DUAL FOCUS ATTENTION NETWORK FOR VIDEO EMOTION RECOGNITION

Haonan Qiu *Liang He* *Feng Wang*[*]

Shanghai Key Laboratory of Multidimensional Information Processing
School of Computer Science and Technology, East China Normal University
haonan.qiu@outlook.com; lhe@cs.ecnu.edu.cn; fwang@cs.ecnu.edu.cn

ABSTRACT

Video emotion recognition is a challenging task due to complex scenes and various forms of emotion expression. Most existing works focus on fusing multiple features over the whole video clips. According to our observations, given a long video clip, the emotion is usually presented by only several actions/objects in a few short snippets, and the meaningful cues are buried in the noisy background. When human judging the emotion in videos, we first find the informative clips and then closely look for emotional cues in the frames. In this paper, we propose **Dual Focus Attention Network** to mimic this process. First, three kinds of features including action, object, and scene are extracted from videos. Second, Two attention modules are used to focus on the visual features of the videos from temporal and spatial dimensions respectively. With our dual focus attention network, we can effectively discover the most emotional frames along the time dimension and the most emotional visual cues in each frame. Our experiments conducted on two widely used datasets Ekman and VideoEmotion show that our proposed approach outperforms the existing approaches.

Index Terms— Video emotion recognition, attention for video, deep learning.

1. INTRODUCTION

With the development of the mobile internet, a huge number of videos are produced and uploaded to the Internet every day. This creates a lot of demands for video content analysis. Video emotion recognition is an important task to identify the emotional impact on the viewers when watching the videos, which is widely used in many applications such as video recommendation and video summarization. Many approaches have been proposed for emotion recognition in videos [1, 2, 3, 4, 5].

Unlike other video recognition tasks, emotion recognition has many unique challenges. First, in a video, only a few frames contain emotional information. Thus, we need to pay attention to the time dimension; otherwise, we may miss the informative keyframes, and can be easily disturbed by other noises. Second, emotion is more abstract and presented by different kinds of visual information including actions, scenes, and objects. For instance, *surprise* and *fair* are two kinds of astonishing state. The presence of a cake or blood may be important cues to distinguish them. To discover the emotion behind the screen, we need to focus on different visual cues in each frame.

To tackle these challenges, we propose a network called **Dual Focus Attention Network (DFAN)**. This network contains two different attention modules namely **Time Series Focus** and **Frame Objects Focus**. *Time Series Focus* module focuses on the time dimension to discover the informative keyframes in the video and locate the segment which can best represent the emotion. *Frame Objects Focus* module focuses on the objects that appear in each frame and looks for the objects which can best represent the emotion attribute. With these two different attention modules, we know when to look and where to look so as to locate the most informative visual cues for effective recognition of emotion. The main contributions of this paper can be summarized as follows:

- We propose two attention modules, namely Time Series Focus Attention and Frame Objects Focus Attention for video emotion recognition. With these two modules, we can discover the most informative visual cues in the video on both time and spatial dimensions.
- We build the emotion recognition framework with our Dual Focus Attention modules which can effectively use different features.
- We conduct comprehensive experiments to verify the effectiveness of our approach and achieve state-of-the-art results on two benchmark video emotion datasets.

[*] Corresponding author.
The work described in this paper was supported by the National Natural Science Foundation of China (No.61375016)
The computation is performed in ECNU Multifunctional Platform for Innovation (001).

978-1-7281-1331-9/20/$31.00

2. RELATED WORKS

2.1. Emotion Recognition in Videos

The early works of emotion recognition are based on the handcrafted features on movie datasets. In [6], Hidden Markov Model with low-level features is employed to recognize the emotion of videos. In [7, 8], visual and audio features are combined to achieve promising results on the dataset consisting of Hollywood movies. Recent research of emotion recognition has shifted from movies to the videos uploaded by the users to the Internet. Jiang *et al.* [1] build the widely used VideoEmotion dataset and employ the traditional image features such as dense SIFT [9], HOG [10], and SSIM [11] with audio feature MFCC [12] to recognize the emotion in videos. In [2], Image Transfer-Encoding (ITE) is proposed which uses an image database to assist video encoding and improve recognition accuracy. In [4], Content Fusing Network is proposed to fuse multiple modalities for emotion recognition. Zhang *et al.* [4] propose a special polynomial kernel function which shows superior discriminative ability than spatial features. Xu *et al.* [5] propose the concept selection approach with different modalities to find the relationship between high-order features and video emotion. Existing works mainly focus on feature extraction, selection, and fusion over the whole videos. As discussed above, the emotion is usually expressed by only several visual cues in a few short snippets of the video. In this paper, we propose the dual focus attention network to discover the most informative features buried in the video sequence on both time and spatial dimensions so as to improve the discriminative ability of the network.

