Face-Focused Cross-Stream Network for Deception Detection in Videos

Mingyu Ding[1,*] An Zhao[1,*] Zhiwu Lu[1,†] Tao Xiang[2,3] Ji-Rong Wen[1]
[1]Beijing Key Laboratory of Big Data Management and Analysis Methods
School of Information, Renmin University of China, Beijing 100872, China
[2]Department of Electrical and Electronic Engineering, University of Surrey, UK
[3]Samsung AI Centre, Cambridge, UK
zhiwu.lu@gmail.com t.xiang@surrey.ac.uk

Abstract

Automated deception detection (ADD) from real-life videos is a challenging task. It specifically needs to address two problems: (1) Both face and body contain useful cues regarding whether a subject is deceptive. How to effectively fuse the two is thus key to the effectiveness of an ADD model. (2) Real-life deceptive samples are hard to collect; learning with limited training data thus challenges most deep learning based ADD models. In this work, both problems are addressed. Specifically, for face-body multimodal learning, a novel face-focused cross-stream network (FFCSN) is proposed. It differs significantly from the popular two-stream networks in that: (a) face detection is added into the spatial stream to capture the facial expressions explicitly, and (b) correlation learning is performed across the spatial and temporal streams for joint deep feature learning across both face and body. To address the training data scarcity problem, our FFCSN model is trained with both meta learning and adversarial learning. Extensive experiments show that our FFCSN model achieves state-of-the-art results. Further, the proposed FFCSN model as well as its robust training strategy are shown to be generally applicable to other human-centric video analysis tasks such as emotion recognition from user-generated videos.

1. Introduction

With the recent rapid development of human-centric AI, human-centric video analysis [48, 49, 54, 30, 32, 61, 27] has also begun to draw much attention from the computer vision community. Other than the conventional video content analysis that focuses on generic semantic concept analysis of video content, human-centric video analysis aims to extract, describe, and organize a wealth of information regarding the main objects of interest in most videos: humans. This topic covers a wide range of research problems such as deception detection [36, 37], emotion recognition in videos [54, 18], personality computing [49, 56], and action recognition [6, 26, 28, 35, 40, 47, 59]. For example, it is often important to recognize the deceptive behaviors [36, 37], emotions [54, 18], or personality traits [49, 56] of the subject of a video in real-world scenarios.

Deception detection [36, 37] is a late addition to human-centric video analysis and still under-studied. Deception is defined as an intentional attempt to mislead others [4]. In our day-to-day life, deceptive behaviors occur in the form of intended lies, fabrications, omissions, misrepresentations, among others. Some deceptive behaviors are simply harmless, but others may have major threats to the society, e.g., those taking place in a courtroom. Detecting real-world human deceptive behaviors is a challenging task even for humans, and often requires well-trained human experts. A large-scale deployment of deception detection thus depends upon automated deception detection (ADD) [36, 37]. An ADD system can find applications in many real-world scenarios including airport security screening, court trial, job interview, and personal credit risk assessment.

The ADD task faces two major challenges. (1) **Multimodal fusion:** As a subtle human behavioral trait, deception is hard to detect in real-life scenarios. Its reliable detection needs to resort to multiple modalities including the visual, verbal, and acoustic [14, 16, 22, 12, 21]. Among them, the visual modality is considered to be the most informative one. Multiple visual cues also exist visually. In particular, facial expressions [58, 33] and body motions [51, 31] are typically the focus of visual analysis. An important problem thus arises: How to effectively fuse these modalities/cues? Such a fusion is not straightforward because they not only have different strengths in each individual video sequence, but also are temporally asynchronized. An example of the asynchronization between the face and body cues is shown in Figure 1. (2) **Data scarcity:** Unlike the conventional physiological and biological methods [44, 5, 19, 45, 8], an

*Equal contribution.
†Corresponding author.

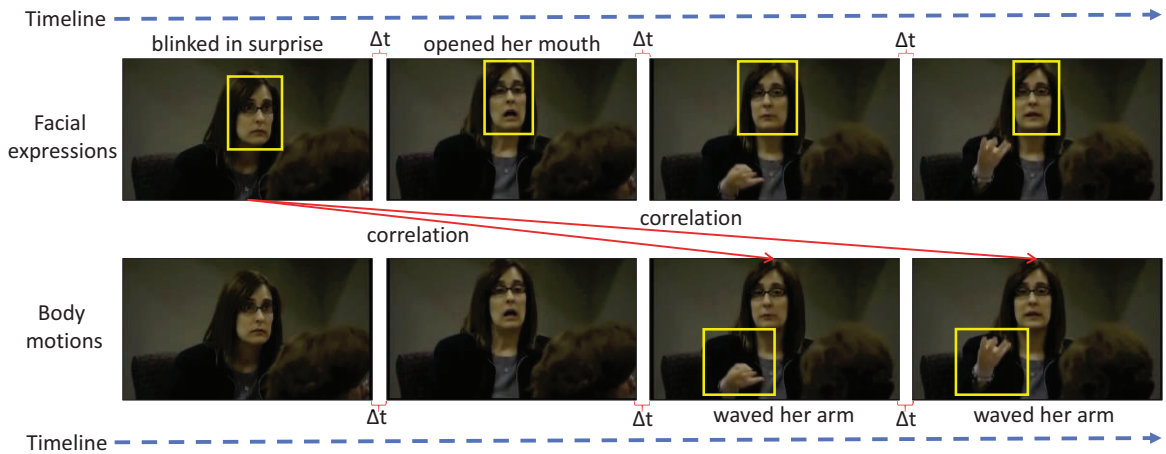

Figure 1. Illustration of the asynchronization between facial expressions and body motions. It can be seen that a subject who lies tends to first have a surprised expression before she/he is aroused to take body/hand movements. Notation: $\Delta t = 2$–3 frames.

ADD model is non-contact and non-invasive. This indirectness means that collecting large quantity of high-quality data containing samples of deceptive behaviors is critical. Earlier data collection efforts focused on human contributors in a lab or in a crowdsourcing setting. In other words, they are staged; the usefulness of these datasets for real-world deployment is thus questionable. Recently, the focus of ADD has been towards detecting deceptive behaviors from real-life data. Particularly, a new multimodal benchmark dataset of real-life videos from court trials is introduced in [36, 37]. However, with only 121 video clips and half of them containing deception, this dataset is insufficient for training a deep neural network based model that has dominated the recent ADD approaches [9, 20].

