Audio-Visual Deception Detection: DOLOS Dataset and Parameter-Efficient Crossmodal Learning

Xiaobao Guo[12*] Nithish Muthuchamy Selvaraj[3*] Zitong Yu[3†] Adams Wai-Kin Kong[2]
Bingquan Shen[4] Alex Kot[3]

[1]Rapid-Rich Object Search (ROSE) Lab, Interdisciplinary Graduate Programme, Nanyang Technological University
[2]School of Computer Science and Engineering, NTU [3]School of Electrical & Electronic Engineering, NTU
[4]DSO National Laboratories, Singapore

{xiaobao001, ms.nithish, zitong.yu, adamskong, eackot}@ntu.edu.sg sbingqua@dso.org.sg

Abstract

Deception detection in conversations is a challenging yet important task, having pivotal applications in many fields such as credibility assessment in business, multimedia anti-frauds, and custom security. Despite this, deception detection research is hindered by the lack of high-quality deception datasets, as well as the difficulties of learning multimodal features effectively. To address this issue, we introduce DOLOS[1], the largest gameshow deception detection dataset with rich deceptive conversations. DOLOS includes 1,675 video clips featuring 213 subjects, and it has been labeled with audio-visual feature annotations. We provide train-test, duration, and gender protocols to investigate the impact of different factors. We benchmark our dataset on previously proposed deception detection approaches. To further improve the performance by fine-tuning fewer parameters, we propose Parameter-Efficient Crossmodal Learning (PECL), where a Uniform Temporal Adapter (UT-Adapter) explores temporal attention in transformer-based architectures, and a crossmodal fusion module, Plug-in Audio-Visual Fusion (PAVF), combines crossmodal information from audio-visual features. Based on the rich fine-grained audio-visual annotations on DOLOS, we also exploit multi-task learning to enhance performance by concurrently predicting deception and audio-visual features. Experimental results demonstrate the desired quality of the DOLOS dataset and the effectiveness of the PECL. The DOLOS dataset and the source codes are available at here.

1. Introduction

Deception is a pervasive and complex phenomenon that occurs in all areas of life, and understanding its nature and impact is crucial in preventing negative consequences. Effective deception detection has crucial implications in the area of border security, anti-fraud, business negotiations and etc. [14, 12, 1, 13, 9, 27, 44]. Deep learning algorithms have achieved comparable or even better performance than human in many complex tasks [22, 41, 40, 7, 38]. One may expect AI models can also bring significant breakthroughs in deception detection. Albeit the fruitful progress in computer vision [48, 10, 31, 46] and audio representation learning [5, 18, 28], it remains a significant challenge to efficiently explore AI ability to process multimodal information in perceiving and predicting human deceptive behaviors.

The performance of AI models in deception detection is heavily reliant on the availability of authentic and effective deception samples from the real world. The deceptive subjects must be spontaneous and motivated [8, 49] such that certain behavioral cues (e.g., vocal pitch and chin raise) could be more pronounced. Public benchmark datasets collected from court trials [36], game shows [42], and lab-based scenarios [15], have contributed to spurring interest and progress in deception detection research. Deception can be affected by a range of factors in diverse real-world situations. Thus, it is necessary to create deception datasets from different scenarios. However, current datasets are still insufficient to drive further progress and inspire novel ideas due to their limitations on both quantity and quality. These limitations include (1) the small number of deceptive samples and subjects, (2) the lack of rich annotated visual and speech attributes, and (3) the variety of protocols. It is imperative to build a larger and richer deception detection dataset. In particular, more deceptive samples and subjects, better annotations for facial movements, gestures, and audio attributes,

*Equal contribution
†Corresponding author
[1]The name "DOLOS" comes from Greek mythology.

and more types of protocols for investigating factors affecting deception detection.

In addition to a high-quality dataset, effective methods are equally important to deception detection. A variety of works have been done towards using visual and acoustic information in videos for deception detection [42, 15, 36]. Current methods fall into two categories: unimodal learning and multimodal fusion. However, both approaches have limitations as they do not fully exploit unimodal features or integrate complementary information from multiple modalities. To make better use of available information, fine-tuning large pre-trained models has shown promising results [39, 26, 20, 47]. However, fully fine-tuning pre-trained models, such as W2V2 [5] and ViT [10], can be inefficient and lead to overfitting, especially if the downstream dataset is limited. Therefore, it is important to consider parameter efficiency when fine-tuning pre-trained models. Adapters [17, 21] offer an efficient approach for fine-tuning models. While originally proposed for language [17] or vision [21] tasks, adapters have not yet been applied to multiple modalities for temporal feature extraction.

Contributions. In this paper, we establish a new deception detection dataset and propose parameter-efficient cross-modal learning for audio-visual deception detection.

As our first contribution, we introduce DOLOS, a new gameshow dataset for audio-visual deception detection. DOLOS has several merits compared with current public datasets. First, the gameshow is a reliable source for collecting deception detection data because all the participants are motivated to cheat and the ground truths are available. Second, the proposed dataset is in a conversational setup, where the deception behaviors are more naturally presented. Third, our dataset is the largest in terms of the number of subjects and also the largest non-lab based dataset in terms of the number of video clips. The dataset has been labeled with fine-grained audio-visual annotations. We also benchmark our dataset on previous deception detection approaches that involve both unimodal and multimodal features. In addition, we offer three distinct protocols, namely train-test, duration, and gender, to explore various factors that may impact deception detection.

Our second contribution is Parameter-Efficient Cross-modal Learning (PECL), a method for deception detection that achieves high performance by fine-tuning a small number of extra learnable weights. Specifically, we introduce a Uniform Temporal Adapter (UT-Adapter) that explores the temporal attention between input embeddings for both visual and audio modalities without the need for delicate modifications. For multimodal fusion, we propose Plug-in Audio-Visual Fusion (PAVF), which utilizes the complementary information between visual and audio features to enhance the overall performance. PECL is designed to be parameter efficient, with only the UT-Adapter, PAVF modules, and the classifier being trainable. Furthermore, we explore the benefits of multi-task learning, a proven method that enhances performance in audio and visual tasks [54, 45]. By simultaneously predicting deception, facial movements, and phonetic features, our proposed method can be further improved.