2.2. Attention Mechanism in Video Recognition

The attention module can improve the network's interpretability by capturing where the model needs to focus on. Therefore, it has been widely used in many video recognition tasks such as action recognition. Attention modules can be used on time or spatial dimensions. Wang *et al.* [13] propose the Non-Local extension module to improve the network's performance through self-attention within spatio-temporal space. Gidhar *et al.* [14] use the classic Transformer structure to aggregate features from the context around the person in the video. In [15], Eidetic 3D LSTM is proposed which employs a gate-controlled self-attention module to store better short-term features and achieves well recognition performance by observing only limited frames. Sydorov *et al.* [16] use the attention module to tightly crop the relevant screen regions in the videos. The most relevant to our Frame Objects Focus is R*CNN [17] which not only uses Faster RCNN [18] to detect the key person box in the frame, but also uses candidate secondary region box to assist action recognition in the frame. In this paper, we propose the attention modules on both time and spatial dimensions. The attention in time dimension is used to find the most informative clips that present emotions and filter out irrelevant clips. The spatial attention is used to look into each frame and pay attention to different objects that best present emotions in the frame to further distinguish different emotions.

3. DUAL FOCUS ATTENTION FOR EMOTION RECOGNITION

In this section, we present our approach for video emotion recognition in detail. Figure 1 illustrates the pipeline of our approach. First, multiple visual features are extracted as described in Section 3.1. Second, we present our dual focus attention network in Section 3.2.

3.1. Multiple Feature Extraction

Emotions are usually presented in different forms in videos. Thus, it is necessary to combine multiple features for emotion recognition. In this section, we extract three kinds of features in order to capture different visual cues for recognizing emotions.

Action Feature: Action is an important feature in videos which plays a key role for the recognition of many emotions. For example, the action *Fighting* is often accompanied by emotion *Anger*, and *Celebrating* implies *Joy* and *Surprise*. In this paper, we employ the Temporal Segment Network [19] with DPN107 [20] as the backbone to extract the action features. The networks are trained on Kinetics-400 [21] action dataset which contains more than 400 different action categories. We train two different action recognition networks with RGB and Optical modalities respectively. The networks' outputs before the last global average pooling layer are extracted. We combine both features of RGB and Optical modalities as our action features F_A ($F_A \in R^{2048}$).

Scene Feature: For video producers, the scene is an important way of presenting emotions. The gloomy and strange atmosphere makes people feel fear, while the sunny scenery brings positive emotions. The information in the scene is useful to identify the emotion. For scene feature extraction, we employ Resnet18 [22] pre-trained on the Place 365 dataset [23] which contains 1.8 million labeled images of 365 scene categories. The network's classification layer's outputs are used as our scene features F_S ($F_S \in R^{365}$).

Object Feature: The objects in each frame contain clues which are very useful for distinguishing the emotions. For example, the rain implies *Sadness*, while a flower means *Surprise* and *Joy*. Unlike some previous works such as [3] where only the global features of the whole frame are used, we employ object detection to locate the objects in each frame. Specifically, the Grid-RCNN [24] is pre-trained on COCO dataset [25] to extract the object boxes. The top K objects are selected from each frame according to the predicted probabilities. Each box's feature is the combination of the predicted