We address both problems in this paper. For the multimodal fusion problem, we propose a novel face-focused cross-stream network (FFCSN). Different from the popular two-stream networks [40, 6, 47, 35], our FFCSN model has two novel components: (a) Face detection is added into the spatial stream subnet to capture the facial expressions explicitly. (b) Correlation learning is performed across the spatial and temporal streams for joint deep feature learning from facial expressions and body motions. Importantly, our model is able to cope with the asynchronization/temporal inconsistency between facial expressions and body motions (see Figure 1). For the training data scarcity problem, we introduce meta learning [39, 52, 25] and adversarial learning [11, 23, 24] into the training process of our FFCSN. Meta learning, based on the principle of learning to learn, is deployed here to improve the generalization ability of the model and avoid overfitting to the insufficient training data. In the meantime, adversarial learning based feature synthesis is adopted as a data augmentation strategy. When these two are combined, our FFCSN can be trained effectively even with the very sparse data in the existing real-life deception detection benchmarks [36, 37].

Our contributions are three-fold: (1) We have proposed a novel face-focused cross-stream network (FFCSN) for joint deep feature learning from facial expressions and body motions in real-life videos. Comparing to existing two-stream networks, our FFCSN model is uniquely able to cope with the asynchronization/temporal inconsistency between facial expressions and body motions. (2) To avoid model overfitting and improve generalization ability, meta learning and adversarial learning are introduced into the training process of FFCSN. (3) We demonstrate that our FFCSN model can be easily extended to other human-centric video analysis problems such as emotion recognition from user-generated videos [54, 18]. Extensive experiments are carried out on benchmark datasets and the results show that our model clearly outperforms existing state-of-the-art alternatives.

2. Related Work

Video-Based Deception Detection. Earlier works on video-based ADD are limited by the datasets which contain only staged deceptive behaviors [14, 16, 22, 12, 21]. Their usefulness for detecting real-life deception is thus in doubt. The change towards deception detection with real-life data was first advocated in [7], where the identification of deception in statements issued by witnesses and defendants is targeted using a corpus collected from hearings in Italian courts (i.e., no visual data was available). In [36, 37], a new multimodal deception dataset of real-life videos from court trials was first introduced, and the combination of features extracted from different modalities is used for deception detection. Thanks to this benchmark dataset, more advancing ADD methods [17, 1, 50] have been developed to leverage multimodal features for detecting deception.

Deep Learning for Deception Detection. Recent ADD methods typically benefit from the latest development in deep neural networks [9, 20]. However, it is noted in [50]

that, given the small size of the real-life ADD benchmark introduced in [36, 37], hand-crafted features are much better than deep features. This is not surprising: deep learning models are known to be data hungry. The real-life ADD dataset in [36, 37] only provides around 100 video clips, which is a number of magnitudes smaller than, for example, those YouTube-collected action recognition benchmark datasets such as UCF101 [41]. Our model differs significantly from existing deep ADD models in that the data scarcity problem is addressed explicitly, based on a meta learning and adversarial learning based training strategy. Adversarial learning [11, 23, 24] has recently been used as a data augmentation strategy to deal with the lack of training data. However, meta-learning [39, 52, 25] was originally proposed for transfer learning. Here, we re-purpose it for learning with scarce data and uniquely combine it with adversarial learning to cope with the extreme challenge of data scarcity in ADD. We show in experiments that our model outperforms [17, 1, 50] by big margin, thanks to the proposed training strategy (see Tables 1 and 2).

Two-Stream Network. Our FFCSN model adopts a two-stream network architecture, one for RGB still frame modeling and the other for optical flow extracted from consecutive frames. Such a two-stream architecture was originally proposed for action recognition in videos and has been popular for many human-centric video analysis tasks [40, 6]. Various improvements such as temporal segment network (TSN) [47] and its variants [60, 62] have been designed by capturing the long-range temporal structure and learning the ConvNet models with limited training samples. Similarly, [35] proposed to add faster R-CNN [38] so that attention can be focused on objects detected in a video. Our FFCSN model is different from existing two-stream models in that: (1) face detection is added into the spatial stream subnet to capture the facial expressions explicitly; (2) correlation learning is performed across the spatial and temporal streams to cope with the temporal inconsistency between facial expressions and body motions for ADD.

Video-Based Emotion Recognition. Deception detection is closely related to emotion recognition: deception could be considered as a specific emotion state of humans, albeit it is much more subtle and harder to detect than others such as happy and angry. Note that emotion recognition from user-generated videos [18] is a challenging problem. Because of the complicated and unstructured nature of user-generated videos and the sparsity of video frames that express the emotion content, it is often hard to understand emotions conveyed in user-generated videos. To address this problem, multi-modal fusion and knowledge transfer approaches have been proposed in recent works [34, 53, 57, 54]. In this paper, we show that our FFCSN model can be easily extended to emotion recognition from user-generated videos, with state-of-the-art results achieved.

3. Methodology

As illustrated in Figure 2, our full FFCSN model for video-based ADD consists of three main modules: face-focused cross-stream network including a facial expression branch as well as a body motion branch, meta learning module, and adversarial learning module. In the following, we give the details of the three main modules.

3.1. Cross-Stream Network Module

In this work, we focus on joint deep feature learning from facial expressions and body motions for video-based ADD. Different from the traditional video-based action recognition, the facial expressions and body motions of a subject are found to be related to his/her deceptive behaviors [58, 33, 51, 31], rather than the whole frame appearance. Therefore, we choose to modify the original two-stream temporal segment network [47] designed for video-based action recognition by replacing its appearance branch with a face expression branch (see Figure 2).

