Video-based Emotion Recognition Using Deeply-Supervised Neural Networks

Yingruo Fan
Department of Electrical and Electronic Engineering,
The University of Hong Kong
Hong Kong, China
fanyingruo@gmail.com

Jacqueline C.K. Lam
Department of Electrical and Electronic Engineering,
The University of Hong Kong
Hong Kong, China
jcklam@eee.hku.hk

Victor O.K. Li
Department of Electrical and Electronic Engineering,
The University of Hong Kong
Hong Kong, China
vli@eee.hku.hk

ABSTRACT

Emotion recognition (ER) based on natural facial images/videos has been studied for some years and considered a comparatively hot topic in the field of affective computing. However, it remains a challenge to perform ER in the wild, given the noises generated from head pose, face deformation, and illumination variation. To address this challenge, motivated by recent progress in Convolutional Neural Network (CNN), we develop a novel deeply supervised CNN (DSN) architecture, taking the multi-level and multi-scale features extracted from different convolutional layers to provide a more advanced representation of ER. By embedding a series of side-output layers, our DSN model provides class-wise supervision and integrates predictions from multiple layers. Finally, our team ranked 3rd at the EmotiW 2018 challenge with our model achieving an accuracy of 61.1%.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**;

KEYWORDS

Emotion Recognition; Deeply-Supervised; Side-output Layers; Convolutional Neural Network; EmotiW 2018 Challenge

ACM Reference Format:
Yingruo Fan, Jacqueline C.K. Lam, and Victor O.K. Li. 2018. Video-based Emotion Recognition Using Deeply-Supervised Neural Networks. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3242969.3264978

1 INTRODUCTION

Video-based emotion recognition is multidisciplinary, and covers the field of psychology, affective computing, and human-robot interaction. In previous EmotiW challenges, most state-of-the-art approaches [1, 7, 11, 16] utilized visual features extracted from facial expression images. In recent years, deep neural networks have been used more widely in visual applications, such as image classification [8, 12], object detection [18, 21] and object tracking [9, 25]. Compared to traditional hand-crafted models [23, 24, 26], an optimal deep learning model is capable of extracting more discriminative features and tackling the expected noises in the sampled database. In the audio-video based sub-challenge, the database consists of video clips extracted from movies or TV shows in real life; this means the expected noises, e.g. head pose, motion blur, face deformation, etc. may occur more frequently in the natural setting.

In 2017 EmotiW Challenge, the winning team [11] put forward a learning strategy, the Supervised Scoring Ensemble (SSE), to connect all probabilities of the fully-connected layers to construct the scoring connection layer. However, the features extracted from the fully-connected layer of CNNs are high-level representations, which might be limited in capturing the full details of facial expressions. Our proposed ensemble method utilizes both deep but semantic and shallow but high-resolution information to provide a more comprehensive representation of features and achieve a better performance. Similar to the runner-up method [16], we also rely on strong CNNs pre-trained on large face recognition datasets and fine-tune them to perform ER. Two potential facial recognition networks are applied to initialize the weights of basic models. As noted in [6], the shallower layers of CNN can learn the rich low-level features, which can be used to help learn the irregular features from the deeper layers. Intuitively, the deeper layers of CNN can capture the semantic information, whilst the shallower layers can capture more profile details. Thus, it is significant to combine both the shallow and the deep layers together to achieve a complementary effect. Inspired by this insight, we propose a Deeply-Supervised CNN (DSN) architecture with a series of side-output layers. A refinement mechanism [10] in cooperation with the skip-connection is employed for fusing all side-output layers. The DSN architecture is shown in Fig. 1. Our novelty includes the use of multi-level and multi-scale features extracted from different layers of a CNN model, and the integration of the upper layers and the lower layers to enhance supervision of the entire framework.

