

# Vid2Int: Detecting Implicit Intention from Long Dialog Videos

Xiaoli Xu<sup>1\*</sup>Yao Lu<sup>1\*</sup>Zhiwu Lu<sup>1,2†</sup>Tao Xiang<sup>3</sup><sup>1</sup>Gaoling School of Artificial Intelligence, Renmin University of China<sup>2</sup>Beijing Key Laboratory of Big Data Management and Analysis Methods<sup>3</sup>University of Surrey, United Kingdom

lmnhsp@gmail.com

luyao777@ruc.edu.cn

luzhiwu@ruc.edu.cn

## Abstract

Detecting subtle intention such as deception and subtext of a person in a long dialog video, or implicit intention detection (IID), is a challenging problem. The transcript (textual cues) often reveals little, so audio-visual cues including voice tone as well as facial and body behaviour are the main focuses for automated IID. Contextual cues are also crucial, since a person's implicit intentions are often correlated and context-dependent when the person moves from one question-answer pair to the next. However, no such dataset exists which contains fine-grained question-answer pair (video segment) level annotation. The first contribution of this work is thus a new benchmark dataset, called Vid2Int-Deception to fill this gap. A novel multi-grain representation model is also proposed to capture the subtle movement changes of eyes, face, and body (relevant for inferring intention) from a long dialog video. Moreover, to model the temporal correlation between the implicit intentions across video segments, we propose a Video-to-Intention network (Vid2Int) based on attentive recurrent neural network (RNN). Extensive experiments show that our model achieves state-of-the-art results.

## 1. Introduction

Implicit intention detection (IID) from a long dialog session consisting of multiple question-answer pairs (see Fig. 1) is a challenging problem that has attracted multiple research disciplines. Specifically, it has been studied by researchers from psychology [7, 2], linguistics [3], sociology [9], and recently computer vision [47, 12]. Note that solving IID automatically without replying on a human expert has many potential real-world applications such as court trial [32] and financial loan risk assessment [49]. For instance, a deception detection model [47, 12] can help a court judge

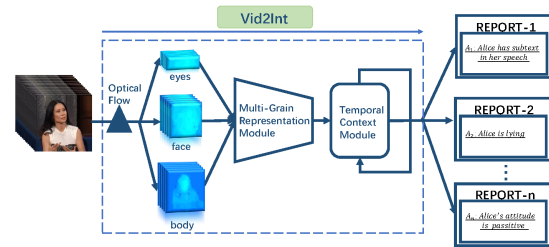


Figure 1. Schematic of implicit intention detection (IID) from long dialog videos. It is essentially a form of Seq2Seq (i.e. Vid2Int) that translates a sequence of video segments containing question-answer pairs to a sequence of intention descriptions.

to make more informed decision during a court trial, and a bank loan officer to detect the occurrence of fraud.

IID covers a number of research directions including deception detection [19, 16] and subtext detection [20]. Deception is an intentional attempt to mislead others [10], and subtext is almost constantly presented in a discourse [4]. From a linguistic point of view, language is rather indirect: people often do not say what they want to say, and tend to express it indirectly (either subconsciously or deliberately). Although the implicit intention is important for daily communication, detecting it is extremely hard. We typically need specific linguistic forms designed by experts to examine it. Hiring a human expert is often expensive and unscalable if IID is to be employed widely. Therefore, there is an urgent need for automated IID methods.

Existing automated IID methods are divided into three groups. The first group exploits verbal cues [21, 23, 27, 10, 20] for intention detection. Compared with the statements of events that a person has experienced, the fake statements are often not vivid enough and the details are scarce. Different from fake statements, the subtext in the discourse is even harder to identify. The second group is based on the micro-expression recognition technique [13, 14, 31, 34], i.e., it relies on analysing visual non-verbal clues. The third group employs physiological measurements (e.g., heart rate) and analysis [35, 36, 43]. These methods thus require profes-

\*Equal Contribution

†Corresponding Author

sional software/hardware and subject cooperation. Among the three groups, the second group, i.e., the visual cue based methods seem most promising because these methods do not need either subject cooperation or the unreliable verbal transcript analysis. Very impressive detection accuracies have been reported [47, 12]. This line of approach is thus the focus of this study.

Despite the promise shown by existing visual cue based IID methods [47, 12], they are severely limited by the lack of suitable datasets. More specifically, a long dialog is composed of multiple question-answer pairs; the dialog video thus can be temporally segmented accordingly. The intention exhibited in each pair/segment typically varies – the subject can tell the truth in some segments whilst lying in others. Importantly, the intention is typically correlated across different segments: a subject often tells more lies to cover up an earlier lie. This suggests that: (1) IID should be performed at a fine-grained video segment level rather than holistically at the video level; and (2) the intention in a specific segment should be inferred with consideration on the temporal context provided by other segments of the video. However, existing datasets such as [32] only contain video-level labels, resulting in existing methods [47, 12] completely ignoring fine-grained IID as well as the temporal correlation between the intentions over different segments of the video. To overcome this limitation and kick-start the research into fine-grained IID, we contribute a new benchmark dataset called Vid2Int-Deception. This dataset contains question-answer pair level annotation and thus enables fine-grained context-aware IID.

To solve the fine-grained IID problem, we first propose a multi-grain representation model. The model is designed to capture the subtle movement changes of eyes, face and body from a long dialog video. After segmenting the long dialog video into segments, each of which contains a question-answer pair, our model takes optical flow as input, because it has proven to be useful for movement change detection [40, 50, 39, 45, 30, 11, 24]. We note that for modeling subtle human facial and body behaviour relevant for intention detection, optical flow is much more effective than the raw RGB frames (see suppl. material). Moreover, to capture the temporal context for inferring intention, we treat a sequence of video segments as input and our goal is to translate the sequence into a sequence of intention labels or descriptions. This is analogous to Seq2Seq [42, 17, 44]. More specifically, the multi-grain representation model is integrated into a Video-to-Intention (Vid2Int) network model for fine-grained IID based on attentive recurrent neural network (RNN) [22, 38]. The overview of our Vid2Int model is shown in Fig. 1<sup>1</sup>.