3.1.1 Cross-Stream Base Network

The spatial stream (i.e. face expression branch) is a face detection model based on the popular faster R-CNN [38]. This branch follows the deep learning framework of faster R-CNN, which has been shown to achieve state-of-the-art results in generic object detection. As illustrated in Figure 2, it essentially consists of two parts: (1) a region proposal network (RPN) for generating a list of region proposals which may contain objects, called regions of interest (RoIs); (2) a R-CNN network for classifying the regions of each frame into objects and refining the boundaries of these regions. The two parts share common parameters in the convolutional layers used for feature extraction, allowing it to accomplish the face detection task efficiently.

In our model, faster R-CNN is generalized for both face detection and expression feature extraction. Note that the traditional faster R-CNN takes only 9 anchors, which sometimes fails to recall small objects. For our face detection task, however, small faces tend to be fairly common. We thus add a size group of 64×64 and increase the number of anchors to 12. In this paper, the RPN batchsize is set to 256, and the ResNet50 [13] is used as the backbone model for the face expression branch.

The temporal stream (i.e. body motion branch) operates on a stack of consecutive warped optical flow fields to capture the motion information. Inspired by the representative work on improved dense trajectories [46], we extract the warped optical flow by first estimating the homography matrix and then compensating the camera motion. This branch can thus avoid concentrating on the camera motion but not on the body motion. As shown in Figure 2, ResNet50 is used to compute the temporal feature maps.

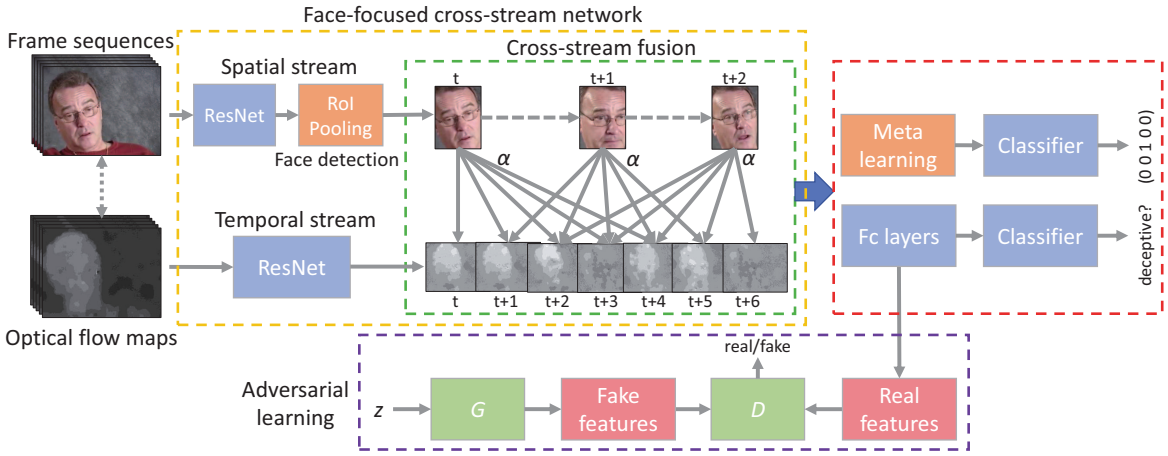

Figure 2. Overview of our full FFCSN model. Note that α collects the correlation scores of matching the spatial and temporal streams. In such cross-stream fusion, we select 5 frames with the same interval. In this figure, we use [t, t+1,..., t+5] for easily understanding.

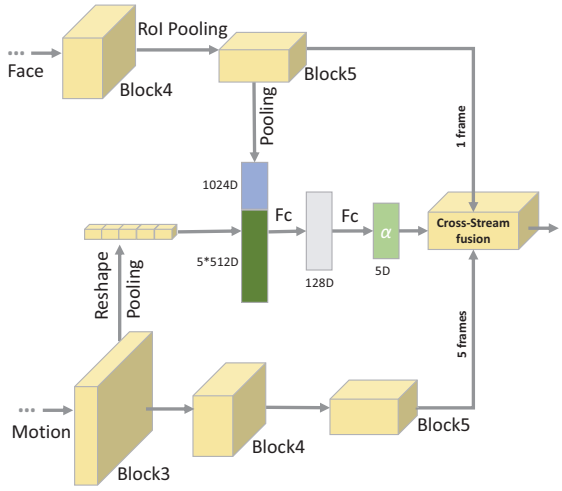

Figure 3. Architecture of our cross-stream fusion block.

3.1.2 Cross-Stream Fusion

In our cross-stream base network, one stream focuses on the face that is only a part of the whole frame, and the other focuses on the body motion that is captured using multiple whole frames. These two parts are clearly complementary to each other. We thus combine the two branches by cross-stream fusion as illustrated in Figure 3. Importantly, the fusion is based on deep correlation analysis rather than simply concatenating the feature vectors extracted from the two streams as in conventional two-stream networks. Specifically, to cope with the asynchronization/temporal inconsistency between facial expressions and body motions (see Figure 1), we choose to learn the correlation among adjacent frames (only 5 adjacent frames are considered here). For the spatial stream, we downsample the feature maps of the final residual block of ResNet50 [13] in the dimension of depth and obtain a 1024-dimension feature vector. For the temporal stream, given that five motion frames are matched to one face frame, we utilize the reshape pooling to obtain one 5×512-dimension feature vector after the third residual block of ResNet50. The outputs of the two streams are then concatenated and fed into two fully-connected layers (with the dimension of 128 and 5, respectively). Finally, we compute the correlation scores $\alpha = [\alpha_1, ..., \alpha_5]$ for the five face-motion pairs using the softmax function, and weight them with α for final two-stream fusion.

