Supplementary Material for 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

1. The DOLOS Dataset

In this section, we describe the data collection, processing, and annotation procedure for the DOLOS dataset in detail. The DOLOS dataset collection is comprised of the extraction of video clips and annotation using the MUMIN coding scheme.

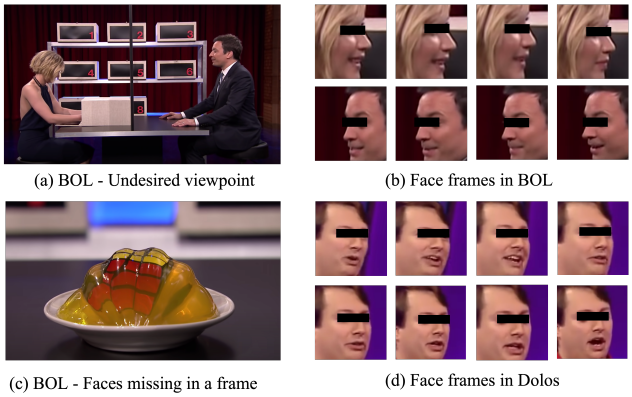
1.1. Extraction of Video Clips

The gameshow episodes from multiple seasons were first discovered on YouTube. To extract the relevant content from each round of an episode, the speakers' utterances (in the form of video clips) were identified by their timestamps. The clips were selected based on specific criteria, such as containing little noise and being of sufficient length to convey meaningful information. Multiple clips could be extracted from each round, all with the same veracity label. Python scripts were used to automate the download of the clips from YouTube, using the timestamps as reference. This process is described in more detail in Section 3.1 of the main paper.

1.2. Data Annotation

The downloaded video clips were annotated manually for facial and speech features described in the main paper (Fig. 2 (b)). The annotations were a nested list of time intervals for each feature. For example, in a 10s video clip, if a speaker smiled during the intervals 3-5s and 7-9s, it was marked as [[3,5],[7,9]].

To ensure consistency in the video clip extraction and annotation tasks, we recruited six annotators and trained them to identify effective video clips that met the requirements. After training, each annotator independently annotated the first 10 episodes of the gameshow, and their results were compared to eliminate any inter-annotator bias. We compared the number of clips extracted and rectified any differences in eligibility criteria through discussion. We then measured the annotations for all extracted clips and realigned any discrepancies between annotators. This process was repeated until we achieved uniformity in clip selection and minimized any inter-annotator bias. The Cohen's Kappa scores for audio and visual features are reported in Table. 1.



(a) BOL - Undesired viewpoint (b) Face frames in BOL

(c) BOL - Faces missing in a frame (d) Face frames in Dolos

Figure 1. Qualitative comparison of DOLOS and Box of Lies (BOL) datasets.

Modality	**Before**	**After**
Audio	0.53	0.67
Vision	0.47	0.63
Average	0.5	**0.65**

Table 1. Cohen's Kappa scores for audio and visual feature annotations on DOLOS.

1.3. Human-level Prediction

We also compared the accuracy of the proposed method on the DOLOS dataset with that of human participants from the gameshow. We manually verified the prediction accuracy of the human participants. The deception detection accuracy of the participants in the gameshow is 41.37%.

*Equal contribution
†Corresponding author

Note that in the gameshow, the human-level prediction was the team captain's decision but not an individual prediction. The individual predictions could not be evaluated as not all the members of the team did predictions. Therefore, we can only regard this as a reference to human-level accuracy. According to Bond *et. al* [3], persons without training can distinguish deception behaviors from truthful ones with an accuracy of 54%. In this paper, the proposed method achieved an accuracy of 64.75% on DOLOS.

1.4. Comparison with Box of Lies

In this section, we qualitatively compared our DOLOS dataset with the previous gameshow dataset Box of Lies (BOL). Note that the BOL dataset was primarily annotated for verbal (text) and MUMIN facial features using the ELAN software [9] by human annotators. Directly using this dataset for multimodal deception detection with visual (face) and audio (speech) modalities faced the following challenges.

In BOL, human annotators transcribed speech and annotated facial features for each utterance interval, regardless of any constraints mentioned in Section 3.2 of the main paper. This means that even if there was background noise or other distractions, the annotators still transcribed the text and annotated the facial expressions. However, this posed a greater challenge for the visual modality due to frequent changes in the viewpoint between speakers and objects during the gameshow, as illustrated in Fig 1 (a) and (c). This resulted in either inconsistent or insufficient face frames when extracting them from the video clips, as shown in Fig 1 (b). For the cross-testing performance discussed in the main paper Section 5.3, we cleaned the BOL dataset and used only qualified samples for training in order to have a fair comparison with DOLOS.

DOLOS data collection constraints inherently prevented the above problems and provided high-quality deceptive samples as shown in Fig 1 (d). There might exist inevitable centisecond level inaccuracies in capturing video clips due to the higher frame rate and manual labeling. However, this will not affect DOLOS providing high-quality deceptive samples with clear speech audio and desired viewpoints.

2. Experiments

2.1. Data Pre-processing

Audio pre-processing. From the video clips, the raw speech audio files were extracted as .wav files, which had a sampling frequency of $44.1KHz$. Using torchaudio [10] library, the audio files were loaded as discrete audio samples and normalized to zero mean and unit variance. The pre-trained 1D convolutional feature extractor in W2V2 architecture [2] with a receptive field of $25ms$ and a stride of $20ms$ was used as the audio feature extractor. This corresponds to 321.89 discrete audio samples for every audio token. Based on these constraints, we resampled each audio file to a fixed sample length of 64×321.81, for which the convolutional feature extractor outputs precisely $L = 64$ tokens.

