.9	52.0
	PCCE	53.6	54.5
VS+VCW+VT+AT	CE	53.9	54.6
	PCCE	54.5	55.3

tention significantly outperform audio attentions (on average more than 10% improvement), which is understandable because in many videos the audio does not change much; (2) adding each one of them introduces performance gains, which demonstrates that all these attentions contribute to the video emotion recognition task; (3) though not performing well alone, combining audio features with visual features can boost the performance with about 1% accuracy gains.

Second, we evaluate the effectiveness of the proposed polarity-consistent cross-entropy loss (PCCE) by comparing with traditional cross-entropy loss (CE). Table 5 shows the results when visual attentions (VS+VCW+VT) and visual+audio attentions (VS+VCW+VT+AT) are considered. From the results, it is clear that for both settings, PCCE performs better. The performance improvements of PCCE over CE for visual attentions and visual+audio attentions are 1.5%, 0.6% and 2.5%, 0.7% on the VE-8 and E-6 datasets, respectively. This demonstrates the effectiveness of emotion hierarchy as prior knowledge. This novel loss can also be easily extended to other machine learning tasks if some prior knowledge is available.

Visualization

In order to show the interpretability of our model, we use the heat map generated by the Gram-Cam algorithm (Selvaraju et al. 2017) to visualize the visual spatial attention obtained by the proposed VAANet. The visual temporal attention generated by our model is also illustrated through the color bar.

As illustrated in Figure 3, the well-trained VAANet can successfully pay more attention not only to the discriminative frames, but also to different salient regions in corresponding frames. For example, in the top left test case, the key object that makes people feel 'disgust' is a caterpillar, and a man is touching it with his finger. The model assigns the highest temporal attention when the finger is removed, and the caterpillar is completely exposed to the camera. In the bottom left case, our model can focus on the person and the dog during the whole video. Further, when the dog rushes out from the bottom right corner and makes the audience feel 'anticipated', the temporal attention becomes larger. In the middle bottom case, our model pays more attention when the 'surprise' comes up.

Conclusion

In this paper, we have proposed an effective emotion recognition method in user-generated videos based on visual and audio attentions. The developed novel VAANet model consists of a novel attention mechanism and a novel cross-entropy loss, with less auxiliary data used. By considering various attentions, VAANet can better focus on the discriminative key segments and their key regions. The polarity-consistent cross-entropy loss can guide the attention generation. The extensive experiments conducted on VideoEmotion-8 and Ekman-6 benchmarks demonstrate that VAANet achieves 2.0% and 0.9% performance improvements as compared to the best state-of-the-art video emotion recognition approach. In future studies, we plan to extend the VAANet model to both fine-tuned emotion classification and emotion regression tasks. We also aim to investigate attentions that can better concentrate on the key frames in each video segment.

Acknowledgments

This work is supported by Berkeley DeepDrive, the National Natural Science Foundation of China (Nos. 61701273, 61876094, U1933114), the Major Project for New Generation of AI Grant (No. 2018AAA010040003), Natural Science Foundation of Tianjin, China (Nos.18JCYBJC15400, 18ZXZNGX00110), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

References

Alarcao, S. M., and Fonseca, M. J. 2017. Emotions recognition using eeg signals: A survey. *IEEE TAFFC*.

Borth, D.; Chen, T.; Ji, R.; and Chang, S.-F. 2013. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *ACM MM*, 459–460.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 961–970.

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 6299–6308.

Chen, W., and Picard, R. W. 2016. Predicting perceived emotions in animated gifs with 3d conolutional neural networks. In *ISM*, 367–368.

Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; and Chua, T.-S. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 5659–5667.

Chen, H.; Ding, G.; Lin, Z.; Zhao, S.; and Han, J. 2018. Show, observe and tell: Attribute-driven attention model for image captioning. In *IJCAI*, 606–612.

Chen, C.; Wu, Z.; and Jiang, Y.-G. 2016. Emotion in context: Deep semantic feature fusion for video emotion recognition. In *ACM MM*, 127–131.

Choutas, V.; Weinzaepfel, P.; Revaud, J.; and Schmid, C. 2018. Potion: Pose motion representation for action recognition. In *CVPR*, 7024–7033.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Ekman, P. 1992. An argument for basic emotions. *Cognition & Emotion* 6(3-4):169–200.

El Ayadi, M.; Kamel, M. S.; and Karray, F. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *PR* 44(3):572–587.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 6546–6555.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Jiang, Y.-G.; Wu, Z.; Wang, J.; Xue, X.; and Chang, S.-F. 2017. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE TPAMI* 40(2):352–364.

Jiang, Y.-G.; Xu, B.; and Xue, X. 2014. Predicting emotions in user-generated videos. In *AAAI*, 73–79.

Jou, B.; Bhattacharya, S.; and Chang, S.-F. 2014. Predicting viewer perceived emotions in animated gifs. In *ACM MM*, 213–216.

Kang, H.-B. 2003. Affective content detection using hmms. In *ACM MM*, 259–262.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv:1705.06950*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Li, L.-J.; Su, H.; Fei-Fei, L.; and Xing, E. P. 2010. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 1378–1386.