To extract deep visual features from a long-term video, our model essentially works on a sequence of short snippets sparsely sampled from the entire video. After each snippet of this sequence predicts its own result, a consensus among all snippets is obtained as the final video-level prediction. For all obtained video-level predictions, we can define a segmental consensus classification loss similar to that of temporal segment network [47]. Specifically, we divide each video into K segments $\{S_1, S_2, \ldots, S_K\}$ of equal duration. From each segment S_k ($k = 1, ..., K$), we then randomly sample a short snippet T_k. In our problem, the short snippet T_k consists of one spatial frame (denoted as $T_k(\mathrm{sf}_1)$) and five temporal frames (denoted as $T_k(\mathrm{tf}_1), ..., T_k(\mathrm{tf}_5)$). Suppose that all snippets/frames have been represented as feature maps here. We thus have $T_k = [T_k(\mathrm{sf}_1), \sum_{j=1}^{5} \alpha_j T_k(\mathrm{tf}_j)]$. Let $\mathcal{F}(T_k; W)$ denote the classification probability predicted by our model with parameters W for T_k. The outputs of all short snippets are combined by the segmental consensus function $\mathcal{E}$ to obtain a consensus of prediction among them. With the softmax loss, the overall loss of our model is defined as:

$$L_{BASE}(y, E) = -\sum_{i=1}^{N_c} y_i (E_i - \log \sum_{j=1}^{N_c} \exp E_j), \qquad (1)$$

where E is the segment consensus computed by $E = \mathcal{E}(\mathcal{F}(T_1; W), \mathcal{F}(T_2; W), \ldots, \mathcal{F}(T_K; W))$, N_c is the number of target classes ($N_c = 2$ in our problem), and y_i is the ground truth label with respect to class i. We define the consensus function $\mathcal{E}$ with average pooling, as in [47].

3.2. Meta Learning Module

Deception detection is a challenging task due to the subtle differences between truthful and deceptive behaviors. Learning to differentiate the two types of behaviors with only a handful of samples of each is extremely challenging. This is especially true when the behaviors are modeled with deep neural networks with a large number of model parameters. To deal with the data scarcity problem, we propose to use meta learning [39, 52, 25, 42] to train our FFCSN (see Figure 2). To best utilize the limited training samples, we introduce pair-wise comparison of them. Specifically, our cross-stream base network can be viewed as the encoding submodule f of our meta learning module. A comparison submodule g is then introduced for meta learning. The meta learning pipeline is illustrated in Figure 4. Examples of the two classes (yellow deceptive and blue truthful) are shown in different colors. In this case, the meta-train set contains five samples (four truthful and one deceptive). The deceptive sample in the meta-validation set is used to form five pairs with the meta-train samples. The final model output, after softmax, is a 5D logit vector supervised to produce a close-to-one value in the third element and close-to-zero values in all other elements. This meta-learning pipeline turns a two-class (deceptive/deceptive) classification problem into a multi-case classification problem and makes full use of the limited training samples.

Formally, in each mini-batch (with the mini-batch size N_b), videos x_i $(i = 1, 2, \ldots, N_b)$ are fed through the encoding submodule f, which outputs the concatenated feature maps $f(x_i)$ $(i = 1, 2, \ldots, N_b)$. We split the videos in the mini-batch into the meta-train and meta-validation sets. A sample x_a is randomly chosen from the meta-validation set. The output $f(x_a)$ is combined with each $f(x_j)$ $(j \neq a)$ in the meta-train set using the operator $\mathcal{C}(f(x_a), f(x_j))$. In our meta learning module, we set $\mathcal{C}(\cdot, \cdot)$ as the concatenation of feature maps in the dimension of depth. The combined feature maps of the sample pairs are fed into the comparison submodule g, which produces a pairwise score representing the similarity between x_a and x_j. We thus generate the pairwise scores for each mini-batch as:

$$r_{a,j} = g(\mathcal{C}(f(x_a), f(x_j))), j \neq a. \quad (2)$$

We train our meta learning module by fitting the pairwise score $r_{a,j}$ to the ground truth pairwise similarity with a cross entropy loss as follows:

$$L_{ML} = \frac{-1}{N_b - 1} \sum_{j \neq a} y_j \log(r_{a,j}) + (1 - y_j) \log(1 - r_{a,j}), \quad (3)$$

where $y_j = 1$ if (x_a, x_j) is an intra-class sample pair, and $y_j = 0$ if (x_a, x_j) is an inter-class sample pair.

As illustrated in Figure 4, the encoding submodule of our meta learning module is just our cross-stream base network. In the following, we give the details of the comparison submodule of our meta learning module. Specifically, the comparison submodule consists of two convolutional blocks and two fully-connected layers: (1) Each convolutional block has a 3×3 convolution layer followed by batch normalization and RELU activation. The number of filters of the convolution layer in the first block is 512 and the number of filters in the second block is 128. The output size of the two blocks is $128 \times 3 \times 3 = 1,152$. (2) The two fully-connected layers are 8 and 2 dimensional, respectively.

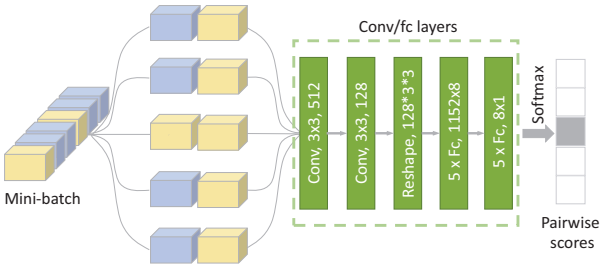

Figure 4. Architecture of the meta learning module used in our full FFCSN model for video-based ADD. See text for details.

3.3. Adversarial Learning Module

In this paper, we aim to synthesize feature vectors for data augmentation in the ADD task. Note that synthesizing raw videos explicitly is an unsolved problem in itself. Therefore, we choose to generate a 256-dimension feature vector for each synthesized video instead, which is a much easier task. In particular, we propose to synthesize *fake feature vectors and attack the classifier* for deception detection during training of our full FFCSN model, in order to overcome the training data scarcity problem.

Adversarial training involves a discriminator and a generator. In our case, the discriminator network aims to classify the inputs into two classes: real or fake. In this paper, the observed variable x is the 256-dimensional vector produced by our cross-stream base network. Given that the discriminator network D consists of 3 fully-connected layers with the ELU activation, $D(x)$ thus denotes the probability that x comes from the real (but not fake) class.

As for the generator network G of our adversarial learning module, the input 32-dimensional noise z is sampled from a zero-mean Gaussian distribution $p_z(z)$ with the standard deviation 1. We use 3 hidden layers to represent G with the size 32, 64, and 256, respectively. The first fully-connected layer uses the ELU activation, and the second fully-connected layer uses the sigmoid activation. $G(z)$ denotes a generated sample drawn from the data space.