To compare and show the advantages of our dataset and the proposed method, we conducted extensive experiments. In the cross-testing with the current gameshow benchmark dataset [42], DOLOS performs better on several test sets in different scenarios. The experimental results on DOLOS also showed that PECL yields superior performance on different protocols. Through our experiments, we uncover valuable insights into multimodal deception detection. We believe that DOLOS, the proposed method, and the benchmarking experiments provide valuable resources to other researchers in advancing research in this area.

2. Related Work

2.1. Deception Detection Datasets

Early works in deception detection research investigated the psychological premises of deception [8, 16, 50, 30, 49] and identified potential cues, such as physiological, visual, vocal, verbal and behavioral cues. The pioneering work on multimodal deception detection emerged with the introduction of the Real-Life Trials (RLT) dataset [36] collected from courtroom trials. The courtroom setting provides a valuable and authentic scenario of people lying, but RLT is small and its performance bottleneck was quickly reached. Pérez-Rosas *et al.* [37] and Kamboj *et al.* [23] later proposed another two new datasets collected from street interviews and political speeches. However, it is difficult to verify the veracity of the responses of the subjects in these datasets. Several lab-based multimodal deception detection datasets were proposed to overcome these shortcomings. The settings of these datasets were typically in situations that speakers were allowed to lie, such as describing an object [15], a personality [32], and face-to-face interviews with the subjects [43]. However, the incentive to deceive may be low, leading to fewer deceptive cues captured.

Reality TV shows based on truth or lie games provide an avenue to collect examples of conversational deception. They also offer a sweet spot among incentive, veracity, and sample size requirements. Often, the participants lie to each other to win the game, and the truth is revealed at the end of the round. Large amounts of gameshow footage are available online. The first gameshow-based deception detection dataset was proposed in [42] and it primarily focused on two modalities: text and manually annotated visual features. However, directly applying this dataset for multimodal deception detection by using audio and visual modalities faces several challenges, which we will describe in Sec. 3.

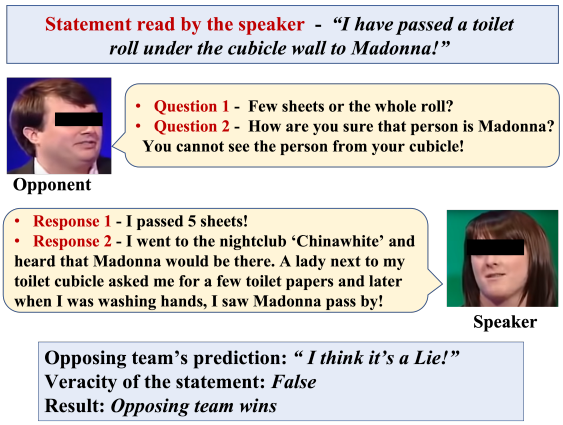


Figure 1. Gameshow format. An example of a deceptive round.

2.2. Multimodal Deception Detection

Previous works on multimodal deception detection include using unimodal and multimodal features and exploring effective fusion mechanism [33, 24, 25, 4, 9, 51, 13]. Gogate *et al.* [13] proposed a deep model that incorporated the audio cues with visual and text modalities to improve the accuracy of deception prediction. Karimi *et al.* [24] explored deceptive cues from RGB images and raw audio in an end-to-end manner. Wu *et al.* [51] utilized several types of features, including micro-expression and IDT (Improved Dense Trajectory) features from RGB images, MFCC (Mel-frequency Cepstral Coefficients) features from the audio, and transcripts. Avola *et al.* [4] addressed the deception detection issue by directly using AUs (facial action units) from video frames and classifying them through an SVM. Ding *et al.* [9] tried to combine facial expressions and body movements to extract more visual deception cues. They also tried to improve the performance by involving audio features and transcripts. Mathur and Matarić [33] focused on facial expressions for deception detection and studied interpretable features from visual, vocal, and verbal modalities. Karnati *et al.* [25] proposed deep networks that learn audio, visual, and EEG representations for deception detection.

Most of the previous works tried to enhance the performance by involving more modalities, more features, and using better feature extraction methods. However, they heavily rely on spatial features rather than temporal information within and between the modalities, which may lose potentially important cues and lead to sub-optimal results.

In terms of fusion methods, the feature-level fusion which concatenates multimodal embeddings and uses the linear layers to extract crossmodal information [24, 4, 9, 51, 13] was the most popular approach. Other works implement decision-level fusion, such as score-level fusion [25, 13]. In all cases, performance can be boosted by a reliable and efficient fusion method.

3. The DOLOS Dataset

3.1. Gameshow Format

We collect data from a British reality comedy gameshow available on YouTube [2], where six participants compete as two teams. In each taking turn, one of their members read out a statement about their personal life from a card. This statement could be either true or false, and only the speaker knows which one it is. The opposing team then asks the speaker a series of questions, and the speaker must provide answers in defense of their statement. The goal of the speaker is to convince the opposing team that their statement is true, while the opposing team tries to determine whether the statement is actually true or false based solely on the speaker's answers. At the end of each round, the veracity of the statement is revealed, and if the opposing team correctly predicts whether it was true or false, they win the round; otherwise, the speaker's team wins. The game continues for several rounds until all participants have played. Notably, if the statement is true, then the speaker's answers will be considered as truthful video clips and vice versa. The gameshow format is given in Fig. 1.