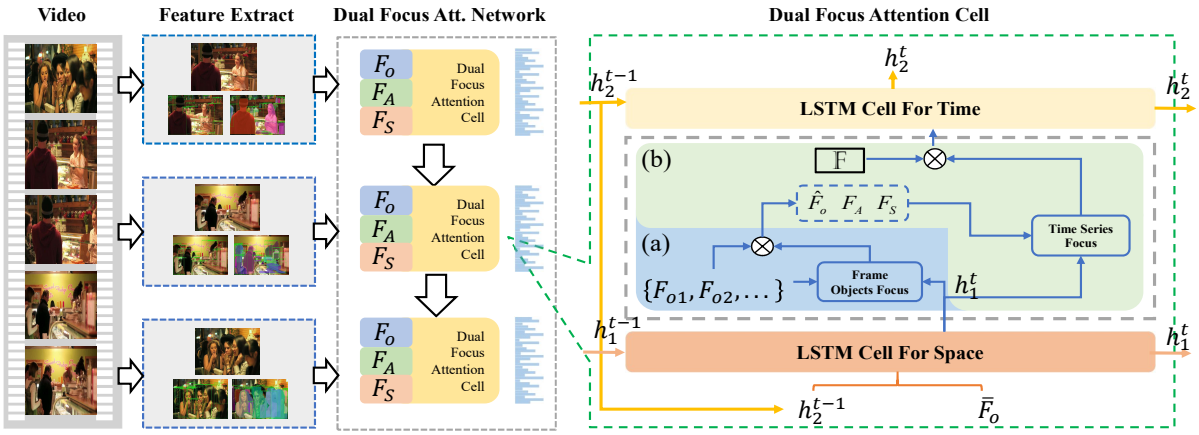


Fig. 1. The main pipeline of our Dual Focus Attention network. Three kinds of features (F_A , F_O and F_S) are extracted from each frame. The two attention modules can adjust the attention weights on the time and spatial dimensions to tell the network when and where to look. ($\hat{F_O}$ is the attention-weighted object feature. $\bar{F_O}$ is the average of the object feature. $\mathbb{F}$ is the pre-extracted action and scene features of the video.)

probability and the encoded coordinates. The final object feature of one frame is F_O ($F_O \in R^{K\times 84}$).

3.2. Dual Focus Attention Network

As illustrated in Figure 1, our Dual Focus Attention network is a recurrent neural network. The Dual Focus Attention (DFA) cell consists of two focus attention modules with two vanilla LSTM cells. For each frame of the video, three different features (F_O, F_A, and F_S) are fed to the DFA cell. The LSTM Cell for Space (LSTM.S) with Frame Objects Focus module (O.Focus) gets the attention weight for each object $\mathcal{A}_o$. The weighted object features $\hat{F}_o$ together with F_A and F_S are fed to the Time Series Focus module (T.Focus) to get the attention weights on the time dimension ($\mathcal{A}_T$). The pre-extracted action and scene features $\mathbb{F}$ are weighted with $\mathcal{A}_T$ and fed to the LSTM Cell for Time (LSTM.T) to predict the emotion of the current time. During the network inference stage, we constantly adjust the attention weights on the time dimension so that the network focuses on the important part of the video and predicts the emotion according to the features of the current frame.

3.2.1. *Frame Object Focus Module*

The main role of this module is to receive the objects' features of the frame and calculate the attention weight of each object. The weighted objects' features $\hat{F}_o$ is then used to help the Time Series Focus module to calculate the attention weights on time dimension. The weighted objects' features $\hat{F}_o$ is calculated by

$$\hat{F}_o = \sum_{i=1}^{K} F_{oi} \times \mathcal{A}o_i \tag{1}$$

where $\mathcal{A}o$ is the attention value of each object at frame t and K is the number of objects selected from each frame. In our implementation, we set $K = 16$.

$$\mathcal{A}o_i = Softmax(\mathcal{G}o_i) \tag{2}$$

where $\mathcal{G}o_i$ is the glimpse features of the objects calculated by

$$\mathcal{G}o = Tanh(W_{oh} * h_1^t + W_{of} * Fo^t + b_o) \tag{3}$$

$$h_1^t = LSTM.S(h_2^{t-1}, \bar{F}_o) \tag{4}$$

where h_1^t is the output hidden values of LSTM.S. LSTM.S accepts not only the average of all object features $\bar{F}_o$, but also the hidden values on time dimension from LSTM.T (h_2^{t-1}). Here W_{oh}, W_{of}, b_o are the network's parameters for training.

3.2.2. *Time Series Focus Module*

In this module, we use the three modality features of frame t to update the attentions $\mathcal{A}_T$ of the entire time series of the video and predict emotions of the current state $\mathcal{P}_t$

$$\mathcal{P}_t = Softmax(W_p h_2^t + b_p) \tag{5}$$

$$h_2^t = LSTM.T(\mathcal{P}_{t-1}, \mathcal{F}_t) \tag{6}$$

where h_2^t is the hidden value of LSTM.T. LSTM.T accepts the prediction at the last frame ($\mathcal{P}_{t-1}$) and the video features weighted by time attention $\mathcal{F}_t$. Here $\mathcal{F}_t$ is:

$$\mathcal{F}_t = \sum_{i=1}^{T} \mathbb{F}_i \times \mathcal{A}_{Ti} \tag{7}$$

where $\mathbb{F}$ is the concatenation of the pre-extracted action and scene features of the video, T is the video length, and $\mathcal{A}_T$ is the time attention value which is calculate by:

$$\mathcal{A}_{Ti} = Softmax(\mathcal{G}_{Ti}), \tag{8}$$

where $\mathcal{G}_T$ is the glimpse feature of the time:

$$\mathcal{G}_T = Tanh(W_{Th} * h_2^{t-1} + W_{Tf} * S_t + b_T) \quad (9)$$

where S_t is concatenation of the three extracted features ($\hat{F_O}$, F_A, and F_S) at frame t. Here W_p, b_p, W_{Th}, W_{Tf}, and b_T are parameters for training.