<sup>1</sup>Note that although our Vid2Int model is a generic IID model, in this study we only evaluate it on the deception detection sub-problem, and leave the subtext detection sub-problem to future work.

Our contributions are: (1) For the first time, we tackle the problem of fine-grained IID from a long dialog video containing multiple question-answer pairs. (2) To facilitate the research in this new problem, we contribute a benchmark dataset called Vid2Int-Deception, which consists of 292 long dialog videos (each with multiple question-answer pairs and segment-level annotation on deception). (3) Apart from the main contributions of defining a new problem and providing a first benchmark dataset to enable the research on the problem, we also formulate an effective Vid2Int model with a number new components introduced. The efficacy of our model is validated through extensive experiments on both Vid2Int-Deception and an existing court video dataset [32]. The results not only show that our model is clearly superior to existing alternatives, but also provide important insight on how to solve the fine-grained IID problem. The code and dataset will be released soon.

## 2. Related Work

**Implicit Intention Detection.** We have discussed the three groups of IID methods. More recent methods mostly focus on the non-intrusive non-contact approaches which rely on verbal, acoustic, visual, and thermal techniques. In particular, visual methods for deception detection typically leverage facial micro-expressions [13, 14, 31, 34] and action recognition [12]. Moreover, some multimodal methods [25, 1, 18, 47, 26] have also been developed for detecting implicit intention from videos. However, these approaches can only handle each video as a whole, and cannot attach class labels (e.g. deception/truth labels) to multiple video segments within the video. This is partially due to the lack of suitable datasets. By contributing the first fine-grained IID dataset, we are able to tackle the segment-level intention detection problem for the first time. **RGB Frames vs. Optical Flow as Input.** Both RGB frames and optical flow have been used as inputs for a video analysis model. Two-stream networks have been popular, which exploit both for video analysis [40, 45, 51, 30]. More recent models, particularly those designed for action recognition, use only RGB frames [6, 28]. However, for IID from long dialog videos, optical flow is found to be far more useful than RGB frames. It is also noted that adding RGB frames to optical flow only yields slight improvements (see the suppl. material). Therefore, in this work, we leverage only optical flow for IID.

**Recurrent Neural Networks.** Recurrent neural networks (RNNs) have often been exploited for sequence modeling. [38] proposed a bidirectional RNN trained in both positive and negative time directions. A modified RNN that utilizes the Long Short-Term Memory (LSTM) [22] structure generally outperforms the vanilla RNN. A number of attentive RNN models [8, 48, 46] have also been proposed for better sequence modeling. In this work, based on attentive RNN, we propose a Vid2Int model for comprehensive understand-

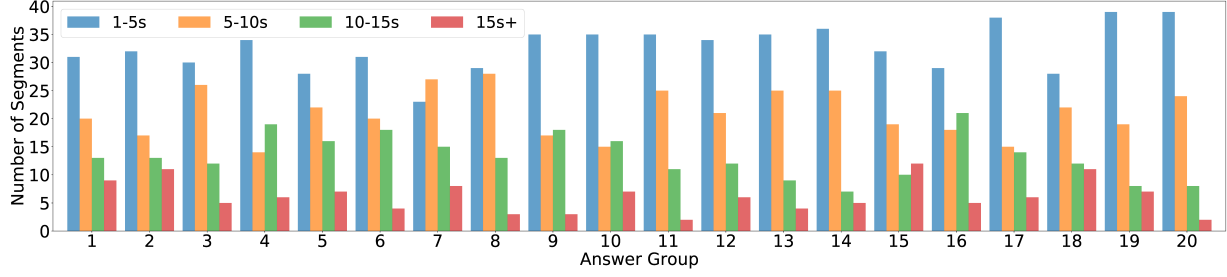


Figure 2. Overview of our Vid2Int-Deception dataset. All video segments are organized into 20 answer groups: answer group  $i$  consists of all answers to question  $i$  ( $i = 1, \dots, 20$ ).

ing of the subject’s implicit intention. However, our Vid2Int is different from existing RNNs in that: (1) It shares the same LSTM for two IID tasks (i.e. IID over video segments and IID over the entire dialog video); (2) Both hidden states and the outputs of IID over video segments are used to generate attention weights (see Fig. 7).

### 3. The Vid2Int-Deception Dataset

The first challenge in IID from long videos is that there exist no suitable datasets for question-answer or video segment level intention detection. The well-known court video dataset [32] used for deception detection has only video level annotation only. In addition, there are a number of other shortcomings: (1) The subjects in this dataset have various angles and significant changes in posture. The videos also have strong background noise and low video quantity. (2) Since the subjects often say a lot of words very quickly, these words may be mixed with truthful/deceptive intentions. However, it is difficult to distinguish which of these words are deceptive. Therefore, a new dataset is needed to study the much harder fine-grained IID problem.

To this end, we contribute a new dialog video dataset named Vid2Int-Deception. Concretely, we invite 73 volunteers (both undergraduate and graduate students) for our data collection, 41 of which are male and 32 are female. The place where we collect the dialog videos is a lab room so that the background of each video is stable. The volunteers are required to sit down and cannot stand up during video recording. Only the upper bodies of the volunteers appear in the recorded videos. A Nikon COOLPIX L110 camera is used for video recording. A polygraph device is used as a reference during our data collection, acting as a strong baseline (widely considered to be more trusted than humans). Each video has a frame size of  $1,280 \times 720$  and a frame rate of 30 FPS.