The adversarial training of D and G can be formulated as the following min-max problem:

$$\min_G \max_D L_{AL}(G, D) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))], \quad (4)$$

where L_{AL} denotes the loss function of our adversarial learning module and $p_{\text{data}}(x)$ denotes the data distribution.

3.4. Training Process

Our full FFCSN model for video-based ADD is trained using an end-to-end training strategy. The loss function of our full FFCSN model is defined as follows:

$$L = L_{BASE} + \beta_1 L_{ML} + \beta_2 L_{AL}, \quad (5)$$

where β_1 and β_2 denote the hyper-parameters. In this paper, we empirically set $\beta_1 = \beta_2 = 1$ in all experiments.

4. Experiments

4.1. Video-Based Deception Detection

4.1.1 Dataset and Setting

Real-Life Dataset. We evaluate our full FFCSN model for deception detection on a real-life multimodal dataset [36]. This dataset consists of 121 court room trial video clips. Since videos from this trial dataset are collected under unconstrained conditions, we need to cope with the change of the viewing angle of the person, the variation in video quality, and the background noise. In this paper, we select a subset of 104 videos from the original trial dataset, including 50 truthful videos and 54 deceptive videos, as in [50].

Evaluation Setting. Our dataset consists of only 58 identities. Since the number of identities is smaller than the number of video clips, the same identity often appear in both deceptive truthful clips. When the videos of the same identity are divided into both the training and test sets, a deception detection method tends to suffer from over-fitting to identities. To address this over-fitting issue, we perform 10-fold cross validation over identities (but not over video samples) as in [50], which ensures that the identities in the test set have no overlap with that in the training set.

Evaluation Metrics. To evaluate the performance of a deception detection method, we compute two metrics as follows: (1) ACC – the classification accuracy (ACC) over the test video samples; (2) AUC – the area under the precision-recall curve (AUC) over the test set, which is originally defined to cope with the imbalance of the positive and negative classes. The former has been widely used in previous research on deception detection [36, 37, 17, 9], while the latter is mainly used in recent works [50, 20].

Network Initialization. We pretrain the face branch of our cross-stream base network using the WIDER-FACE [55] and CK+ [29] datasets, and then pretrain the motion branch of our cross-stream base network as in [47] on the UCF101 [41] dataset. Moreover, for G and D of the adversarial learning module, we adopt the Kaiming initialization. All the other layers are randomly initialized by drawing weights from a zero-mean Gaussian distribution with the standard deviation 0.01 (along with zero biases).

Implementation Details. After network initialization, our full FFCSN model is trained in an end-to-end manner using back-propagation and stochastic gradient descent. The learning rate is set to 0.0005 for the first 10 epochs, and then reduced to one tenth with a step size of 10 (epochs). The maximum number of epochs is set to 100. A momentum of 0.9 and a weight decay of 0.01 are also set for model training. We train our full FFCSN model on two Tesla K40 GPUs, with the batch size 12. Our implementation is developed within the PyTorch framework.

Model	ACC	AUC
Face	84.33	84.11
Motion	86.00	88.63
Face+Motion	88.21	90.57
Face+Motion+CL	89.16	91.89
Face+Motion+CL+ML	92.33	95.83
Face+Motion+CL+ML+AL	**93.16**	**96.71**

Table 1. Ablation study results (%) for our full FFCSN model.

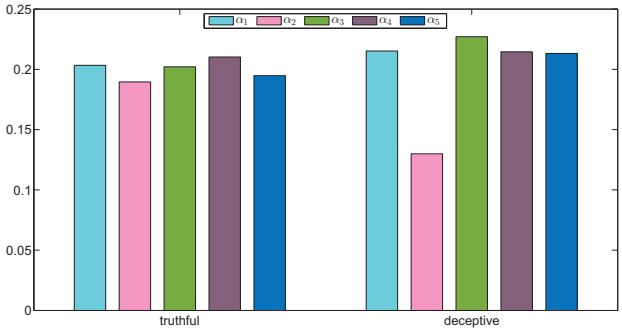

Figure 5. Illustration of the two mean correlation distributions (i.e. $\alpha = [\alpha_1, ..., \alpha_5]$) obtained with our cross-stream network averaged over the truthful and deceptive test samples, respectively.

4.1.2 Ablation Study Results

To show the contribution of each main module of our full FFCSN model, we make comparison to its five simplified versions: (1) Face - Only the face branch of our cross-stream base network used for ADD; (2) Motion - Only the motion branch of our cross-stream base network used for ADD; (3) Face+Motion – our cross-stream base network including the face and motion branches (but without cross-stream correlation learning); (4) Face+Motion+CL – our cross-stream base network with cross-stream correlation learning (CL); (5) Face+Motion+CL+ML – our cross-stream base network further boosted with meta learning (ML). Our full model including adversarial learning (AL) is denoted as Face+Motion+CL+ML+AL.

The ablation study results are presented in Table 1. It can be seen that: (1) The performance continuously increases when more modules are used for ADD, showing the contribution of each module. (2) The improvements achieved by Face+Motion over Face/Motion show that both face expression and body motion are important cues for ADD. (3)

Model	ACC	AUC
[36] (visual+verbal)	75.20	–
[37] (visual+verbal)	77.11	–
[17] (visual+acoustic+verbal)	78.95	–
[9] (visual+acoustic+verbal)	96.42	–
[50] (visual+acoustic+verbal)	–	92.21
[20] (visual+acoustic+verbal)	96.14	97.99
Ours (visual)	93.16	96.71
Ours (visual+acoustic+verbal)	**97.00**	**99.78**

Table 2. Comparative results (%) for video-based ADD. Note that extra *human annotated* micro-expressions are used in [50, 20].

Both ML and AL clearly lead to performance improvements, which provides evidence that they have a good ability of alleviating the training data scarcity. (4) The effectiveness of cross-stream correlation learning is validated by the comparison Face+Motion+CL vs. Face+Motion. This is further supported by Figure 5, where our cross-stream correlation learning is found to learn quite different correlation distributions for the truthful/deceptive classes. That is, the learned correlations indeed improve the discriminativeness of deep visual features for deception detection.