2 THE PROPOSED METHOD

2.1 Deeply-Supervised CNN

The ultimate goal of the proposed deeply-supervised CNN is to enrich the feature map of each layer, by combining the connections across the side-output layers. To this end, we adopt de-convolution techniques [19] in the upsampling operation depicted in Fig. 1,

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3264978

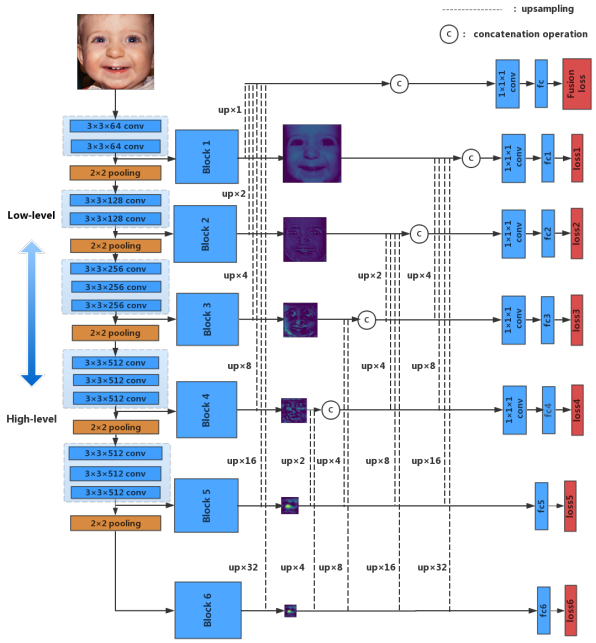

Figure 1: Architecture of the proposed Deeply-Supervised CNN (DSN). Our model takes full advantage of multi-level and multi-scale representations at different convolutional layers, jointly trained to enhance the supervision of the entire framework.

which can take input of an arbitrary size and produce correspondingly size output. Let $A_{side}^{(m)}$ denote the side activations at the mth side-output layer before the upsampling operations are implemented. M is the number of side-output layers. Then new side activations $R_{side}^{(m)}$ can be defined as

$$R_{side}^{(m)} = \begin{cases} \sum_{i=m+1}^{M} r_i^m R_{side}^{(i)} + A_{side}^{(m)}, m = 1, 2, \ldots, M-1 \\ A_{side}^{(m)}, m = M \end{cases}, \quad (1)$$

where r_i^m is the weight of the connection from the side-output layer i to the side-output layer m ($i > m$). When r_i^m is set to 0, the corresponding connection is dropped.

Each side-output layer is optimized by a softmax loss function. Hence, the objective function of all side-output layers is given by

$$L_{side}(W, r) = \sum_{m=1}^{M} \alpha_m l_{side}^{(m)}(W, r), \quad (2)$$

where α_m is the weight of the mth side loss, $l_{side}^{(m)}(\cdot)$ represents the softmax loss function for the mth side output, and W indicates the collection of all standard network layer parameters. Note that $r = r_i^m$, $i > m$. Besides, the fusion loss (Fig. 1) is formulated as

$$L_{fuse}(W, r) = h(\sum_{m=1}^{M} R_{side}^{(m)}), \quad (3)$$

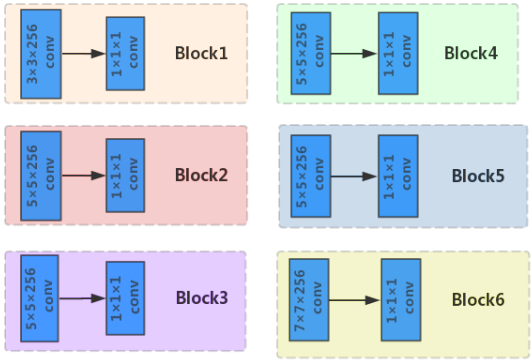

Figure 2: Supervision blocks in VGG-Net.

where $h(\cdot)$ is the standard softmax loss, and $R_{side}^{(m)}$ are activations of the mth side output. The final loss function is calculated by the summation of L_{fuse} and L_{side}. During the training process, L_{fuse} and L_{side} are optimized jointly to enforce supervision across all convolutional layers, thus improving the learning efficiency of network parameters.