The videos usually contain many irrelevant scenes and still shots which affect the discrimination of the attention module. To avoid the time series focus attention module overfitting, we add the attention scrambler on time series during the training stage. The adjacent attention weights $\mathcal{A}_T$ will be mixed after each prediction. At the t-th prediction, we adjust the attention by:

$$\begin{cases} \mathcal{A}_{T(t-1)} = \beta \mathcal{A}_{T(t-1)} + (1-\beta)\mathcal{A}_{Tt} \\ \mathcal{A}_{Tt} = (1-\beta)\mathcal{A}_{T(t-1)} + \beta \mathcal{A}_{Tt} \end{cases} \quad (10)$$

where $\beta \in [0, 1]$ is the hyperparameter to adjust the confusion level. In our implementation, we empirically set $\beta = 0.6$.

4. EXPERIMENTS

4.1. Datasets & Settings

Video Emotion Dataset [1] This dataset contains 1,101 videos collected from YouTube and Flicker. The videos are labeled with eight emotion categories according to Plutchik's theory [26]. Each category has at least 110 videos, and the average length of each video is 100 seconds. Following the previous work [1], our experiments are conducted ten times. Each time we randomly choose 2/3 of the data for training and another 1/3 for validation. The average of recognition accuracies over ten experiments are used to evaluate the performance.

Ekman Dataset [2] This is the largest video emotion dataset to date. The videos are collected from social video-sharing websites. It contains 1,637 videos labeled with Ekman's six basic emotions [27]. Each class has at least 220 videos. The average duration of a video is 112 seconds. The dataset provides the training and validation splits. There are 819 videos in the training dataset and 818 videos in the validation set. Classification accuracy is used to evaluate the recognition performance.

Implementation Details. We implement our approach with the PyTorch framework. to reduce the computation cost, one frame is uniformly sampled at every eight frames for feature extraction. Both the action and the scene feature extraction networks change the final classification layer and use the training set for fine-tuning. During the inference stage, we count the predictions of the last ten frames and select the emotion that appears the most as the final prediction.

F_A	F_S	F_O	VideoEmotion	Ekman
✓			51.51	56.42
	✓		39.08	45.47
		✓	18.74	23.16
✓	✓		52.78	56.88
✓	✓	✓	53.34	57.37

Table 1. The recognition accuracies with different features.

Fusion Strategy	VideoEmotion	Ekman
Max. Pool	49.48	53.12
Avg. Pool	50.02	54.28
T.Focus	51.88	55.93
T.Focus + O.Focus	53.34	57.37

Table 2. The accuracies with different fusion strategies. (For Max and Average pooling, we uniformly sample 10 frames from visual features for pooling)

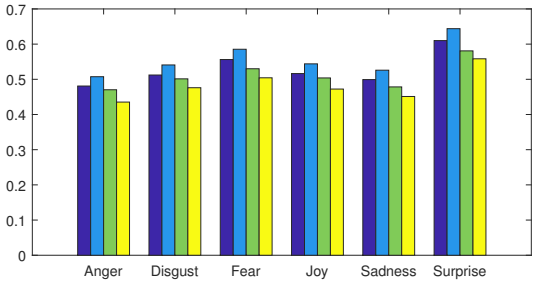
4.2. Multi-modal Feature Fusion

In this section, we evaluate the emotion recognition performance of our proposed approach with different features. As can be seen from Table 4.2, among the three kinds of features, the action feature plays the most important role. Scene feature can also achieve decent results. The action feature together with the scene feature already achieve state-of-the-art performance. Only using the object features from ROI boxes is usually not enough. However, if we fuse it with action and scene features, the object features can help the network to reduce the effect of noise in the complex videos and gain about 0.5% improvement on the final recognition accuracy.