3.2. Data Collection and Annotation Procedure

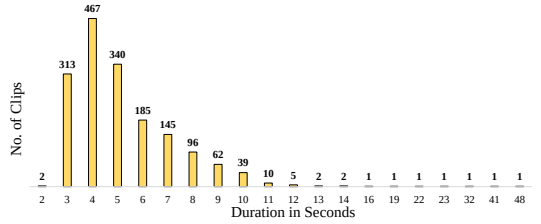
For each episode of the gameshow videos, the video clips that satisfy the following requirements are extracted:

- The participant speaks only the relevant content (*i.e.*, telling the truth or lies) in a clear voice and without strong background noise.
- The face of the participant is clearly visible without occlusion.

From a total of 84 episodes, we extracted 1675 video clips from 213 (141 Male and 72 Female) participants. The duration of the video clips ranges from 2-19 seconds. The extracted video clips are manually annotated for non-verbal deceptive cues using the popular MUMIN coding scheme [2], where we particularly focus on the visual (25 facial) and vocal (5 speech) features. The non-verbal features in the MUMIN coding scheme and the distributions of video clips are shown in Fig. 2.

To eliminate bias among the human annotators, all six human annotators underwent calibration by annotating the same subset of the dataset, and their performance was evaluated using Cohen's Kappa scores [6]. If significant discrepancies are found in visual and audio annotations between annotators, they reconciled their differences through discussion and then re-annotated the subset. This iterative process was repeated until inter-annotator agreement was achieved. Before calibration, the average Cohen's Kappa scores were 0.5 (0.53 for audio and 0.47 for visual annotations). After

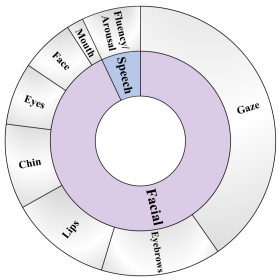
[2] Videos are collected from YouTube under the fair use policy



(a) Distribution of clips (by duration)

MUMIN Coding Scheme			
Lips Corners up Corners down Protruded Retracted	**Gaze** Towards interlocutor Up Down Sideways	**Chin** Move Upwards Move Downwards **Eyebrows** Frowning Raising	**Face** Smile Laughter Scowl
Eyes Frequent Blinking Exaggerated Opening Exaggerated Closing	**Mouth** Exaggerated Opening Closed (awkward pauses)	**Fluency** Non-ah-mmh speech disturbances Word and phrase repetitions Silent pauses	**Arousal** Loudness Tension

(b) Annotations - facial and speech features



(c) Distribution of features

	Deception	Truthful	Total
Male	746	578	1324
Female	153	198	351
Total	899	776	1675

(d) Distribution of clips (by veracity & gender)

Figure 2. DOLOS dataset for multimodal deception detection. Zoom in for a better view.

Dataset	Hand annotated features	#Subjects	Samples				Scenario
			Total	Deceptive	Truthful	Deception/Truth Ratio	
Real Life Trials[36]	✓	56	121	61	60	1.02	Court-room
Real Life Deception[37]		-	56	31	25	1.24	Street interviews
Political Deception[23]		88	180	-	-	-	Political speech
Bag of Lies[15]		35	325	162	163	0.99	Lab
MU3D[32]		80	320	160	160	1	Lab
Deception Detection and remote PPG[43]		70	1680	630	1050	0.6	Lab
Box of Lies[42]	✓	26	1049	862	187	4.61	Gameshows
DOLOS (Ours)	✓	213	1675	899	776	1.16	Gameshows

Table 1. Comparison of multimodal deception detection datasets

calibration, the average scores increased to 0.65 (0.67 for audio and 0.63 for visual annotations).

Based on the statistics in Fig. 2, we organize DOLOS dataset into three different protocols. First, we provide a train-test protocol with a 3-fold split to evaluate the performance of the deception detector for DOLOS dataset. Second, to reflect the variability in speaking duration in real-world scenarios, we provide a duration protocol with short clips (2-4s) and long clips (5-10s) based on the statistics from our dataset. We also provide a gender (male and female) protocol to investigate different factors in the deception detection task.

3.3. Comparison with Box of Lies and Other Datasets

We compare our DOLOS with Box of Lies (BOL) [42], which was also collected from gameshow videos. In terms of quantity, BOL has fewer video clips and fewer subjects. BOL contains only 26 subjects, which is 17.9% of DOLOS. In terms of quality, among 1,049 video clips in BOL, 537 video clips missed the speaker's face. In the other video clips, the speaker and the audio of some video clips do not match. In BOL, the host of the gameshow has more video clips than other subjects, which introduces a large bias. In comparison, the data collection requirements mentioned in Sec. 3.2 inherently prevent these problems. Therefore, DOLOS is more suitable for audio-visual deception detection research (see Sec. 5.3) than BOL.

We also compare DOLOS with other publicly available datasets. A summary of the comparisons is shown in Table 1. Briefly, DOLOS offers the following advantages:

- DOLOS is the largest in terms of number of subjects. It is also the largest in terms of video clips in the non-lab scenario.
- Compared to the Box of Lies gameshow dataset, DOLOS contains a larger number of video clips and is more well-balanced in terms of the proportion of deceptive and truthful samples.
- DOLOS provides manually annotated MUMIN features for all video clips. These annotations can be directly used for deception detection task. They can also be applied for other related tasks (*e.g.*, facial expression prediction) or multi-task learning.

4. Methodology

To efficiently leverage the prior knowledge from large-scale generic audio and visual tasks, we propose Parameter-Efficient Crossmodal Learning (PECL) for audio-visual deception detection. PECL's overall framework is illustrated in Fig. 3 with the main network architecture (Fig. 3 (a)) and several components with detailed structures (Fig. 3 (b)-(e)).