At the testing phase, by fusing multiple feature maps, the predicted scores produced by the fusion layer are computed by

$$P_{fuse} = \sigma(\sum_{m=1}^{M} R_{side}^{(m)}), \quad (4)$$

where $\sigma(\cdot)$ denotes the softmax function. On the other hand, each side output is represented as $P_{side}^{(m)} = \sigma(R_{side}^{(m)})$. We re-formulate the final predicted scores by including P_{fuse} and $P_{side}^{(m)}$ as follows

$$P_{final} = \frac{1}{1+M}(P_{fuse} + \sum_{m=1}^{M} P_{side}^{(m)}). \quad (5)$$

Overally, DSN is trained progressively via supervised learning from the progressively improved information of the intermediate layers. In other words, the entire DSN framework is optimized by the fusion loss and the companion loss of the side-output layers.

2.2 Networks

We experiment with three state-of-the-art networks : VGG-Face [20], ResNet-50 [8] and DenseNet-121 [12]. Note that we only integrate supervision blocks in VGG-Face and ResNet-50. In the case of VGG-Face, the connection pattern of side-output layers is shown in Fig. 1, and the details of supervision blocks are shown in Fig. 2. According to Eq. 1, the side activations of VGG-Face can be formulated as

$$\tilde{R}_{side}^{(m)} = \begin{cases} \Sigma_{i=3}^{6} r_i^m \tilde{R}_{side}^{(i)} + \hat{A}_{side}^{(m)}, m = 1, 2 \\ r_5^m \tilde{R}_{side}^{(5)} + r_6^m \tilde{R}_{side}^{(6)} + \hat{A}_{side}^{(m)}, m = 3, 4 \\ \hat{A}_{side}^{(m)}, m = 5, 6 \end{cases}. \quad (6)$$

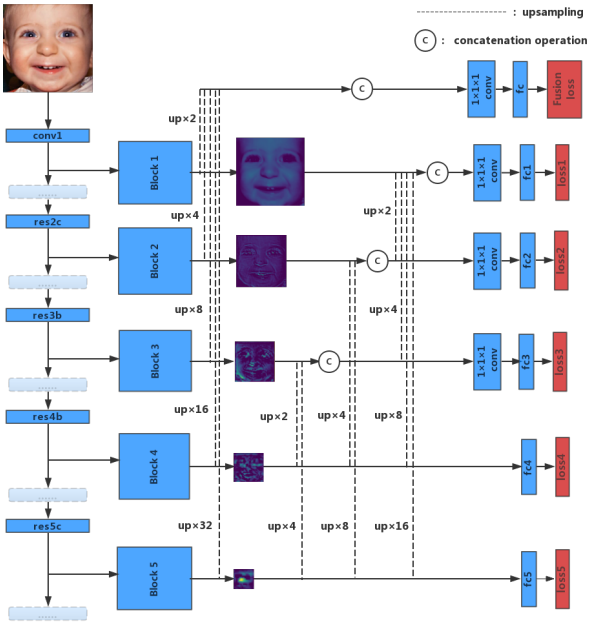

Figure 3: A deeply-supervised CNN architecture based on ResNet-50, consisting of five side-output layers.