4.1.3 Comparative Results

We further make comparison to the state-of-the-art alternatives [17, 9, 50, 20]. Since all of these methods are multimodal, we also include the acoustic and verbal modalities:
Acoustic Feature Learning. We extract the spectrum map from each wave audio of 44,100 Hz sampling rate, and convert each spectrum map into images of fixed size using a sliding window with the window size 300. By taking only the last 300 dimensions along the spectrum height, we obtain a set of samples of the size 300*300. These samples are finally used to train ResNet50. For robust training, ML and AL are similarly exploited for acoustic feature learning.
Verbal Feature Learning. We segment the transcript of each video to words, and then employ the word2vec technique [10] to convert each word into a 300-dimensional feature vector. The feature vectors of all words are averaged as the verbal feature vector of a video. The average vector is fed into three layers of fully connected layers (of the size 300*128, 128*64, and 64*32), resulting in a final vector of 32 dimensions. For robust verbal feature learning, ML and AL are also used like visual feature learning.

The comparative results on the real-life benchmark dataset [36] are given in Table 2. We observe that: (1) Our robust deep feature learning approach clearly performs the best under the multimodal setting, validating the effectiveness of exploiting ML and AL for addressing the training data scarcity issue associated with real-life ADD. (2) When only the visual modality is concerned, our robust deep feature learning approach even outperforms the state-of-the-art

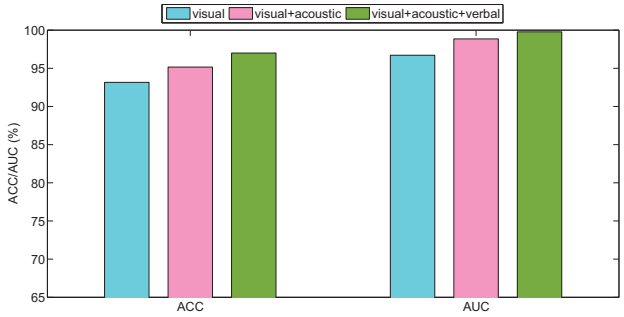

Figure 6. Comparative results obtained by multi-modality fusion.

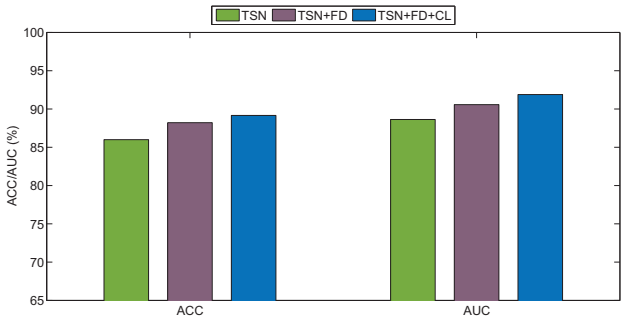

Figure 7. Comparison to temporal segment network (TSN).

multimodal deception detection method [50]. (3) Our multimodal approach achieves performance improvements over the latest deep learning methods [9, 20], due to the extra use of ML and AL in our approach. In addition, we also provide the comparative results of modality fusion for our approach in Figure 6. As expected, our approach is shown to obtain more significant improvements when more modalities are used for deception detection in videos.

4.1.4 Further Evaluations

Comparison to Temporal Segment Network. Different from the state-of-the-art temporal segment network (TSN) [47], our FFCSN model has two novel components: face detection and correlation learning. To show the contribution of these two components, we obtain two variants of our FFCSN model by adding face detection (FD) and correlation learning (CL) into TSN: (1) TSN+FD: face detection is added to the spatial stream of TSN; (2) TSN+FD+CL: cross-stream correlation learning is further used to boost TSN+FD. The comparative results in Figure 7 clearly show that both components are effective for ADD.
Model Selection for Meta Learning. As illustrated in Figure 4, the number of sample pairs in each sampled task in the meta-learning pipeline is empirically set to 5. To evaluate the impact of the task size on the model performance, Figure 8 compares different task sizes. It can be clearly seen that our model approaches the peak at 5, but it is in general insensitive to the task size selection.

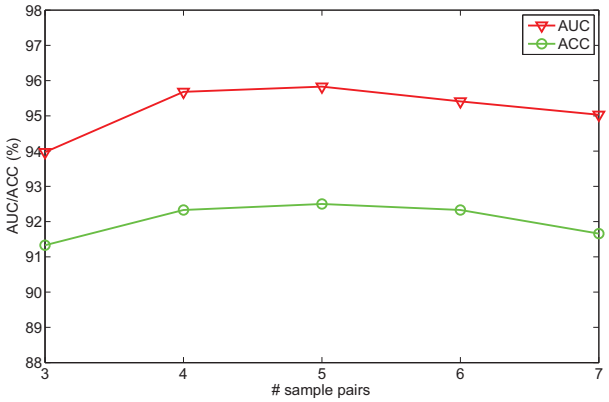

Figure 8. Illustration of the effect of the number of sample pairs used for pairwise comparison on the performance of meta learning.

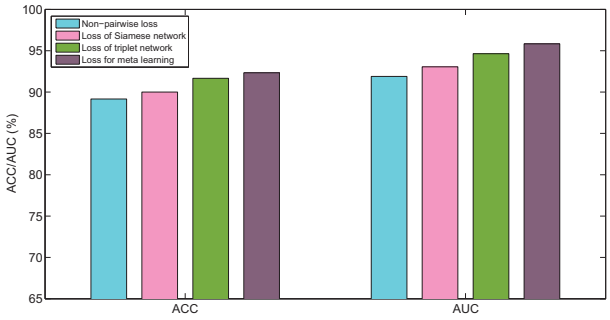

Figure 9. Comparative results obtained by employing different losses for pairwise comparison.