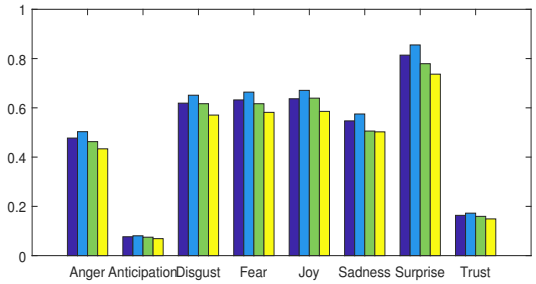
Figure 4.2 shows the effect of sampling different number of objects in each frame. We evaluate the accuracies of emotion recognition by selecting 2,4,8,16 objects in each frame. We can see that the highest accuracy is achieved when we choose four different objects in each frame. This is because that in the actual situation, a video only contains a few informative objects for emotion presentation. The noise may be introduced when too many irrelevant objects are selected.

4.3. Effectiveness of Dual Focus Attention Module

In this section, we evaluate the effectiveness of our dual focus attention module and visualize the attentions. The main role of the attention focus module is to find the most important visual cues for emotion recognition in the complex videos. Traditional video recognition approaches such as TSN [19] just use simple strategies (e.g. average pooling or max pooling) to get video recognition results from every frame or snippet. This may work when the action is simple. In emotion recognition task, most videos are collected from the Internet such as YouTube, and the qualities of the videos uploaded by the users vary a lot. Only using the simple fusion strategy may miss the key information and thus affect the performance.



(a) Ekman



(b) VideoEmotion

Fig. 2. The recognition accuracies with different number of objects selected in each frame. From left to right, the histogram shows the recognition accuracies when selecting top 2, 4, 8, 16 objects in each frame.

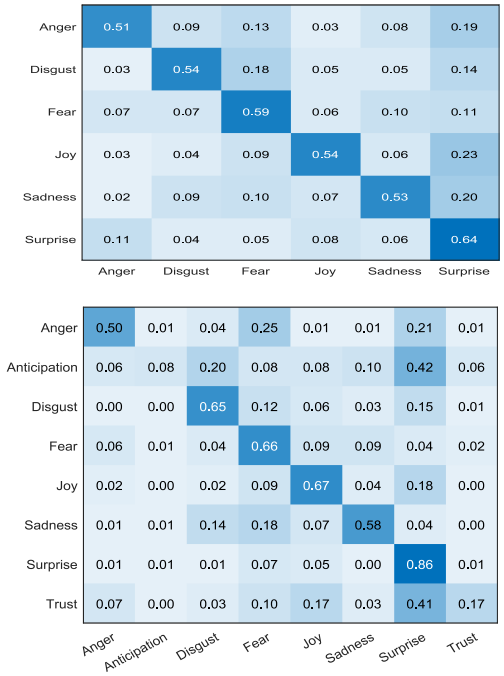
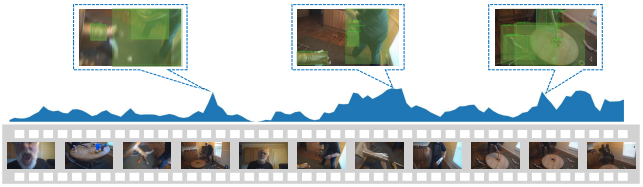


Fig. 3. Confusion matrices for Dual Focus Attention Network on Ekman (top) and VideoEmotion (bottom).



(a) The angry man smashing things

(b) The warm scenes of family members

Fig. 4. Visualization of the attentions in the video. Our module not only finds the important part of the video, but also can focus on the important object in each frame. (The blue curve illustrates the attentions along the time dimension, and the saturation degrees of the green boxes indicate the attentions of the objects.)

Table 4.3 compares the performances of different fusion strategies. After using the time series focus module, we can get 1.86% and 1.65% improvement on two datasets respectively compared with Average pooling. After using both the time series focus module and frame objects focus module, we can achieve the best performance, which can improve the recognition accuracy by 3.32% and 3.09% on two datasets respectively.

In Figure 4.3, we visualize the attentions generated by our dual focus attention modules. As can be seen in Figure 4.3, our approach can find the most informative clips of the video and focus on the useful visual cues in the frame. By simulating the recognition process of human, we can discover the most emotional cues buried in complex backgrounds, and thus improve the recognition accuracies compared with the existing sampling or pooling strategies.

4.4. Comparison with State-of-the-Art Approaches

Table 4.4 compares our approach with recent works on video emotion recognition. Just using time series focus, we can get state-of-the-art results on both datasets. By further integrating the frame objects focus, we can achieve the best results. The accuracy of recognition on the Ekman dataset is 57.37% and on the VideoEmotion is 53.34%.