Similarly, Fig. 3 and Fig. 4 display the configuration details for ResNet-50, the side activations of which are defined as

$$\tilde{R}_{side}^{(m)} = \begin{cases} \Sigma_{i=2}^{5} r_i^m \tilde{R}_{side}^{(i)} + \hat{A}_{side}^{(m)}, m = 1 \\ r_4^m \tilde{R}_{side}^{(4)} + r_5^m \tilde{R}_{side}^{(5)} + \hat{A}_{side}^{(m)}, m = 2, 3 \\ \hat{A}_{side}^{(m)}, m = 4, 5 \end{cases} . \quad (7)$$

For convenience, we refer to the modified networks above as DSN-VGG-Face and DSN-ResNet-50, respectively. The connection coefficient r_i^m is set to 1 in both cases (Eq. 6 and Eq. 7). The pre-trained models, i.e. VGG-Face [20], ResNet-50 [8] and DenseNet-121 [12], are employed to initialize the network parameters. Besides, we utilize a facial recognition network which is trained on around six million images of human faces. This network, denoted as FR-Net, is empirically fine-tuned on the emotion datasets, replacing the last fully connected layer with a new layer in the following experiments. The same applies to DSN-VGG-Face , DSN-ResNet-50, and DenseNet-121.

3 EXPERIMENTS

The training and testing processes are performed on NVIDIA GeForce GTX 1080Ti 11G GPUs, using the deep learning framework Caffe[13]. In this section, we first introduce the facial expression recognition datasets used for this challenge and the corresponding data processing. Next, we describe the implementation details, including data augmentation strategy, parameter setting and training method. Finally, results of our model presented to EmotiW 2018 are shown.

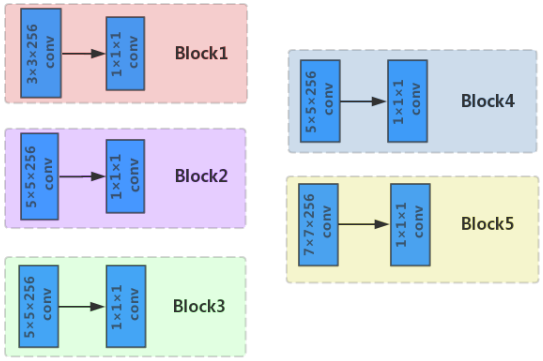

Figure 4: Supervision blocks in ResNet-50.

3.1 Data Pre-processing

3.1.1 Datasets. Recently, Real-world Affective Faces Database (RAF-DB) [17],which collects the real-world facial expression images from thousands of individuals, has been released to encourage more real-world research on facial expression recognition. RAF-DB contains about 12271 training samples and 3068 test samples downloaded from the Internet, providing spontaneous expressions under different environmental conditions. The other database, Acted Facial Expressions in the Wild (AFEW) [3], is established for the Emotion Recognition in the Wild Challenge (EmotiW). It consists of training (773), validation (383) and test (653) video clips collected from TV shows or movies. The samples in the two datasets are labelled with seven emotions: angry, disgust, fear, happy, sad, surprise and neutral.

3.1.2 Face Detection. In the AFEW training set, we use the open-source multimedia tool FFmpeg[1] to split the video into individual frames, which are processed further by filtering the blurred facial images. To extract and align faces from the video frames, we employ a C++ library, Dlib [15] face detector to locate the 68 facial landmarks. Aligned and cropped face images (224 × 224 pixels) are generated in a uniform template with a affine transformation, in order to reduce the influence of in-plane rotation. The same operation is also applied to images in RAF-DB.

3.2 Implementation details

The four networks aforementioned are first fine-tuned on the training set of RAF-DB. We augment the training images by flipping, rotating each by ±2°, ±4° and ±6°, and adding Gaussian white noises with variances of 0.001, 0.005, 0.01 and 0.015. In addition to offline data augmentation, the input samples are also randomly cropped from the corners and center of the images during the training process. Meanwhile, during the testing phase, the prediction value is the average of the values of the five cropped parts obtained from each sample. All four networks are trained with a linear learning rate decay in stochastic gradient descent (SGD) [2] optimizer, with

[1] https://www.ffmpeg.org/

Table 1: The accuracy (%) of our top 4 submissions to EmotiW 2018.