Our Vid2Int-Deception dataset consists of 292 dialog videos, each of which has five question-answer pairs. By regarding each answer as a video segment, we have totally 1,460 video segments. Note that each video segment contains both the questioning and answering periods because the volunteer has subtle movement changes even when the

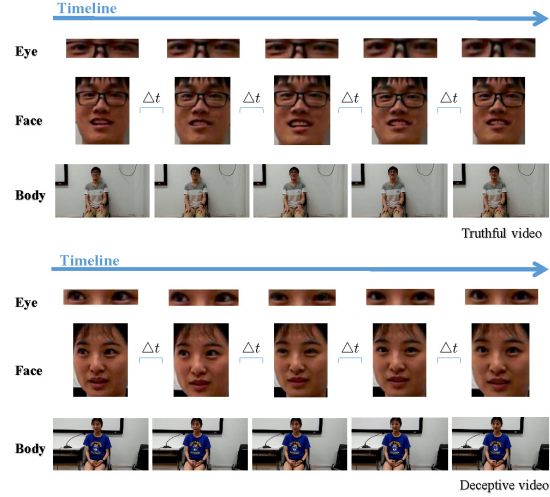


Figure 3. Two dialog video examples. Each example is decomposed into three videos by focusing on different parts of the subject. Notation:  $\Delta t = 5$  frames.

questioning phase just begins. After video recording, we ask the volunteer to give a truthful/deceptive label to each answer according to whether it is consistent with the truth. As a result, our dataset has 148 deceptive videos (at least one answer in each video is deceptive) and 144 truthful videos (all answers in each video are truthful). The distribution of duration (second) of video segments is shown in Fig. 2. All video segments are organized into 20 answer groups: answer group  $i$  consists of all answers to question  $i$  ( $i = 1, \dots, 20$ ). In this work, the candidate question set used for our data collection has 20 questions in total. Moreover, two dialog video examples are shown in Fig. 3, where each example is decomposed into three videos by focusing on different parts of the volunteer.

The correlation matrix between candidate questions (i.e. answer groups) is illustrated in Fig. 4. In this work, based on a form of entropy, we define the correlation matrix  $C = [c(i, j)]_{20 \times 20}$  between answer groups as follows:

$$c(i, j) = 1 + \frac{1}{2} \sum_{s=1}^4 p_s(i, j) \log_2 p_s(i, j), \quad (1)$$

where  $p_s(i, j)$  is the probability of the combination state  $s$

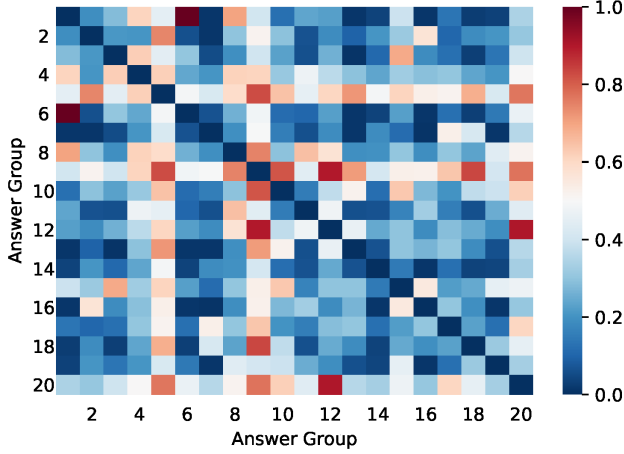


Figure 4. Illustration of the correlation matrix between candidate questions (i.e. answer groups). The correlation used here is defined in Eq. (1).

of answer group pair  $(A_i, A_j)$ . When  $i = j$ , we simply define  $c(i, i) = 0$ . Given that  $A_i$  (or  $A_j$ ) has two states: truthful state  $T$  and deceptive state  $D$ , the combination states of  $(A_i, A_j)$  are listed as follows:  $\{1 : (T, T), 2 : (T, D), 3 : (D, T), 4 : (D, D)\}$ . If there is no correlation between  $A_i$  and  $A_j$ ,  $p_s(i, j) = 1/4$  for the combination state  $s$  ( $s = 1, \dots, 4$ ) and thus  $c(i, j) = 0$ . If there is strong correlation between  $A_i$  and  $A_j$ ,  $p_s(i, j) = 1$  and  $p_{s'}(i, j) = 0$  for other states  $s' \neq s$ , resulting in  $c(i, j) = 1$ . From Fig. 4, we find that there indeed exist a number of strongly-correlated question pairs, indicating that the deception states across different segments are correlated that need to be analysed jointly rather than independently.

The value of our new dataset for real-world application-s lies in several aspects: (1) This dataset can facilitate the research on fine-grained IID in long dialog videos, which is crucial for developing a human-centered AI system. We will soon release it on GitHub. (2) We took great care during data collection to simulate the real-world environment: a ‘natural’ question-answer dialog is conducted with almost unconstrained conditions; a large variety of subjects are recruited; and a polygraph device is used as a reference. The model trained on our dataset can thus be used as a good initialization for real-world applications such as job interview and financial loan risk assessment.

## 4. Methodology

Our Vid2Int model consists of a voice activity detection module (VADM), a multi-grain representation module (MGRM), and a temporal context module (TCM), as shown in Fig. 5. We give the details of each module below.

### 4.1. Voice Activity Detection Module

Voice Activity Detection (VAD) [41] is also known as voice endpoint detection or speech boundary detection, which aims to identify and eliminate long silent periods from the sound signal stream to save the channel resources without degrading the quality of the service. In our Vid2Int model, VAD is used to automatically split a long dialog video into multiple video segments, each of which corresponds to a single answer of the subject. Specifically, we reduce the sampling rate to 8kHz, compute the sub-band energy with Gaussian mixture model, and output the probability of silence and speech of each audio window.

In this work, we first extract the audio of the long dialog video and then split it into  $m$  audio segments by VAD. The long dialog video is also split into  $m$  video segments according to the boundary points of the audio. Therefore, from the long dialog video, we obtain a sequence of video segments and a sequence of audio segments. In particular, we represent  $m$  audio segments using Mel-frequency cepstral coefficients (MFCC) [29]. Formally, the audio’s feature representation is defined as:  $\mathcal{U} = \{\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_m\}$ , where  $\mathcal{U}_i$  is a 32-dimension feature vector obtained by applying a fully-connected layer on the MFCC features of audio segment  $i$  ( $i = 1, \dots, m$ ). This enables us to study whether the audio cues complement to those visual cues.

### 4.2. Multi-Grain Representation Module

After each dialog video is divided by VADM, there are  $m$  video segments, which are denoted as  $\{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_m\}$ , as illustrated in Fig. 5. The obtained video segments are further fed into a face detector and an eye detector to extract the face area and the eye area, respectively. Therefore, each dialog video is decomposed into three videos by focusing on different parts of the subject: the body part  $\{\mathcal{S}_1^b, \mathcal{S}_2^b, \dots, \mathcal{S}_m^b\}$ , the face part  $\{\mathcal{S}_1^f, \mathcal{S}_2^f, \dots, \mathcal{S}_m^f\}$ , and the eye part  $\{\mathcal{S}_1^e, \mathcal{S}_2^e, \dots, \mathcal{S}_m^e\}$ . Since each video segment often has hundreds of frames (with much redundant information), the computational cost is huge when all frames are used. Therefore, we sample each video segment with  $K$  sub-segments at equal intervals for the subsequent optical flow extraction, as in [45].