Specifically, in Fig. 3 (a), visual inputs are tokenized by a 2D-CNN module and audio inputs are tokenized by a

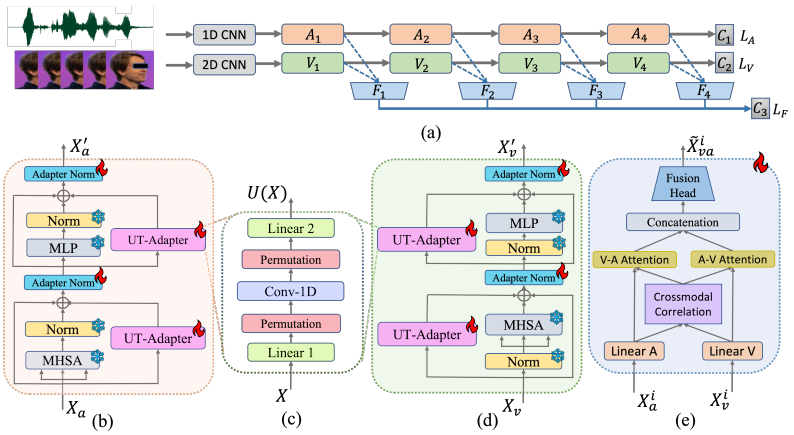


Figure 3. Parameter-Efficient Crossmodal Learning (PECL) Framework. (a) is the main architecture. (b) is the encoder layer A_i for audio features. (c) is the UT-Adapter with a shared structure between audio and visual features. (d) is the encoder layer V_i for visual features. (e) is the PAVF module F_i. The trainable modules and layers are marked by "fire", and the frozen layers are marked by "snowflakes" .

1D-CNN module (See Sec. 5.1) with the same dimensions, which are denoted as $\boldsymbol{X_v} \in \mathbb{R}^{L\times D}$ and $\boldsymbol{X_a} \in \mathbb{R}^{L\times D}$, where L is the length of sequences and D is the dimension of each sequence embedding. V_i and A_i represent transformer-based visual and audio modality encoder layers, where we adopt W2V2 [5] as the backbone network for audio modality and ViT [10] for visual modality. F_i are fusion modules between each pair of unimodal encoder layers and $i = 1, 2, 3, 4$. C_1, C_2, and C_3 are classifiers for audio, visual, and fusion networks, respectively. L_A, L_V, and L_F are cross-entropy losses for audio, visual, and fused modalities, respectively.

4.1. Uniform Temporal Adapter (UT-Adapter)

To improve the parameter efficiency and alleviate the overfitting issue, we insert learnable adapter layers [17, 21] in the pre-trained visual and audio models. During training, all the weights, except those in the adapter layers and the classification layers are frozen. Unlike NLP-Adapter for language tasks [17] and Conv-Adapter for vision tasks [21], our adapter layer, named Uniform Temporal Adapter (UT-Adapter), is to explore temporal attention for both visual and audio modalities.

Let F_V and G_A represent the visual and audio encoder layers in Fig. 3. Their outputs are denoted as

$$\boldsymbol{X_v^{'}} = F_V(\boldsymbol{X_v}) \; and \; \boldsymbol{X_a^{'}} = G_A(\boldsymbol{X_a}), \tag{1}$$

where $\boldsymbol{X_v^{'}} \in \mathbb{R}^{L\times D}$ and $\boldsymbol{X_a^{'}} \in \mathbb{R}^{L\times D}$. F_V and G_A are formed by a Multi-Head Self-Attention (MHSA) module, UT-Adapters, a Multi-Layer Perception (MLP) module, and LayerNorm (LN). To simplify the notations, let $\boldsymbol{X}$ be one of the input features, *i.e*, $\boldsymbol{X_v}$ and $\boldsymbol{X_a}$ (in Fig. 3 (b) and (d)). $\boldsymbol{X}$ is projected to a query $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W_Q}$, a key $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W_K}$, and a value $\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W_V}$, where $\boldsymbol{W_Q}$, $\boldsymbol{W_K}$, and $\boldsymbol{W_V} \in \mathbb{R}^{D\times D}$ are pre-trained projection weights. For a single head $H_j, j \in \{1, 2, \cdots, N\}$, the self-attention operation is formulated as

$$H_j(\boldsymbol{X}) = Softmax\left(\frac{\boldsymbol{Q_j}\boldsymbol{K_j}^{\intercal}}{\sqrt{D}}\right)\boldsymbol{V_j}, \tag{2}$$

and the outputs of all the $H_j, j \in \{1, 2, \cdots, N\}$ are then concatenated and projected to a single output $H(\boldsymbol{X})$.

UT-Adapters are placed in parallel with MHSA and MLP modules. In particular, a UT-Adapter is a stack of linear and 1D-convolutional layers. As shown in Fig. 3 (c), the output of a UT-Adapter is operated as

$$U(\boldsymbol{X}) = L_2(P(C(P(L_1(\boldsymbol{X}; \boldsymbol{W_1})); \boldsymbol{W_C})); \boldsymbol{W_2}), \tag{3}$$

where L_1 and L_2 are *Linear 1* and *Linear 2* layers in Fig. 3 (c) and P and C are *Permutation* and *Conv-1D*, respectively. $\boldsymbol{W_1} \in \mathbb{R}^{D\times 128}$ and $\boldsymbol{W_2} \in \mathbb{R}^{128\times D}$ are trainable weights of L_1 and L_2. $\boldsymbol{W_C}$ is the 1D-convolutional weight with the kernel size of 3. L_1 projects the input $\boldsymbol{X} \in \mathbb{R}^{L\times D}$ to $\boldsymbol{X} \in \mathbb{R}^{L\times 128}$, Permutation layer P shifts $\boldsymbol{X} \in \mathbb{R}^{L\times 128}$ to $\boldsymbol{X} \in \mathbb{R}^{128\times L}$ and C is applied along the temporal dimension to capture the temporal dynamics. P layer and L_2 then project $\boldsymbol{X} \in \mathbb{R}^{128\times L}$ to $\boldsymbol{X} \in \mathbb{R}^{L\times D}$.

To be specific, for F_V, the output from MHSA and a UT-adapter can be represented as:

$$\boldsymbol{X_v^{''}} = AN(\boldsymbol{X_v} + H(LN(\boldsymbol{X_v})) + U(\boldsymbol{X_v})), \tag{4}$$

and $\boldsymbol{X_v^{'}}$ in Eq. 1 is obtained by:

$$\boldsymbol{X_v^{'}} = AN(\boldsymbol{X_v^{''}} + MLP(LN(\boldsymbol{X_v^{''}})) + U(\boldsymbol{X_v^{''}})), \tag{5}$$

where AN is a normalization layer introduced after UT-Adapter, which is trainable.