Alternative Losses for Pairwise Comparison. In this paper, our loss defined in Eq. (3) is used for pairwise comparison. To show its effectiveness, we compare it to two typical pairwise losses under the same setting: loss of Siamese network [3, 43], and loss of triplet network [2, 15]. The conventional non-pairwise loss is also included as the baseline. The comparative results in Figure 9 show that: (1) All three pairwise losses clearly lead to better results than the conventional non-pairwise loss, validating the effectiveness of pairwise comparison for deception detection. (2) The loss defined in Eq. (3) performs the best among the three pairwise losses, i.e., the meta learning module is more capable of modelling the complicated relationships among video samples than the Siamese network and triplet network.

4.2. Video-Based Emotion Recognition

4.2.1 Dataset and Setting

The YouTube-8 dataset [18] is used for performance evaluation. This dataset consists of 1,101 videos (downloaded from YouTube) annotated with 8 basic emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. We randomly generate 10 train/test splits, each using 2/3 of the dataset for training and 1/3 for testing. The averaged recognition accuracy (ACC) over 10 random train/test splits is used as the evaluation metric. Our FFCSN model is trained exactly the same as in Section 4.1, and only visual features are extracted from raw videos for emotion recognition.

Model	multimodal	ACC
[18]	visual+acoustic+attribute	46.1
[34]	visual+acoustic+attribute	51.1
[57]	visual+attribute	52.5
[54]	visual+acoustic	52.6
[53]	visual+acoustic	52.6
Ours	visual	**57.8**

Table 3. Comparative results (%) of video-based emotion recognition on the YouTube-8 dataset.

4.2.2 Comparative Results

We compare our FFCSN model to the state-of-the-art alternatives [18, 34, 53, 57, 54]. The comparative results are presented in Table 3. We have the following observations: (1) Our FFCSN model achieves significant improvements over the state-of-the-art models, validating the effectiveness of our face-focused cross-stream network for emotion recognition from user-generated videos. Note that the biggest challenge of this emotion recognition task lies in the complicated and unstructured nature of user-generated videos and the sparsity of video frames that express the emotion content. Our FFCSN model is clearly effective in overcoming this challenge. (2) The improvements obtained by our FFCSN model are really impressive, given that only visual features are extracted by our model, whilst at least two modalities are used by all other models.

5. Conclusion

In this paper, we have investigated the challenging problem of deception detection from real-life videos. For joint deep feature learning from facial expressions and body motions, we have proposed a novel face-focused cross-stream network (FFCSN). Importantly, different from existing two-stream networks, our FFCSN model enables us to cope with the temporal inconsistency between facial expressions and body motions for ADD. Moreover, we have also developed a new training approach for our FFCSN model by inducing meta learning and adversarial learning into the training process of our base model. As a result, our FFCSN model can be trained effectively even with only a handful of training samples. Extensive experiments show that the proposed FFCSN model achieves state-of-the-art results.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (61573363 and 61832017), and the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (15XNLQ01).

References

[1] Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Trans. Information Forensics and Security*, 12(5):1042–1055, 2017. 2, 3

[2] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, pages 403–412, 2017. 8

[3] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, pages 539–546, 2005. 8

[4] Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. Cues to deception. *Psychological Bulletin*, 129(1):74–118, 2003. 1

[5] Maarten Derksen. Control and resistance in the psychology of lying. *Theory & Psychology*, 22(2):196–212, 2012. 1

[6] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941, 2016. 1, 2, 3

[7] Tommaso Fornaciari and Massimo Poesio. Automatic deception detection in Italian court cases. *Artificial Intelligence and Law*, 21(3):303–340, 2013. 2

[8] Matthias Gamer. Mind reading using neuroimaging: Is this the future of deception detection? *European Psychologist*, 19(3):172, 2014. 1

[9] Mandar Gogate, Ahsan Adeel, and Amir Hussain. Deep learning driven multimodal fusion for automated deception detection. In *IEEE Symposium Series on Computational Intelligence*, pages 1–6, 2017. 2, 6, 7

[10] Yoav Goldberg and Omer Levy. word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014. 7

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 2, 3

[12] Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Frank Enos, Julia Hirschberg, and Sachin Kajarekar. Combining prosodic lexical and cepstral systems for deceptive speech detection. In *ICASSP*, 2006. 1, 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 4

[14] Julia Hirschberg, Stefan Benus, Jason M Brenier, et al. Distinguishing deceptive from non-deceptive speech. In *European Conference on Speech Communication and Technology*, pages 1833–1836, 2005. 1, 2

[15] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92, 2015. 8

[16] David M Howard and Christin Kirchhübel. Acoustic correlates of deceptive speech–an exploratory study. In *International Conference on Engineering Psychology and Cognitive Ergonomics*, pages 28–37, 2011. 1, 2

[17] Mimansa Jaiswal, Sairam Tabibu, and Rajiv Bajpai. The truth and nothing but the truth: Multimodal analysis for deception detection. In *ICDM Workshops*, pages 938–943, 2016. 2, 3, 6, 7

[18] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. Predicting emotions in user-generated videos. In *AAAI*, volume 14, pages 73–79, 2014. 1, 2, 3, 8

[19] F Andrew Kozel, Kevin A Johnson, Qiwen Mu, Emily L Grenesko, Steven J Laken, and Mark S George. Detecting deception using functional magnetic resonance imaging. *Biological Psychiatry*, 58(8):605–613, 2005. 1

[20] Gangeshwar Krishnamurthy, Navonil Majumder, Soujanya Poria, and Erik Cambria. A deep learning approach for multimodal deception detection. *arXiv preprint arXiv:1803.00344*, 2018. 2, 6, 7

[21] Sarah Ita Levitan, Guozhen An, Min Ma, Rivka Levitan, Andrew Rosenberg, and Julia Hirschberg. Combining acoustic-prosodic, lexical, and phonotactic features for automatic deception detection. In *INTERSPEECH*, pages 2006–2010, 2016. 1, 2

[22] Sarah I Levitan, Guzhen An, Mandi Wang, Gideon Mendels, Julia Hirschberg, Michelle Levine, and Andrew Rosenberg. Cross-cultural production and detection of deception from speech. In *ACM Workshop on Multimodal Deception Detection*, pages 1–8, 2015. 1, 2