For optical flow extraction, we operate on a stack of consecutive warped optical flow fields to capture the motion information. In this work, each dialog video are represented with three types of optical flow: body flow  $\mathcal{F}^b = \{\mathcal{F}_1^b, \mathcal{F}_2^b, \dots, \mathcal{F}_m^b\}$ , face flow  $\mathcal{F}^f = \{\mathcal{F}_1^f, \mathcal{F}_2^f, \dots, \mathcal{F}_m^f\}$ , and eye flow  $\mathcal{F}^e = \{\mathcal{F}_1^e, \mathcal{F}_2^e, \dots, \mathcal{F}_m^e\}$ .

Our MGRM for fusing the above three optical flows is shown in Fig. 6, where MobileNetV2 [37] is used as the backbone network. Specifically, the three optical flows are first encoded into three feature vectors of different grains (i.e. body, face, eye) with three backbone networks. Further, the coarse-grained features are combined with more

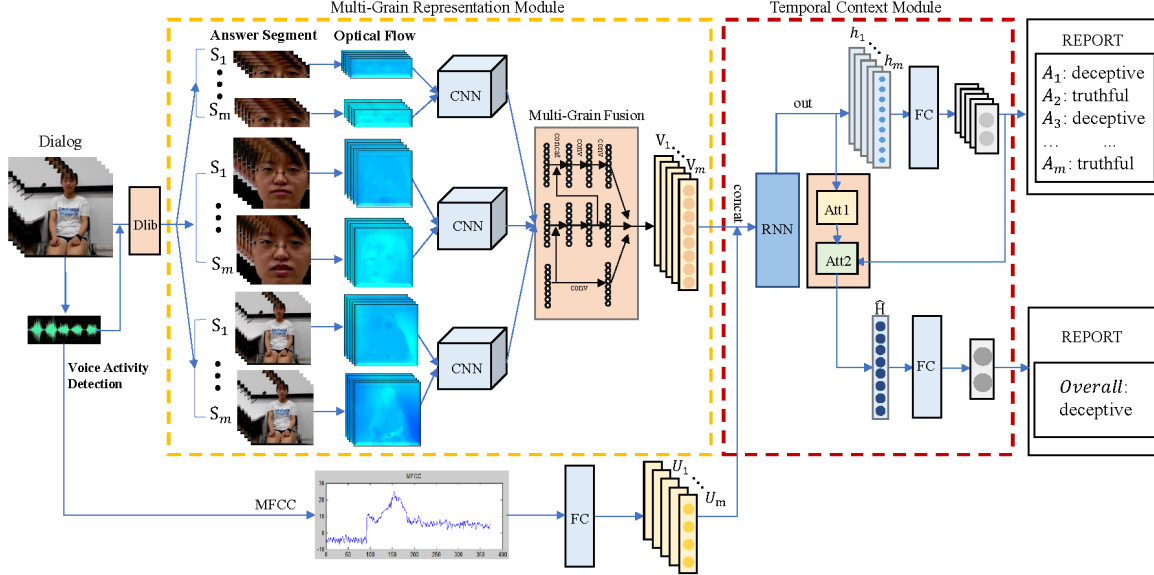


Figure 5. Overview of our full Vid2Int model. Voice activity detection is used to split the long dialog video into multiple video segments.

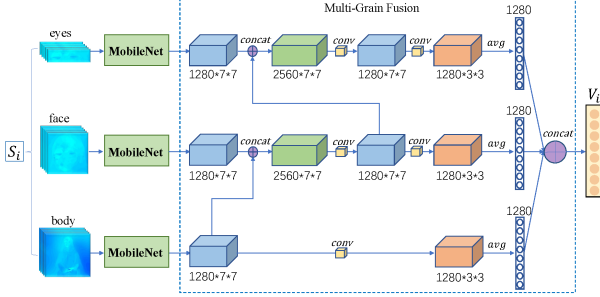


Figure 6. Illustration of the network architecture of our MGRM.

fine-grained features before pooling layers. In this work, we pay more attention to the face flow and eye flow, because these two local flows make it easier to detect the subtle movement changes of the subject.

Formally, after the three flows ( $\mathcal{F}^b, \mathcal{F}^f, \mathcal{F}^e$ ) are fed into the backbone networks, we extract  $7 \times 7$  feature maps ( $\mathbf{f}_{7 \times 7}^b(i), \mathbf{f}_{7 \times 7}^f(i), \mathbf{f}_{7 \times 7}^e(i)$ ) of 1,280 channels for video segment  $i$  ( $i = 1, \dots, m$ ). Importantly, different from existing multi-stream fusion networks [40, 15], our model employs multi-grain fusion rather than simple concatenation. First, we concatenate the feature map  $\mathbf{f}_{7 \times 7}^b(i)$  for body flow to the feature map  $\mathbf{f}_{7 \times 7}^f(i)$  for face flow, and reduce the channel from 2,560 ( $2 \times 1,280$ ) to 1,280 using  $1 \times 1$  convolution kernel. We thus obtain a new feature map  $\hat{\mathbf{f}}_{7 \times 7}^f(i)$  for face flow. Second, we concatenate the feature map  $\hat{\mathbf{f}}_{7 \times 7}^f(i)$  for face flow to the feature map  $\mathbf{f}_{7 \times 7}^e(i)$  for eye flow, and adopt the same channel reduction operation to obtain a new feature map  $\hat{\mathbf{f}}_{7 \times 7}^e(i)$  for eye flow. Third, we define a small down-sampling block which has two  $3 \times 3$  convolution layers followed by batch normalization and RE-LU activation. For the three flows,  $7 \times 7$  feature maps ( $\mathbf{f}_{7 \times 7}^b(i), \hat{\mathbf{f}}_{7 \times 7}^f(i), \hat{\mathbf{f}}_{7 \times 7}^e(i)$ ) are transformed into  $3 \times 3$  feature