G_A shares a similar structure with F_V. The output $\boldsymbol{X_a}$ is computed by

$$\begin{aligned} \boldsymbol{X_a^{''}} &= AN(\boldsymbol{X_a} + LN(H(\boldsymbol{X_a})) + U(\boldsymbol{X_a})), \\ \boldsymbol{X_a^{'}} &= AN(\boldsymbol{X_a^{''}} + LN(MLP(\boldsymbol{X_a^{''}})) + U(\boldsymbol{X_a^{''}})). \end{aligned} \tag{6}$$

The input $\boldsymbol{X_v}$ and $\boldsymbol{X_a}$ are sequence embeddings from vision and audio. The MHSA module in the transformer encoder facilitates global attention, allowing the model to effectively learn both spatial and temporal attention from the input. However, this may not be optimal for learning local temporal attention and local spatial information. Thus, UT-Adapter is proposed to capture the local temporal attention in parallel with MHSA and MLP modules. The 2D-CNN captures the local spatial information and the 1D-CNN captures temporal information. We empirically show that this *adapter-transformer* architecture is parameter-efficient tuning and delivers better performance.

4.2. Plug-in Audio-Visual Fusion (PAVF)

To explore the audio-visual interactions, we propose a Plug-in Audio-Visual Fusion (PAVF) module that learns crossmodal attention for fusion.

We denote the output features from V_i and A_i as $\boldsymbol{X_v^i} \in \mathbb{R}^{L\times D}$ and $\boldsymbol{X_a^i} \in \mathbb{R}^{L\times D}$ and propose to learn crossmodal attention in an embedding space with lower dimension, which is more efficient and reduces computational costs. $\boldsymbol{X_v^i}$ and $\boldsymbol{X_a^i}$ are first projected to $\boldsymbol{X_v^i} \in \mathbb{R}^{L\times D'}$ and $\boldsymbol{X_a^i} \in \mathbb{R}^{L\times D'}$, $D' < D$, by linear projection layers. Crossmodal correlation $\boldsymbol{P^i} \in \mathbb{R}^{L\times L}$ is learned to indicate the importance between visual sequences and audio sequences. A specific visual sequence may have a higher correlation to certain audio sequences in a video clip, which offers important cues to the deception detection task. To calculate $\boldsymbol{P^i}$, we introduce a trainable crossmodal correlation weight matrix $\boldsymbol{W_P^i} \in \mathbb{R}^{D'\times D'}$ and the calculation of $\boldsymbol{P^i}$ is

$$\boldsymbol{P^i} = \boldsymbol{X_a^i} \boldsymbol{W_P^i} \boldsymbol{X_v^{i\intercal}}. \tag{7}$$

We further use the learned crossmodal correlation $\boldsymbol{P^i}$, to perform $\boldsymbol{X_v^i} \rightarrow \boldsymbol{X_a^i}$ and $\boldsymbol{X_a^i} \rightarrow \boldsymbol{X_v^i}$ attentions. The attended features are

$$\begin{aligned} \tilde{\boldsymbol{X}}_{\boldsymbol{v}}^{\boldsymbol{i}} &= Softmax(\boldsymbol{P^i}) \boldsymbol{X_v^i} + \boldsymbol{X_v^i}, \\ \tilde{\boldsymbol{X}}_{\boldsymbol{a}}^{\boldsymbol{i}} &= Softmax(\boldsymbol{P^{i\intercal}}) \boldsymbol{X_a^i} + \boldsymbol{X_a^i}, \end{aligned} \tag{8}$$

where $\boldsymbol{P^{i\intercal}}$ is the transpose of $\boldsymbol{P^i}$, $\tilde{\boldsymbol{X}}_{\boldsymbol{v}}^{\boldsymbol{i}} \in \mathbb{R}^{L\times D'}$, and $\tilde{\boldsymbol{X}}_{\boldsymbol{a}}^{\boldsymbol{i}} \in \mathbb{R}^{L\times D'}$. Softmax is conducted column-wise. The attended features are concatenated and inputted into a fusion head. The fusion head consists of a linear projection L_p that further reduces the embedding dimension to D'', a normalization layer LN, and a non-linear activation $ReLU$. It can be denoted as

$$\tilde{\boldsymbol{X}}_{\boldsymbol{va}}^{\boldsymbol{i}} = ReLU(LN(L_p(\tilde{\boldsymbol{X}}_{\boldsymbol{v}}^{\boldsymbol{i}} \oplus \tilde{\boldsymbol{X}}_{\boldsymbol{a}}^{\boldsymbol{i}}))), \tag{9}$$

where $\tilde{\boldsymbol{X}}_{\boldsymbol{va}}^{\boldsymbol{i}} \in \mathbb{R}^{L\times D''}$ and $\oplus$ is a concatenation operation.

PAVF modules are inserted into each pair of V_i and A_i to perform fusion, in which richer fused information can be learned. The outputs $\tilde{\boldsymbol{X}}_{\boldsymbol{va}}^{\boldsymbol{i}}$ from each PAVF module are then concatenated and used for prediction.

4.3. Multi-task Learning

Multi-task learning has demonstrated promising performance in various audiovisual tasks [19, 45], effectively enhancing the performance of multiple tasks through mutual learning. With DOLOS' MUMIN features (Fig. 2 (b)), we enhance performance through multi-task learning, simultaneously conducting two prediction tasks using fused multimodal features. More clearly, for K MUMIN features labels and one deception-truth label, we predict $K+1$ labels. Note that all the labels are binary. We denote the fused audio-visual embedding as $\tilde{\boldsymbol{X}}_{\boldsymbol{va}}$. The classification layer takes $\tilde{\boldsymbol{X}}_{\boldsymbol{va}}$ as input to produce prediction scores $S \in \mathbb{R}^{1\times(K+1)}$. Given the ground-truth label $Y \in \mathbb{R}^{1\times(K+1)}$, we use cross-entropy loss to perform multi-task learning. It can be formulated as

$$\mathcal{L}_M = -\sum_{k=1}^{K+1} \left(Y_k \log(S_k) + (1-Y_k)\log(1-S_k)\right), \tag{10}$$

where $\mathcal{L}_M$ is the sum of the cross-entropy losses for each element in S and Y.