Figure 3 shows the confusion matrices of the recognition results on both datasets with our dual focus attention network. We achieve impressive recognition accuracies on the emotion categories which have obvious visual features such as *Surprise*, *Fear*, and *Joy*. However, some introverted emotions such as *Anticipation* and *Trust* are difficult to distinguish by visual features and can be easily confused with other emotion

Method	VideoEmotion	Ekman
Jiang et.al[1] (Visual Only)	41.90	-
Jiang et.al[1] (Visual+Audio)	46.10	-
ITE [2]	43.80	50.90
Content Fusion Network [3]	50.60	51.80
Kernelized Feature [4]	49.70	54.40
Concept Selection [5]	51.48	55.62
Our + T.Focus	51.88	55.93
Our + T.Focus + O.Focus	**53.34**	**57.37**

Table 3. Comparison with state-of-the-art approaches. ('-' indicates the results are unavailable in the corresponding papers.)

categories. For these emotions, more effective features are needed to improve the recognition accuracy. Although only visual features are employed in this paper, our framework can easily incorporate new features such as audio.

5. CONCLUSIONS

We have presented our novel approach namely Dual Focus Attention module for video emotion recognition, which has two attention modules to focus on time and spatial dimensions of the video respectively. Based on the dual focus attention module, we build a network which can easily combine multi-modal features such as action, scene, and objects, and discover the most useful cues for emotion recognition. The experiments conducted on two widely used datasets demonstrates the effectiveness of our dual focus attention network compared with state-of-the-art approaches. For future work, we will employ more effective features such as audio to further improve the performance of video emotion recognition.

6. REFERENCES

[1] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue, "Predicting emotions in user-generated videos," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[2] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal, "Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization," 2016, vol. 9, pp. 255–270.

[3] Chen Chen, Zuxuan Wu, and Yu-Gang Jiang, "Emotion in context: Deep semantic feature fusion for video emotion recognition," in *Proceedings of the 24th ACM International Conference on Multimedia*. ACM, 2016, pp. 127–131.

[4] Haimin Zhang and Min Xu, "Recognition of emotions in user-generated videos with kernelized features," 2018, vol. 20, pp. 2824–2835.

[5] Baohan Xu, Yingbin Zheng, Hao Ye, Caili Wu, Heng Wang, and Gufei Sun, "Video emotion recognition with concept selection," in *2019 IEEE International Conference on Multimedia and Expo*, 2019, pp. 406–411.

[6] Hang-Bong Kang, "Affective content detection using hmms," in *Proceedings of the Eleventh ACM International Conference on Multimedia*. ACM, 2003, pp. 259–262.

[7] René Marcelino Abritta Teixeira, Toshihiko Yamasaki, and Kiyoharu Aizawa, "Determination of emotional content of video clips by low-level audiovisual features," *Multimedia Tools and Applications*, vol. 61, no. 1, pp. 21–49, 2012.

[8] Min Xu, Changsheng Xu, Xiangjian He, Jesse S Jin, Suhuai Luo, and Yong Rui, "Hierarchical affective content analysis in arousal and valence dimensions," *Signal Processing*, vol. 93, no. 8, pp. 2140–2150, 2013.

[9] David G Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE International Conference on Computer Vision*. IEEE, 1999, vol. 2, pp. 1150–1157.

[10] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition)*. IEEE, 2005, vol. 1, pp. 886–893.

[11] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[12] Beth Logan et al., "Mel frequency cepstral coefficients for music modeling.," in *ISMIR*, 2000, vol. 270, pp. 1–11.

[13] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[14] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman, "Video action transformer network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 244–253.

[15] Yunbo Wang, Lu Jiang, Ming Hsuan Yang, Li Jia Li, Mingsheng Long, and Li Fei-Fei, "Eidetic 3d lstm: A model for video prediction and beyond," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[16] Vladyslav Sydorov, Karteek Alahari, and Cordelia Schmid, "Focused attention for action recognition," 2019.

[17] Georgia Gkioxari, Ross Girshick, and Jitendra Malik, "Contextual action recognition with r* cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1080–1088.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[19] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[20] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng, "Dual path networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 4467–4475.

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[23] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[24] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan, "Grid r-cnn," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7363–7372.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[26] Robert Plutchik and Henry Kellerman, *Emotion, theory, research, and experience*, Academic Press, 1980.

[27] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*, vol. 11, Elsevier, 2013.