Model	Validation	Test
DSN-VGG-Face + DSN-ResNet-50 + DenseNet-121 + FG-Net	55.13	58.04
DSN-VGG-Face + VGG-Face + DSN-ResNet-50 + DenseNet-121 + FG-Net	56.05	59.57
3rd fusion of DSN-VGG-Face + VGG-Face + DSN-ResNet-50 + DenseNet-121 + FG-Net	56.97	59.87
5th fusion of DSN-VGG-Face + VGG-Face + DSN-ResNet-50 + DenseNet-121 + FG-Net	57.43	61.10

Table 2: Network training parameters (Arc. = Architecture, ILR = Initial learning rate, BS = Batch size, MI = Maximum iterations, WD = weight decay, M = momentum).

Arc.	ILR	BS	MI	WD	M
DSN-VGG-Face	0.0001	16	30k	0.0005	0.9
DSN-ResNet-50	0.01	16	30k	0.0001	0.9
DenseNet-121	0.01	8	50k	0.0001	0.9
FG-Net	0.001	32	20k	0.0001	0.9

Table 3: Model accuracy (%) of the validation dataset (Original refers to the basic network without embedding the side-output layers).

Model	Original	Deeply-supervised framework
VGG-Face	45.16	48.04
ResNet-50	40.99	43.86
DenseNet-121	44.13	-
FG-Net	47.00	-

the training parameters shown in Table 2. Specially, for DSN-based networks, i.e., DSN-VGG-Face and DSN-ResNet-50, the α_m in Eq. 2 is set to 1 for all side-output layers. During this stage, the networks should achieve an accuracy of $79 - 84\%$ based on the validation set of RAF-DB. Next, we use the processed data w.r.t. AFEW to re-perform the training process, with the same network training parameters shown in Table 2. Finally, the summation of the predicted emotion scores over all frames is taken as the final score of this model for each testing video clip.

Regarding the parameters of the supervision blocks (Fig. 2 and Fig. 4), assuming that all the first convolutional layers have C channels, we experiment at different values of C (C = 128, 256 and 512) and obtain the best performance of our model when C = 256.

3.3 Results on EmotiW 2018

As expected, incorporating the deeply-supervised blocks in VGG-Face and ResNet-50 (Table 3) boosts the performance on the local validation set in EmotiW 2018. Our submissions to the audio-video based emotion recognition sub-challenge are shown in Table. 1. We adopt the class-wise ensemble method, where multiple scores for each emotion category are summed up based on the corresponding weights. Different assigned weights are investigated based on grid searching. It can be concluded that the single models given in Table. 1 exhibit different recognition capabilities, such that the final ensemble can obtain better performance as a result of the complementary effects.

For additional insight concerning the category classification, Fig. 5 displays the confusion matrix of our best submitted model. Our final ensemble model performs relatively better in recognizing the category of Happy and Surprise. This result is consistent with previous studies [17]. To provide clearer insights, we employ a facial action coding system (FACS) analysis [22], where the changes in facial behaviors are labeled with the codes of action units (AUs) [5]. Given that Happy and Surprise facial expressions might consist of more distinguishable AUs, e.g. the FACS codes of Happy are

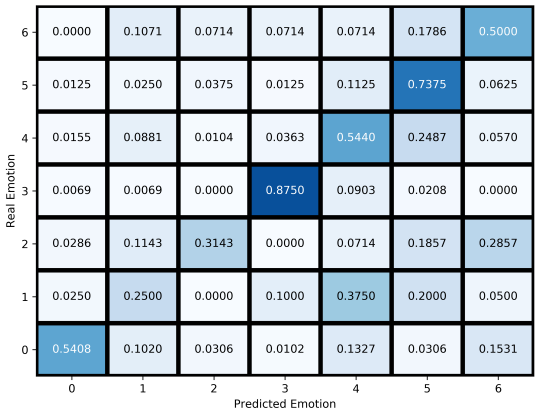

Figure 5: Confusion matrix obtained with the best result of the test set. Note that the numbers represent seven facial expressions: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise and 6=Neutral. 'Real Emotion' denotes the true value based on human labelling while 'Predicted Emotion' denotes the predicted value.