5. Experimental Results

5.1. Implementation Details

Data pre-processing. For each video clip, we evenly sampled $L = 64$ images and cropped the face areas using MTCNN face detector [53]. These images were normalized and resized to 160×160 pixels. The raw speech audio was resampled such that the W2V2 feature extractor outputs $L = 64$ tokens. No data augmentation was applied to both modalities.

Model details. We used ImageNet pre-trained ViT as the backbone network for visual modality. We tokenized face images with a 2D-CNN module, which resulted in a feature with a dimension of 64×256. For the audio modality, we adopted the pre-trained W2V2 model. The raw audio was tokenized by the 1D-CNN module and the feature size was 64×512 for each audio sample. Using a linear projection layer, visual and audio tokens were projected to 64×768 dimensions. We empirically utilized the first 4 transformer encoder layers from ViT and W2V2 models to extract features because deception detection can benefit from the low-level features [44, 29, 13]. The UT-Adapter and PAVF modules were inserted in these four encoder layers (in Fig. 3 (a)). The one-dimensional convolution layer (in Fig 3 (c)) in the UT-Adapter for both visual and audio encoders had a kernel size of 3 and a stride of 1.

Training and evaluation. The models were trained with cross-entropy loss. We used Adam optimizer and trained the models for 20 epochs with a learning rate of 3e-4 and a batch size of 16. We evaluated our method with accuracy (ACC), F1 score (F1), and Area Under Curve (AUC).

Modalities	3-Fold Average			Duration Protocol			Gender Protocol		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Visual	61.44	69.42	58.89	61.03	**72.01**	56.51	59.37	64.19	54.94
Audio	59.19	**73.46**	52.54	58.24	71.83	52.38	52.62	63.22	51.08
Concatenation	61.62	70.2	60.5	60.8	69.63	57.79	58.0	66.22	53.75
Fusion (PAVF)	64.75	71.2	62.71	62.43	70.04	59.92	58.28	65.41	53.31
PAVF + Multi-task	**66.84**	73.35	**64.58**	**64.48**	71.09	**62.44**	**59.04**	**66.84**	**55.1**

Table 2. Multimodal deception detection on DOLOS dataset. The metrics are ACC (%), F1 (%), and AUC(%).

	Method	Train Data - DOLOS		Train Data - Box of Lies	
		Test BgOL	Test RLT	Test BgOL	Test RLT
V	RN18+LSTM	55.08 (+4.62)	56.62 (+3.85)	50.46	52.77
	3D RN	55.69 (+3.38)	**57.87** (+**5.53**)	52.31	52.34
	Ours (Vision)	**56.92** (+4.30)	55.32 (+2.55)	52.62	52.77
A	RN18 †	**55.84** (+**4.91**)	53.53 (+1.64)	50.93	51.89
	MLP (MFCC)	52.31 (+2.46)	52.34 (+0.42)	49.85	51.92
	Ours (Audio) ‡	55.39 (+4.00)	**59.15** (+**4.68**)	51.39	54.47
V+A	3D RN+RN18	50.77 (+1.23)	54.47 (+1.70)	49.54	52.77
	Ours (PAVF)	**57.54** (+**4.31**)	**56.17** (+**2.98**)	53.23	53.19

Table 3. Cross-testing on Bag-of-Lies (BgOL) and Real-Life-trials (RLT) datasets. The metric is ACC (%). RN stands for ResNet. All the visual features are from face frames. † is trained on mel-spectrogram and ‡ is trained on raw audio.

	Features	Method	ACC (%)
V	Open Face [33, 29]	LSTM	56.81
	Action Units (AU) [33, 4]	LSTM	57.47
	Facial Affect [33, 24]	LSTM	57.67
	MUMIN Features [36, 37]	MLP	58.84
	Face [25, 9, 51, 13, 29, 24]	RN18+LSTM	59.55
	Face	Ours (Vision)	**61.44**
A	MFCC [51]	MLP	56.86
	Open SMILE [33, 13, 52]	MLP	57.49
	W2V2 Audio [25, 9, 29, 24]	Ours (Audio)	**59.19**
V+A	Open Face + Open SMILE [33]	Score-Fusion	59.75
	Face + Open SMILE [13, 29, 52]	Score-Fusion	60.00
	Face + W2V2 Audio [25, 24, 9]	Ours(Concat)	**61.62**
	Face + W2V2 Audio	Ours(PAVF)	**64.75**

Table 4. Benchmarking DOLOS dataset on visual and audio features. The weighted average is used for score fusion. RN18 stands for ResNet-18.

5.2. Audio-visual Deception Detection in DOLOS

The performance of PECL model on DOLOS is presented in Table 2. The average results were calculated from the 3 folds defined in the train-test protocol. Duration and gender protocol results were cross-tested averages of training/testing on (long/short + short/long) and (male/female + female/male), respectively. For unimodal results, the visual modality showed slightly better performance than the audio across all the protocols. To study the performance of a simple feature-level fusion, we concatenated the last output embeddings from visual and audio encoders to perform prediction. The concatenation usually draws the average of unimodal predictors. To overcome this issue, the proposed PAVF module learned the correlation between visual and audio features from different encoder layers and improved the performance. The multi-task learning further boosted the accuracy across all the protocols. However, by comparing the overall performances, we observed that the duration and gender factors showed an adverse effect on performance. These results pointed out the importance of duration and gender factors in deception detection.