Mirsamadi, S.; Barsoum, E.; and Zhang, C. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *ICASSP*, 2227–2231.

Mnih, V.; Heess, N.; Graves, A.; et al. 2014. Recurrent models of visual attention. In *NIPS*, 2204–2212.

Pang, L.; Zhu, S.; and Ngo, C.-W. 2015. Deep multimodal learning for affective analysis and retrieval. *IEEE TMM* 17(11):2008–2020.

Plutchik, R., and Kellerman, H. 1980. *Emotion, theory, research, and experience*. Academic press.

Schlosberg, H. 1954. Three dimensions of emotion. *Psychological Review* 61(2):81.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.-F.; and Pantic, M. 2017. A survey of multimodal sentiment analysis. *IVC* 65:3–14.

Song, S.; Lan, C.; Xing, J.; Zeng, W.; and Liu, J. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 4263–4270.

Tu, G.; Fu, Y.; Li, B.; Gao, J.; Jiang, Y.-G.; and Xue, X. 2019. A multi-task neural approach for emotion attribution, classification and summarization. *IEEE TMM*.

Wang, H. L., and Cheong, L.-F. 2006. Affective understanding in film. *IEEE TCSVT* 16(6):689–704.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 20–36.

Woo, S.; Park, J.; Lee, J.-Y.; and So Kweon, I. 2018. Cbam: Convolutional block attention module. In *ECCV*, 3–19.

Xu, M.; Xu, C.; He, X.; Jin, J. S.; Luo, S.; and Rui, Y. 2013. Hierarchical affective content analysis in arousal and valence dimensions. *SIGPRO* 93(8):2140–2150.

Xu, B.; Fu, Y.; Jiang, Y.-G.; Li, B.; and Sigal, L. 2018. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE TAFFC* 9(2):255–270.

Yang, J.; She, D.; Lai, Y.-K.; Rosin, P. L.; and Yang, M.-H. 2018a. Weakly supervised coupled networks for visual sentiment analysis. In *CVPR*, 7584–7592.

Yang, J.; She, D.; Lai, Y.-K.; and Yang, M.-H. 2018b. Retrieving and classifying affective images via deep metric learning. In *AAAI*.

Yang, Z.; Zhang, Y.; and Luo, J. 2019. Human-centered emotion recognition in animated gifs. In *ICME*.

Yao, X.; She, D.; Zhao, S.; Liang, J.; Lai, Y.-K.; and Yang, J. 2019. Attention-aware polarity sensitive embedding for affective image retrieval. In *ICCV*, 1140–1150.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *CVPR*, 4651–4659.

You, Q.; Jin, H.; and Luo, J. 2017. Visual sentiment analysis by attending on local image regions. In *AAAI*, 231–237.

Zhan, C.; She, D.; Zhao, S.; Cheng, M.-M.; and Yang, J. 2019. Zero-shot emotion recognition via affective structural embedding. In *ICCV*, 1151–1160.

Zhang, H., and Xu, M. 2018. Recognition of emotions in user-generated videos with kernelized features. *IEEE TMM* 20(10):2824–2835.

Zhang, L.; Wang, S.; and Liu, B. 2018. Deep learning for sentiment analysis: A survey. *DMKD* 8(4):e1253.

Zhao, S.; Yao, H.; Gao, Y.; Ji, R.; and Ding, G. 2017. Continuous probability distribution prediction of image emotions via multi-task shared sparse regression. *IEEE TMM* 19(3):632–645.

Zhao, S.; Ding, G.; Huang, Q.; Chua, T.-S.; Schuller, B. W.; and Keutzer, K. 2018a. Affective image content analysis: A comprehensive survey. In *IJCAI*, 5534–5541.

Zhao, S.; Yao, H.; Gao, Y.; Ding, G.; and Chua, T.-S. 2018b. Predicting personalized image emotion perceptions in social networks. *IEEE TAFFC* 9(4):526–540.

Zhao, S.; Zhao, X.; Ding, G.; and Keutzer, K. 2018c. Emotiongan: unsupervised domain adaptation for learning discrete probability distributions of image emotions. In *ACM MM*, 1319–1327.

Zhao, S.; Gholaminejad, A.; Ding, G.; Gao, Y.; Han, J.; and Keutzer, K. 2019a. Personalized emotion recognition by personality-aware high-order learning of physiological signals. *ACM ToMM* 15(1s):14.

Zhao, S.; Jia, Z.; Chen, H.; Li, L.; Ding, G.; and Keutzer, K. 2019b. Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In *ACM MM*, 192–201.

Zhao, S.; Lin, C.; Xu, P.; Zhao, S.; Guo, Y.; Krishna, R.; Ding, G.; and Keutzer, K. 2019c. Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions. In *AAAI*, 2620–2627.

Zhao, S.; Wang, S.; Soleymani, M.; Joshi, D.; and Ji, Q. 2019d. Affective computing for large-scale heterogeneous multimedia data: A survey. *ACM ToMM*.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. In *NIPS*, 487–495.

Zhu, X.; Dai, J.; Yuan, L.; and Wei, Y. 2018. Towards high performance video object detection. In *CVPR*, 7210–7218.