maps ( $\mathbf{f}_{3 \times 3}^b(i), \mathbf{f}_{3 \times 3}^f(i), \mathbf{f}_{3 \times 3}^e(i)$ ) using the down-sampling block. We impose average pooling on the  $3 \times 3$  feature maps to output 1280-dimensional feature vectors ( $\mathbf{f}_i^b, \mathbf{f}_i^f, \mathbf{f}_i^e$ ). We thus represent video segment  $i$  as:

$$\mathcal{V}_i = \Phi(\mathbf{f}_i^b, \mathbf{f}_i^f, \mathbf{f}_i^e), \quad (2)$$

where  $\Phi$  denotes the concatenation operation and  $\mathcal{V}_i$  is a 3,840-dimensional feature vector. The output of our MGRM is thus represented as  $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_m\}$ .

By concatenating the audio's feature representation  $\mathcal{U} = \{\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_m\}$  with  $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_m\}$ , we represent each dialog video as  $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_m\}$ , as shown in Fig. 5. Formally,  $\mathcal{D}_i$  ( $i = 1, \dots, m$ ) is formulated as:

$$\mathcal{D}_i = \Phi(\mathcal{V}_i, \mathcal{U}_i), \quad (3)$$

where  $\Phi$  is the concat operation and  $\mathcal{D}_i$  is a 3,872-dimensional feature vector.

### 4.3. Temporal Context Module

Temporal context exists in a sequence of video segments. In our dataset, the volunteers are asked successively with a number of related questions, and the deception states are thus heavily correlated (see Fig. 4). Consequently, we need to model the temporal context across a sequence of answer segments to obtain better IID results. In this work, a Temporal Context Module (TCM) is designed to learn the temporal context from a long dialog video (see Fig. 7).

We firstly feed the fused representation  $\mathcal{D}$  into an LSTM module, since it has the capacity of learning the temporal context information. The total hidden state of LSTM can be defined with  $H = [h_1, h_2, \dots, h_m]$ , where  $h_i$  is the hidden state of each hidden layer in LSTM. We can use  $h_i$  to predict the subject's implicit intention within the video segment  $\mathcal{D}_i$ .



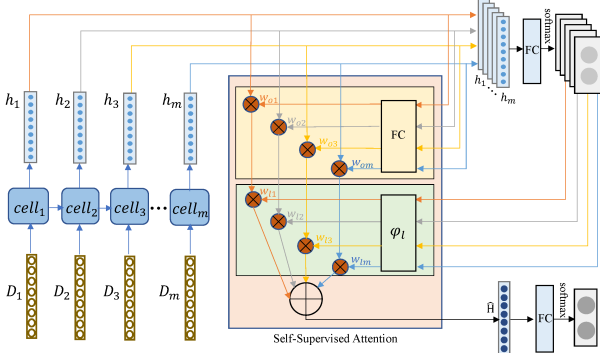


Figure 7. Illustration of the network architecture of our TCM.

Let each cell state be defined as  $C_i$  ( $i = 1, \dots, m$ ). The hidden state of each video segment is given by:

$$h_i = \text{LSTM}(h_{i-1}, \mathcal{D}_i, C_i), \quad (4)$$

where  $h_i$  is a 128-dimensional feature vector in our setting.

Note that each hidden state has a different influence on IID for the whole dialog video and the prediction of each video segment can also exert their influence. For example, given a deceptive dialog video, the answer segments with truthful predictions should have less importance for IID on the whole video. Therefore, we propose an attentive-LSTM module which consists of two forms of self-supervised attention, as illustrated in Fig. 7. Formally, we define the attentive representation  $\text{att}(h_i)$  of  $h_i$  as follows:

$$\text{att}(h_i) = w_o(h_i) \cdot w_l(h_i), \quad (5)$$

where  $w_o(h_i)$  is the first attention and  $w_l(h_i)$  is the second attention. Specifically,  $w_o(h_i)$  is defined as:

$$w_o(h_i) = FC(h_i), \quad (6)$$

where  $FC$  denotes the fully-connected layer used for attention. Moreover,  $w_l(h_i)$  is defined with the predicted labels:

$$w_l(h_i) = \varphi_l(p_i) = p_i[1 - \alpha \cdot p_i(1 - p_i)], \quad (7)$$

where  $p_i$  is the probability of  $h_i$  being classified as the deceptive class, and  $\alpha$  denotes a hyperparameter for adjusting  $w_l$  ( $\alpha = 4$  in this work). We have  $p_i = \text{softmax}(FC(h_i))$ , where  $FC$  denotes the fully-connected layer used for label prediction. Note that the term  $[1 - \alpha \cdot p_i(1 - p_i)]$  measures the reliability of the label prediction for  $h_i$ .

Finally, we multiply the total hidden state  $H$  and attentive weights  $\text{att}(H) = [\text{att}(h_1), \text{att}(h_2), \dots, \text{att}(h_m)]^T$  to generate the attentive representation of each dialog video. The feature vector  $\hat{H}$  of the dialog video is defined as:

$$\hat{H} = H \times \text{att}(H). \quad (8)$$

There are two IID tasks concerned in our Vid2Int model: IID for each video segment, and IID for each long dialog

| Method          | Answer-ACC   | Dialog-ACC   |
|-----------------|--------------|--------------|
| Two-Stream [40] | 62.38        | 71.15        |
| ST [50]         | 63.46        | 73.08        |
| ShuttleNet [39] | 64.61        | 73.08        |
| TSN [45]        | 65.38        | 76.92        |
| I3D [5]         | 66.92        | 78.85        |
| STPN [30]       | 67.31        | 78.85        |
| Two-Stream+LSTM | 65.77        | 75.00        |
| ST+LSTM         | 67.69        | 76.92        |
| ShuttleNet+LSTM | 69.23        | 78.85        |
| TSN+LSTM        | 70.38        | 82.69        |
| I3D+LSTM        | 70.77        | 84.62        |
| STPN+LSTM       | 72.69        | 84.62        |
| Vid2Int (ours)  | <b>76.53</b> | <b>92.31</b> |

Table 1. Comparative accuracies (%) on our Vid2Int-Deception dataset.

video. Let  $L_s$  be the loss defined over video segments and  $L_d$  be the loss defined over dialog videos. Our total loss for model training can be defined as follows:

$$L = L_s + \beta L_d, \quad (9)$$

where  $\beta$  denotes a hyperparameter for weighting the two losses. In this work, we empirically set  $\beta = 1$ .