5.3. Comparison with Box of Lies

To compare DOLOS with the current gameshow dataset BOL, we conducted cross-testing, where DOLOS and BOL were used for training and other deception detection datasets, *i.e.*, Bag of Lies and Real Life Trials, for testing. It was challenging because these datasets were collected from different deceptive scenarios. As shown in Table 3, we compared several methods for audio, visual, and fusion and found that models trained on DOLOS achieved higher accuracies overall. PECL model trained on DOLOS also showed better test accuracy than that trained on BOL, (*e.g.*, Fusion (PAVF) improved 4.31% on BgOL and 2.98% on RLT). These results suggest that models trained on DOLOS can be more effective in identifying deceptive content in other scenarios than those trained on BOL.

5.4. Benchmarking DOLOS

We benchmark DOLOS in Table 4, following the deception detection literature. It is worth noting that the existing literature on deception detection is not standardized with protocols, data processing, and feature extraction methodologies. To highlight and cover the most representative works in deception detection literature, we benchmarked DOLOS on the popular visual and audio features used before for deception detection. The OpenFace features (2D facial landmarks, eye gaze, and head pose) and Action Units (AU) features were extracted using the OpenFace toolkit [3]. The facial affect (emotion) features were extracted using the Affectnet [34] model. For audio, the Mel Cepstral Frequency Coefficients (MFCC) and openSMILE features were extracted by the openSMILE toolkit [11].

Performance. For visual modality, the RGB face images gave the best performance, since they contained more deceptive cues. This was followed by manually annotated MUMIN features, affect features, and AU features, which all represented high-level human expressions. The low-level facial landmark features did not perform well. For audio modality, raw speech audio with rich deceptive cues performed the best. The openSMILE features performed better than MFCC, since it extracted nearly three times the audio

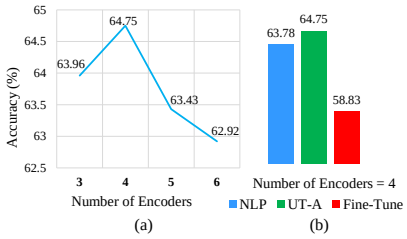


Figure 4. Ablation on (a) encoder depth and (b) type of adapter.

Method	Trainable			Total	Trainable/Total
	Adapter	PAVF	Total		
Fine-Tune	–	1.982	70.180	70.180	100%
NLP [17]	5.345	1.982	7.327	73.347	9.99%
UT-A	3.077	1.982	**5.059**	71.079	**7.12%**

Table 5. Comparison of the number of trainable parameters and total parameters (Millions).

Position	|| MHSA || FFN	|| MHSA	|| FFN	MHSA △ FFN
ACC (%)	**64.75**	63.88	63.37	64.39

Table 6. Ablation study on UT-Adapter positions. \|\| indicates "in parallel with" and △ indicates "between".

features compared with MFCC. Score-level fusion was conducted for the best-performing features of the two modalities. The proposed transformer-based detector achieved the best in both unimodal and multimodal settings.

5.5. Ablation Study

The ablation studies were conducted on different components of PECL model based on the train-test protocol.

Impact of encoder depth and adapter types. Fig. 4 (a) indicated that the proposed PECL performed best with four layers and may overfit when additional layers were added. Fig. 4 (b) showed that UT-Adapter outperformed the NLP adapter [17] and full fine-tuning, which may be due to better temporal modeling with Conv1D layers in UT-Adapter.

Comparison of parameter amount. Table 5 demonstrated that our proposed method achieved the best performance while utilizing the least amount of trainable parameters when compared to both the NLP adapter and full fine-tuning approaches. Note that UT-Adapter performed best with 128 dimensions while NLP adapter with 256 dimensions.

Impact of adapter position. The results in Table 6 showed that the UT-Adapter worked the best when parallel to both MHSA and FFN layers [21], as they both captured the temporal attention. The performance dropped when parallel to only either layer. The performance was also close for UT-Adapter placed in-between MHSA and FFN layers.

Impact of PAVF modules. The number of PAVF modules in Fig. 5 corresponds to the last number of audio-visual encoders that operate with them. The advantage of fusion in earlier stages [35] can be clearly seen with a substantial increase in performance for four PAVF modules, where the

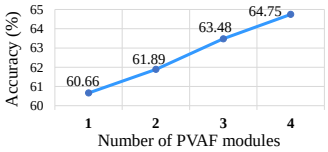


Figure 5. Ablation on numbers of PAVF modules.

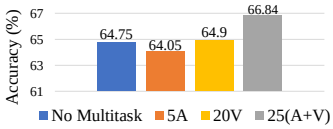


Figure 6. Results of multi-task learning with different features.

low-level crossmodal interactions were captured. The results revealed that four PAVF modules are particularly helpful in capturing audio-visual deceptive cues.

Impact of multi-task learning. In Fig. 6, we illustrated the ablation study of multi-task learning described in Sec 4.3, for different audio-visual features (Refer Fig 2 (b)). The results showed that using 20 visual features (20V) achieved better results than using 5 audio features (5A) and using 25 audio-visual features further boosted accuracy. The results demonstrated that the manually annotated features in DOLOS can benefit deception detection performance.

6. Conclusion

In this paper, we introduced DOLOS, the largest gameshow deception dataset with fine-grained audio-visual feature annotations. We also proposed Parameter-Efficient Crossmodal Learning, where the UT-Adapter learns temporal attention for both visual and audio modalities, and the PAVF module captures correlation information between audio and visual modalities. We benchmarked our dataset on the existing deception detection approaches. Extensive experiments showed that DOLOS has better quality than the Box of Lies and PECL model achieved better results on DOLOS and can be applied to other deceptive content.

Ethical Consideration. Deception detection research is crucial in credibility assessment, anti-frauds, border security, etc. However, the misuse of deception detection may cause potential negative impacts such as privacy invasion, discrimination, and/or harmful surveillance. This work was completed under an ethics clearance (IRB-2022-901). Data was collected from YouTube under the fair use policy and the subjects are public personalities.