[23] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017. 2, 3

[24] Guangzhen Liu, Jun Hu, An Zhao, Mingyu Ding, Yuqi Huo, and Zhiwu Lu. Insightgan: Semi-supervised feature learning with generative adversarial network for drug abuse detection. In *International Conference on Neural Information Processing*, pages 411–422, 2018. 2, 3

[25] Zhiwu Lu, Jiechao Guan, Aoxue Li, Tao Xiang, An Zhao, and Ji-Rong Wen. Zero and few shot learning with semantic feature synthesis and competitive learning. *arXiv preprint arXiv:1810.08332*, 2018. 2, 3, 5

[26] Zhiwu Lu and Yuxin Peng. Latent semantic learning by efficient sparse coding with hypergraph regularization. In *AAAI*, pages 411–416, 2011. 1

[27] Zhiwu Lu and Yuxin Peng. Latent semantic learning with structured sparse representation for human action recognition. *Pattern Recognition*, 46(7):1799–1809, 2013. 1

[28] Zhiwu Lu, Yuxin Peng, and Horace HS Ip. Spectral learning of latent semantics for action recognition. In *ICCV*, pages 1503–1510, 2011. 1

[29] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, pages 94–101, 2010. 6

[30] Daniel McDuff and Mohammad Soleymani. Large-scale affective content analysis: Combining media content features and facial reactions. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 339–345, 2017. 1

[31] Nicholas Michael, Mark Dilsizian, Dimitris Metaxas, and Judee K Burgoon. Motion profiles for deception detection using visual cues. In *ECCV*, pages 462–475, 2010. 1, 3

[32] Takahiro Ogawa, Yoshiaki Yamaguchi, Satoshi Asamizu, and Miki Haseyama. Human-centered video feature selection via mRMR-SCMMCCA for preference extraction. *IEICE Trans. Information and Systems*, 100(2):409–412, 2017. 1

[33] Michel Owayjan, Ahmad Kashour, Nancy Al Haddad, Mohamad Fadel, and Ghinwa Al Souki. The design and development of a lie detection system using facial micro-expressions. In *International Conference on Advances in Computational Tools for Engineering Applications*, pages 33–38, 2012. 1, 3

[34] Lei Pang, Shiai Zhu, and Chong-Wah Ngo. Deep multimodal learning for affective analysis and retrieval. *IEEE Trans. Multimedia*, 17(11):2008–2020, 2015. 3, 8

[35] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream R-CNN for action detection. In *ECCV*, pages 744–759, 2016. 1, 2, 3

[36] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Deception detection using real-life trial data. In *International Conference on Multimodal Interaction*, pages 59–66, 2015. 1, 2, 3, 6, 7

[37] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, CJ Linton, and Mihai Burzo. Verbal and non-verbal clues for real-life deception detection. In *EMNLP*, pages 2336–2346, 2015. 1, 2, 3, 6, 7

[38] 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*, pages 91–99, 2015. 3

[39] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, pages 4967–4976, 2017. 2, 3, 5

[40] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014. 1, 2, 3

[41] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3, 6

[42] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, pages 1199–1208, 2018. 5

[43] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated Siamese convolutional neural network architecture for human re-identification. In *ECCV*, pages 791–808, 2016. 8

[44] Aldert Vrij. *Detecting Lies and Deceit: The Psychology of Lying and Implications for Professional Practice*. Wiley Series on the Psychology of Crime, Policing and Law. Wiley, Hoboken, NJ, USA, 2001. 1

[45] Aldert Vrij, Ronald Fisher, Samantha Mann, and Sharon Leal. Detecting deception by manipulating cognitive load. *Trends in cognitive sciences*, 10(4):141–142, 2006. 1

[46] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558, 2013. 3

[47] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016. 1, 2, 3, 4, 6, 7

[48] Shangfei Wang and Qiang Ji. Video affective content analysis: a survey of state of the art methods. *IEEE Trans. Affective Computing*, 6(4):410–430, 2015. 1

[49] Xiu-Shen Wei, Chen-Lin Zhang, Hao Zhang, and Jianxin Wu. Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Trans. Affective Computing*, 9(3):303–315, 2018. 1

[50] Zhe Wu, Bharat Singh, Larry S. Davis, and V. S. Subrahmanian. Deception detection in videos. In *AAAI*, pages 1695–1702, 2018. 2, 3, 6, 7

[51] Fan Xia, Hong Wang, and Junxian Huang. Deception detection via blob motion pattern analysis. In *International Conference on Affective Computing and Intelligent Interaction*, pages 727–728, 2007. 1, 3

[52] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. In *ECCV*, pages 811–826, 2018. 2, 3, 5

[53] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. Video emotion recognition with transferred deep feature encodings. In *ICMR*, pages 15–22, 2016. 3, 8

[54] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE Trans. Affective Computing*, 9(2):255–270, 2018. 1, 2, 3, 8

[55] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. WIDER FACE: A face detection benchmark. In *CVPR*, pages 5525–5533, 2016. 6

[56] Chen-Lin Zhang, Hao Zhang, Xiu-Shen Wei, and Jianxin Wu. Deep bimodal regression for apparent personality analysis. In *ECCV*, pages 311–324, 2016. 1

[57] Haimin Zhang and Min Xu. Recognition of emotions in user-generated videos with kernelized features. *IEEE Trans. Multimedia*, 20(10):2824–2835, 2018. 3, 8

[58] Zhi Zhang, Vartika Singh, Thomas E Slowe, Sergey Tulyakov, and Venugopal Govindaraju. Real-time automatic deceit detection from involuntary facial expressions. In *CVPR*, pages 1–6, 2007. 1, 3

[59] Qiong Zhao, Zhiwu Lu, and Horace HS Ip. Action recognition based on learnt motion semantic vocabulary. In *Pacific-Rim Conference on Multimedia*, pages 193–202, 2010. 1

[60] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, pages 803–818, 2018. 3

[61] Liuyang Zhou, Zhiwu Lu, Howard Leung, and Lifeng Shang. Spatial temporal pyramid matching using temporal sparse representation for human motion retrieval. *The Visual Computer*, 30(6-8):845–854, 2014. 1

[62] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. ECO: Efficient convolutional network for online video understanding. In *ECCV*, pages 695–712, 2018. 3