## 5. Experiments

### 5.1. Fine-Grained IID from Dialog Videos

**Dataset and Setting** 1) **Vid2Int-Deception**. The new Vid2Int-Deception is used in this experiment. As mentioned earlier, it consists of 292 long dialog videos. The participant in each video is required to answer five related questions. Each video has a frame size of  $1,280 \times 720$  and a frame rate of 30 FPS.

2) **Data Split**. Our Vid2Int-Deception dataset is split into the training/test set at the ratio 4:1 according to the participants, which ensures that participants in the test set have no overlap with those in the training set. Further, we adopt two approaches to augment the training set: three brightness adjustments and four frame random deletions (at the beginning of each video). After data augmentation, there are 2,803 dialog videos in the training set.

3) **Evaluation Setting**. For performance evaluation, we define two metrics as: (1) Answer-ACC: the classification accuracy computed over visual answers (i.e. video segments) in the test set; (2) Dialog-ACC: the classification accuracy computed at the dialog video level in the test set.

4) **Implementation Details**. Our full model for fine-grained IID is trained end-to-end using back-propagation and adaptive moment estimation. The learning rate is set to 0.005 at first epochs and then reduced by 0.5 on plateau with patience of 30 epochs. The maximum number of total epochs is set to 120. We train our full model on one TITIAN XP, with the batch size 5. Our implementation is developed within the PyTorch framework.

| Method                       | Answer-ACC   | Dialog-ACC   |
|------------------------------|--------------|--------------|
| Face+Softmax                 | 66.00        | 71.15        |
| Face+LSTM                    | 71.92        | 76.92        |
| Concat (Face+Eye)+LSTM       | 72.69        | 80.77        |
| Concat (Face+Eye+Body)+LSTM  | 74.23        | 84.61        |
| MGRM (Face+Eye)+LSTM         | 74.23        | 86.54        |
| MGRM (Face+Eye+Body)+LSTM    | 75.38        | 88.46        |
| MGRM (Face+Eye+Body)+AT-LSTM | <b>76.53</b> | <b>92.31</b> |

Table 2. Ablation study results for our full Vid2Int model.

| Method                       | Answer-ACC   | Dialog-ACC   |
|------------------------------|--------------|--------------|
| MGRM (no cross-grain fusion) | 74.62        | 85.68        |
| MGRM (body-to-face fusion)   | 75.00        | 86.54        |
| MGRM (face-to-eye fusion)    | 75.00        | 88.46        |
| MGRM (full)                  | <b>76.53</b> | <b>92.31</b> |

Table 3. Ablation study results for our MGRM.

**Comparative Results** The comparative results for fine-grained IID on the Vid2Int-Deception dataset are shown in Table 1. For competitors, we focus on those optical-flow based methods because RGB frames are found to be much weaker for fine-grained IID (see more supports in the suppl. material). It can be seen that: (1) Our Vid2Int model significantly outperforms the representative/state-of-the-art alternatives, validating the effectiveness of both multi-grain representation and attentive LSTM for fine-grained IID. (2) Adding LSTM into all compared methods consistently leads to improvements, indicating that the temporal context is indeed crucial for solving the fine-grained IID problem.

**Ablation Study Results** 1) **Ablation Study for Our Full Model.** To show the contribution of each module of our full Vid2Int model, we compare six simplified versions: (1) Face+Softmax – only face flow is used, followed by softmax-based prediction; (2) Face+LSTM – only face flow is used, followed by LSTM-based prediction; (3) Concat (Face+Eye)+LSTM – both face flow and eye flow are used, followed by LSTM-based prediction; (4) Concat (Face+Eye+Body)+LSTM – All three flows are used, followed by LSTM-based prediction; (5) MGRM (Face+Eye)+LSTM – our multi-grain representation module using face flow and eye flow, followed by LSTM-based prediction. (6) MGRM (Face+Eye+Body)+LSTM – our MGRM using all three flows, followed by LSTM-based prediction. Note that our full Vid2Int model including attentive-LSTM is denoted as MGRM (Face+Eye+Body)+AT-LSTM.

The ablative results for our full Vid2Int model are shown in Table 2. We have the following observations: (1) Adding more modules into our model consistently leads to performance improvements, demonstrating the contribution of each module. (2) The margins between MGRM (Face+Eye+Body)+LSTM and Concat (Face+Eye+Body)+LSTM validate the effectiveness of our MGRM (also see Table 3); (3) Our attentive LST-

| Method         | Answer-ACC   | Dialog-ACC   |
|----------------|--------------|--------------|
| LSTM           | 75.38        | 88.46        |
| LSTM+ATT1      | 75.76        | 90.38        |
| LSTM+ATT2      | 76.15        | 90.38        |
| LSTM+ATT1+ATT2 | <b>76.53</b> | <b>92.31</b> |

Table 4. Ablation study results for our TCM.

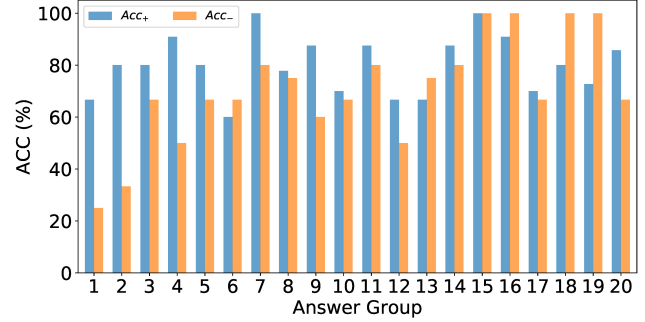


Figure 8. Illustration of the positive and negative accuracies per answer group.

M clearly yields better results, according to the comparison MGRM (Face+Eye+Body)+AT-LSTM vs. MGRM (Face+Eye+Body)+LSTM (also see Table 4).