Acknowledgments. This work was done at Rapid-Rich Object Search (ROSE) Lab, EEE, Nanyang Technological University. This research is supported in part by the NTU-PKU Joint Research Institute (a collaboration between the Nanyang Technological University and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation), and the DSO National Laboratories, Singapore, under the project agreement No. DSOCL21238.

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. 1

[2] Jens Allwood, Loredana Cerrato, Laila Dybkjaer, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. The mumin multimodal coding scheme. *NorFA yearbook*, 2005:129–157, 2005. 3

[3] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2):20, 2016. 7

[4] Danilo Avola, Luigi Cinque, Gian Luca Foresti, and Daniele Pannone. Automatic deception detection in rgb videos using facial action units. In *Proceedings of the 13th International Conference on Distributed Smart Cameras*, pages 1–6, 2019. 3, 7

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020. 1, 2, 5

[6] Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics*, 27(1):3–23, 1999. 3

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[8] Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. Cues to deception. *Psychological bulletin*, 129(1):74, 2003. 1, 2

[9] Mingyu Ding, An Zhao, Zhiwu Lu, Tao Xiang, and Ji-Rong Wen. Face-focused cross-stream network for deception detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7802–7811, 2019. 1, 3, 7

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, pages 0–7, 2021. 1, 2, 5

[11] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010. 7

[12] Tommaso Fornaciari and Massimo Poesio. Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21(3):303–340, 2013. 1

[13] Mandar Gogate, Ahsan Adeel, and Amir Hussain. Deep learning driven multimodal fusion for automated deception detection. In *2017 IEEE symposium series on computational intelligence (SSCI)*, pages 1–6. IEEE, 2017. 1, 3, 6, 7

[14] Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Frank Enos, Julia Hirschberg, and Sachin Kajarekar. Combining prosodic lexical and cepstral systems for deceptive speech detection. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages 0–7. IEEE, 2006. 1

[15] Viresh Gupta, Mohit Agarwal, Manik Arora, Tanmoy Chakraborty, Richa Singh, and Mayank Vatsa. Bag-of-lies: A multimodal dataset for deception detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–7, 2019. 1, 2, 4

[16] Julia Bell Hirschberg, Stefan Benus, Jason M Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, et al. Distinguishing deceptive from non-deceptive speech. 2005. 2

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2, 5, 8

[18] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 1

[19] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021. 6

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[21] Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*, 2022. 2, 5, 8

[22] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. 1

[23] Manvi Kamboj, Christian Hessler, Priyanka Asnani, Kais Riani, and Mohamed Abouelenien. Multimodal political deception detection. *IEEE MultiMedia*, 28(1):94–102, 2020. 2, 4

[24] Hamid Karimi, Jiliang Tang, and Yanen Li. Toward end-to-end deception detection in videos. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1278–1283. IEEE, 2018. 3, 7

[25] Mohan Karnati, Ayan Seal, Anis Yazidi, and Ondrej Krejcar. Lienet: a deep convolution neural networks framework for detecting deception. *IEEE Transactions on Cognitive and Developmental Systems*, 2021. 3, 7

[26] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 2

[27] Chenqi Kong, Kexin Zheng, Shiqi Wang, Anderson Rocha, and Haoliang Li. Beyond the pixel world: a novel acoustic-based face anti-spoofing system for smartphones. *IEEE Transactions on Information Forensics and Security*, 17:3238–3253, 2022. 1

[28] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 1

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

[30] Timothy R Levine and Steven A McCornack. Theorizing about deception. *Journal of Language and Social Psychology*, 33(4):431–440, 2014. 2

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[32] E Paige Lloyd, Jason C Deska, Kurt Hugenberg, Allen R McConnell, Brandon T Humphrey, and Jonathan W Kunstman. Miami university deception detection database. *Behavior research methods*, 51(1):429–439, 2019. 2, 4

[33] Leena Mathur and Maja J Matarić. Introducing representations of facial affect in automated multimodal deception detection. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 305–314, 2020. 3, 7

[34] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 7

[35] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 8

[36] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 59–66, 2015. 1, 2, 4, 7

[37] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, CJ Linton, and Mihai Burzo. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2336–2346, 2015. 2, 4, 7

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[39] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf, 2018. 2

[40] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016. 1

[41] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. 1

[42] Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1768–1777, 2019. 1, 2, 4

[43] Jeremy Speth, Nathan Vance, Adam Czajka, Kevin W Bowyer, Diane Wright, and Patrick Flynn. Deception detection and remote physiological monitoring: A dataset and baseline experimental results. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 2, 4

[44] Anastasis Stathopoulos, Ligong Han, Norah Dunbar, Judee K Burgoon, and Dimitris Metaxas. Deception detection in videos using robust facial features. In *Proceedings of the Future Technologies Conference*, pages 668–682. Springer, 2020. 1, 6

[45] Fei Tao and Carlos Busso. End-to-end audiovisual speech recognition system with multitask learning. *IEEE Transactions on Multimedia*, 23:1–11, 2020. 2, 6

[46] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021. 1

[47] Nik Vaessen and David A Van Leeuwen. Fine-tuning wav2vec2 for speaker recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7967–7971. IEEE, 2022. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[49] Aldert Vrij and Pär Anders Granhag. Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, 1(2):110–117, 2012. 1, 2

[50] Gemma Warren, Elizabeth Schertler, and Peter Bull. Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*, 33(1):59–69, 2009. 2

[51] Zhe Wu, Bharat Singh, Larry Davis, and V Subrahmanian. Deception detection in videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 3, 7

[52] Jun-Teng Yang, Guei-Ming Liu, and Scott C-H Huang. Multimodal deception detection in videos via analyzing emotional state-based feature. *arXiv preprint arXiv:2104.08373*, 2021. 7

[53] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 6

[54] Zixing Zhang, Bingwen Wu, and Björn Schuller. Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6705–6709. IEEE, 2019. 2