AU6 (cheek raiser) and AU12 (lip corner puller), these two expressions would exhibit distinctive intensity variations. Hence, CNN can learn these distinguishable AUs to identify the Happy and Surprise emotion. Particularly, our DSN ensemble model has obtained a much higher accuracy of 73.75% for Surprise, which significantly outperforms the results reported by previous top-winning teams [11, 14, 16] (e.g. 73.75% vs. 19.57% [11]). This can be partially attributed to the rich AUs involved in expressing Surprise in RAF-DB. In addition, imbalanced distribution of facial expression classes in different datasets may also contribute to the relatively better ER for Surprise.

4 CONCLUSION

We present our approach to the Audio-video sub-challenge in EmotiW 2018 [4]. Our proposed DSN model learns multi-level and multi-scale representations to fully exploit the powerful hierarchical features of CNNs. Moreover, our model provides class-wise supervision to integrate the prediction results from multiple layers, which takes full advantage of the complementary characteristics. Finally, our ensemble method achieves an outstanding performance based on the testing set provided by EmotiW 2018, with an accuracy of 61.10%. Last but not the least, facial expression is one of the many modes for detecting emotions. In the future, we shall combine this method with other ER methods based on other modalities, e.g. electroencephalogram (EEG) based on heart rates, to develop a more complementary and high-level framework for ER.

ACKNOWLEDGEMENTS.

This research is supported in part by the Theme-based Research Scheme of the Research Grants Council of Hong Kong, under Grant No. T41-709/17-N.

REFERENCES

[1] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang. 2016. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 433–436.

[2] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 177–186.

[3] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia* 19, 3 (2012), 34–41.

[4] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. 2018. EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction (in press). In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM.

[5] Paul Ekman. 2002. Facial action coding system (FACS). *A human face* (2002).

[6] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. 2018. Multi-Region Ensemble Convolutional Neural Network for Facial Expression Recognition. *arXiv preprint arXiv:1807.10575* (2018).

[7] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 445–450.

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

[9] Zhiqun He, Yingruo Fan, Junfei Zhuang, Yuan Dong, and HongLiang Bai. 2017. Correlation Filters with Weighted Convolution Responses.. In *ICCV Workshops*. 1992–2000.

[10] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. 2017. Deeply supervised salient object detection with short connections. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 5300–5309.

[11] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. 2017. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 553–560.

[12] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Vol. 1. 3.

[13] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 675–678.

[14] Dae Ha Kim, Min Kyu Lee, Dong Yoon Choi, and Byung Cheol Song. 2017. Multimodal emotion recognition using semi-supervised learning and multiple neural networks in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 529–535.

[15] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10, Jul (2009), 1755–1758.

[16] Boris Knyazev, Roman Shvetsov, Natalia Efremova, and Artem Kuharenko. 2017. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. *arXiv preprint arXiv:1711.04598* (2017).

[17] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2584–2593.

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[20] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. 2015. Deep Face Recognition.. In *BMVC*, Vol. 1. 6.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[22] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2017. Learning bases of activity for facial expression recognition. *IEEE Transactions on Image Processing* 26, 4 (2017), 1965–1978.

[23] Caifeng Shan, Shaogang Gong, and Peter W McOwan. 2009. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing* 27, 6 (2009), 803–816.

[24] Karan Sikka, Tingfan Wu, Josh Susskind, and Marian Bartlett. 2012. Exploring bag of words architectures in the facial expression domain. In *European Conference on Computer Vision*. Springer, 250–259.

[25] Chong Sun, Dong Wang, Huchuan Lu, and Ming-Hsuan Yang. 2018. Learning spatial-Aware regressions for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8962–8970.

[26] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. 2012. Learning active facial patches for expression analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2562–2569.