2) **Ablation Study for Our MGRM.** The ablative results for our MGRM are shown in Table 3. The same TCM is used for all compared methods. It can be seen that: our cross-grain fusion (body-to-face fusion/face-to-eye fusion) clearly outperforms the simple concat fusion without using cross-grain fusion, and combining the two cross-grain fusion methods leads to significant improvements.

3) **Ablation Study for Our TCM.** The ablation results for our TCM are shown in Table 4. The same MGRM is used for all compared methods. We can observe that both forms of self-supervised attention, i.e. ATT1 in Eq. (6) and ATT2 in Eq. (7), benefit the conventional LSTM and combining them yields further performance improvements.

**Further Evaluation** 1) **Performance Analysis over Answer Groups.** The positive and negative accuracies per answer group (i.e.  $Acc_+$  and  $Acc_-$ ) obtained by our Vid2Int model are shown in Fig. 8. In particular, given a question from the set of candidate questions,  $Acc_+$  is the prediction accuracy over the group of answers to this question that have truthful ground-truth labels, and  $Acc_-$  is the prediction accuracy over the group of answers to this question that have deceptive ground-truth labels. We can observe that our model performs the worst over Q1 and Q2 by taking both  $Acc_+$  and  $Acc_-$  on board, and performs the best over Q15 and Q16. The four questions are given below:

**Q1:** What (e.g. character, good looking, height, or family condition) is the most important when you choose a boyfriend/girlfriend? Please also explain why.

**Q2:** Have your parents done any disappointing thing to you? What is the most disappointing?






| Timeline       |   |   |   |   |   |
|----------------|---|---|---|---|---|
| Questions      | Please describe the job of your parents.  | Please describe your height, weight, looks.                                       | Please describe your major and research direction.                                | Do you like your parents' career? Why?  | If your favorite gift was given to others by your mother. What will you do?       |
| Video Segments |  |  |  |  |  |
| Answers        | They are CEO and earn much money.   | I have an ordinary looking with normal weight and medium height.                  | I'm major in economics and learn how to earn more money.                          | Yes, I do like it due to it's high salary.  | I'll buy a new one.   |
| Ground Truth   | Deceptive   | Deceptive   | Deceptive   | Truthful  | Truthful  |
| Prediction     | Truthful  | Deceptive   | Deceptive   | Truthful  | Truthful  |

Figure 9. An example of fine-grained IID with our Vid2Int model.

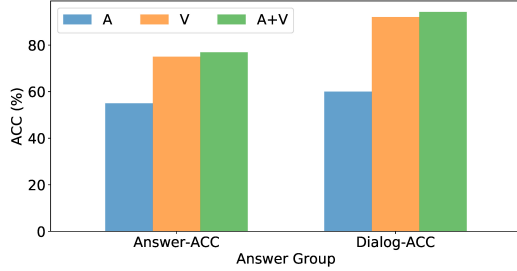


Figure 10. Ablative results obtained for multi-modal fusion.

**Q15:** Please describe your family.

**Q16:** Please describe the jobs of your parents.

The above observation can be explained as: (1) Although the four questions are closely related, **Q1** and **Q2** are clearly more private/personal than **Q15** and **Q16**; (2) Since people are more likely to hide their thoughts when asked more private questions, the IID task becomes harder.

**2) Qualitative Results.** Fig. 9 provides an example of fine-grained IID with our Vid2Int model. It shows that only the first prediction is wrong and the rest labels are predicted correctly by our model. This suggests that LSTM is indeed a good choice for IID from long dialog videos. Moreover, a demo video is also included in the suppl. material to simulate the realistic scenario.

**3) Multi-Modal Fusion.** Fig. 10 shows the ablative results for multi-modal fusion. The notations are: A – acoustic modality; V – visual modality; A+V – both modalities. It can be seen that the audio modality brings further improvements when it is fused with the visual modality for IID.

## 5.2. IID from Court Videos

**Dataset and Setting** We also evaluate our Vid2Int on a court video dataset [32] which has truthful and deceptive videos collected from public court trails. This dataset includes 121 short videos, along with their transcriptions. As in [12], we select a subset of 104 videos from the original dataset. The subset has 50 truthful videos and 54 deceptive videos. Different from our Vid2Int-Deception dataset, each video in the court video dataset has a single label, without the labels of video segments. Therefore, only the MGRM of our Vid2Int model is used on this dataset.

| Method                           | ACC          | AUC          |
|----------------------------------|--------------|--------------|
| [32] (visual+verbal)             | 75.20        | –            |
| [33] (visual+verbal)             | 77.11        | –            |
| [25] (visual+acoustic+verbal)    | 78.95        | –            |
| [18] (visual+acoustic+verbal)    | 96.42        | –            |
| [47] (visual+acoustic+verbal)    | –            | 92.21        |
| [26] (visual+acoustic+verbal)    | 96.14        | 97.99        |
| [12] (visual)                    | 93.16        | 96.71        |
| [12] (visual+acoustic+verbal)    | 97.00        | 99.78        |
| Vid2Int (visual)                 | 94.84        | 97.32        |
| Vid2Int (visual+acoustic+verbal) | <b>97.84</b> | <b>99.83</b> |

Table 5. Comparative results (%) for IID on the real-life court video dataset. Three modalities (i.e. visual, acoustic, and verbal) can be used for IID on this dataset.

Moreover, we perform 10-fold cross validation over subjects (but not over videos) as in [12], which ensures that the subjects in the test set have no overlap with those in the training set. Two evaluation metrics are computed on the test set: (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. These metrics have been widely used in previous works [12, 26].

**Comparative Results** We compare our Vid2Int model to the state-of-the-art models on the public court video dataset. The comparative results are shown in Table 5. We have the following observations: (1) Our Vid2Int model performs the best, validating the effectiveness of our multi-grain representation module for IID. (2) Our model outperforms the state-of-the-art model [12], showing that our multi-grain fusion is more effective to combine multiple visual clues than the cross-stream fusion proposed in [12]. (3) It is obvious from Table 5 that the results obtained on this dataset is getting saturated. This means that a new and more challenging dataset is needed to study the much harder fine-grained IID problem. This is exactly where our first contribution lies.

## 6. Conclusion

In this paper, we investigated the challenging problem of fine-grained IID from long dialog videos. Studying this problem is made possible for a new dataset contributed by this work. We also proposed a multi-grain representation model over optical flow to capture the subtle movement changes of eyes, face, and body for fine-grained IID. Moreover, for modeling temporal context, a Vid2Int model based on attentive RNN was proposed. Extensive experiments show that our model achieves state-of-the-art results.

**Acknowledgement** Zhiwu Lu is partially supported by National Natural Science Foundation of China (61976220 and 61832017), and Beijing Outstanding Young Scientist Program (BJJWZYJH012019100020098).



## References

- [1] Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions on Information Forensics and Security*, 12(5):1042–1055, 2016.
- [2] Icek Ajzen and Nilanjana Dasgupta. Explicit and implicit beliefs, attitudes, and intentions. In *The Sense of Agency*, volume 115. Social Cognition and Social Ne, 2015.
- [3] John Atkinson, Anita Ferreira, and Elvis Aravena. Discovering implicit intention-level knowledge from natural-language texts. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 249–262, 2008.
- [4] Elizabeth Brondolo and Kristy-Lee Jean-Pierre. “You said, I heard”: Speaking the subtext in interracial conversations. In *Remediation in Medical Education*, pages 131–156. Springer, 2014.
- [5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.
- [6] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Multi-fiber networks for video recognition. In *ECCV*, pages 352–367, 2018.
- [7] Guillaume Chevance, Yannick Stephan, Nelly Héraud, and Julie Boiché. Interaction between self-regulation, intentions and implicit attitudes in the prediction of physical activity among persons with obesity. *Health Psychology*, 37(3):257–261, 2018.
- [8] Sumit Chopra, Michael Auli, and Alexander M Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL*, pages 93–98, 2016.
- [9] Nilanjana Dasgupta and Luis M Rivera. When social context matters: The influence of long-term contact and short-term exposure to admired outgroup members on implicit attitudes and behavioral intentions. *Social Cognition*, 26(1):112–123, 2008.
- [10] 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.
- [11] Mingyu Ding, Zhe Wang, Bolei Zhou, Jianping Shi, Zhiwu Lu, and Ping Luo. Every frame counts: Joint learning of video segmentation and optical flow. In *AAAI*, pages 10713–10720, 2020.
- [12] Mingyu Ding, An Zhao, Zhiwu Lu, Tao Xiang, and Ji-Rong Wen. Face-focused cross-stream network for deception detection in videos. In *CVPR*, pages 7802–7811, 2019.
- [13] Paul Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003.
- [14] Paul Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. WW Norton & Company, 2009.
- [15] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941, 2016.
- [16] Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. In *ACL*, pages 171–175, 2012.
- [17] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *ICML*, pages 1243–1252, 2017.
- [18] 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.
- [19] Pär Anders Granhag and Maria Hartwig. A new theoretical perspective on deception detection: On the psychology of instrumental mind-reading. *Psychology, Crime & Law*, 14(3):189–200, 2008.
- [20] Peter Gräsch and Alexander Felfernig. On the importance of subtext in recommender systems. *icom*, 14(1):41–52, 2015.
- [21] Julia Hirschberg, Stefan Benus, Jason M Brenier, Frank Enos, , et al. Distinguishing deceptive from non-deceptive speech. In *INTERSPEECH*, pages 1833–1836, 2005.
- [22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [23] 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.
- [24] Yuqi Huo, Xiaoli Xu, Yao Lu, Yulei Niu, Mingyu Ding, Zhiwu Lu, Tao Xiang, and Ji-rong Wen. Lightweight action recognition in compressed videos. In *ECCV Workshop*, 2020.
- [25] 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.
- [26] Gangeshwar Krishnamurthy, Navonil Majumder, Soujanya Poria, and Erik Cambria. A deep learning approach for multimodal deception detection. *arXiv preprint arXiv:1803.00344*, 2018.
- [27] 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.
- [28] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093, 2019.
- [29] Beth Logan. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, pages 1–11, 2000.
- [30] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*, pages 6752–6761, 2018.
- [31] 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 *ACTEA*, pages 33–38, 2012.
- [32] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Deception detection using real-life trial data. In *ICMI*, pages 59–66, 2015.

- [33] 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.
- [34] Tomas Pfister and Matti Pietikäinen. Automatic identification of facial clues to lies. *SPIE Newsroom*, 2012.
- [35] John A Podlesny and David C Raskin. Physiological measures and the detection of deception. *Psychological Bulletin*, 84(4):782–799, 1977.
- [36] John A Podlesny and David C Raskin. Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology*, 15(4):344–359, 1978.
- [37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.
- [38] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [39] Yemin Shi, Yonghong Tian, Yaowei Wang, Wei Zeng, and Tiejun Huang. Learning long-term dependencies for action recognition with a biologically-inspired deep network. In *ICCV*, pages 716–725, 2017.
- [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.
- [41] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE signal processing letters*, 6(1):1–3, 1999.
- [42] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.
- [43] Panagiotis Tsiamyrtzis, Jonathan Dowdall, Dvijesh Shastri, Ioannis T Pavlidis, MG Frank, and P Ekman. Imaging facial physiology for the detection of deceit. *International Journal of Computer Vision*, 71(2):197–214, 2007.
- [44] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *ICCV*, pages 4534–4542, 2015.
- [45] 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.
- [46] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based LSTM for aspect-level sentiment classification. In *EMNLP*, pages 606–615, 2016.
- [47] Zhe Wu, Bharat Singh, Larry S Davis, and VS Subrahmanian. Deception detection in videos. In *AAAI*, pages 1695–1702, 2018.
- [48] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.
- [49] Jiaqi Yan, Xin Li, Yani Shi, Sherry Sun, and Huaqing Wang. The effect of intention analysis-based fraud detection systems in repeated supply chain quality inspection: A context of learning and contract. *Information & Management*, page 103177, 2019.
- [50] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *AAAI*, pages 1655–1661, 2017.
- [51] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. ECO: Efficient convolutional network for online video understanding. In *ECCV*, pages 695–712, 2018.