Visual pre-processing. By running MTCNN [11] face detector on the video clips, the face frames were directly extracted at a rate of $25fps$. For each video clip, we sampled $L = 64$ face frames, resized them to 160×160 pixels, and normalized them to ImageNet normalization statistics mean = $[0.485, 0.456, 0.406]$ and std = $[0.229, 0.224, 0.225]$.

2.2. Model Details

W2V2. We used the W2V2 model pre-trained on the Librispeech Corpus [7] as backbone. The convolutional feature extractor in W2V2 extracted features with the size of 64×512 and a linear projection layer with position embedding projected the features to tokens with a dimension of 64×768. We used the first four transformer encoder layers of W2V2.

ViT. We adopted an ImageNet pre-trained ViT and discarded the CNN patch extraction and position embedding layers since they were originally designed for image classification tasks. For facial feature extraction, we utilized a shallow 3-layer CNN with residual connections to extract the feature vectors with the size of 64×256 for $L = 64$ face images, where L was the number of frames. Using a linear projection and a position embedding layer, the face features were projected to a space with the size of 64×768. We adopted the first four encoder layers from ViT.

In deception detection, the temporal information in visual and audio features is important, as the deceptive cues are dynamically present in the videos. Previous methods usually focused on spatial information by adopting strong pre-trained networks, *e.g.*, VGG network pre-trained on face images, and then applied shallow temporal layers such as LSTM. However, this may not be sufficient as the temporal information in earlier stages of feature extraction was ignored. To better explore temporal attention, we applied a shallow CNN feature extractor for spatial features and more focused on improving temporal attention based on transformer-based networks. We also considered the efficiency of the method. With our best configuration of the proposed method, where the model had four W2V2 and ViT encoders, UT-Adapters, and four PAVF fusion modules, the total number of parameters was 71.08M and the trainable parameter was 5.06M.

2.3. Benchmarking

In this section, we describe the feature extraction methodologies for benchmarking DOLOS and provide the implementation details. The complete set of visual and au-

Visual			Audio	
Open Face [1]	Affect [6]	AU [1]	MFCC [10]	Open SMILE [4]
Eye Gaze and Pose	Happy	AU 1 - Inner brow raiser	28 Mel Filter Banks	Feature Set eGeMAPSv02
Eye Landmarks	Sad	AU 2 - Outer brow raiser	Sampling at 16 KHz	Total = 88 features
2D Facial Landmarks	Surprise	AU 4 - Brow lowerer		
Head Pose	Fear	AU 5 - Upper lid raiser		
	Disgust	AU 6 - Cheek raiser		
	Anger	AU 7 - Lid tightener		
	Contempt	AU 9 - Nose wrinkler		
		AU 10 - Upper lip raiser		
		AU 12 - Lip corner puller		
		AU 14 - Dimpler		
		AU 15 - Lip corner depressor		
		AU 17 - Chin raiser		
		AU 20 - Lip stretched		
		AU 23 - Lip tightener		
		AU 25 - Lips part		
		AU 26 - Jaw drop		
		AU 28 - Lip suck		
		AU 45 - Blink		

Table 2. Visual and audio features for DOLOS benchmarking

Features	Feature Dimensions	Projection	Model
Open Face	64x262	✓	2 layer LSTM, I/O_dim = 128, hidden_dim=64
AU	64x36	✓	2 layer LSTM, I/O_dim = 32, hidden_dim=16
Affect	64x7	✓	2 layer LSTM, I/O_dim = 8, hidden_dim=8
MUMIN	1x25		MLP, hidden_dim=16
RGB Face	64x3x160x160	✓	Resnet18 with Avg Pooling, 2 layer LSTM, I/O_dim = 128, hidden_dim=64
MFCC	Tx28	✓	2 layer LSTM, I/O_dim = 32, hidden_dim=16
Open SMILE	1x88		MLP, hidden_dim=64

Table 3. Implementation details for DOLOS benchmarking

Method	Backbone Fine-tuning Layers			
	(4)	(3,4)	(2,3,4)	(All)
ACC (%)	**59.00**	57.87	58.30	58.83
F1 (%)	66.34	**72.71**	72.43	72.11
AUC (%)	**56.78**	51.13	52.06	53.02
No. of Parameters (M)	25.658	39.838	54.018	68.198

Table 4. Comparions on fine-tuning the different number of encoder layers on the backbone networks.

dio features used to benchmark DOLOS is presented in Table 2. The OpenFace features and Action Units (AU) features were extracted using the OpenFace toolkit [1]. The facial affect (emotion) features were extracted using the Affectnet [6] model, which continuously predicted the likelihood of seven emotion categories. For audio, the Mel Cepstral Frequency Coefficients (MFCC) and OpenSMILE features were extracted by using the OpenSMILE toolkit [5]. The dimension of the features and the model implementation details are presented in Table 3. For all features, a linear classifier was used for deception detection.

2.4. Multi-task Learning

Due to limited space in the main paper, we are presenting the complete results for multitask learning in Table 6. Our findings indicate that utilizing all 25 visual-audio features in multitask learning yielded the best performance among the options evaluated. Fig. 2 displays the binary classification accuracies obtained using these 25 MUMIN features. Specifically, we observed that 11 features achieved an accuracy of over 90%, 8 features scored between 80% and 90%, while 6 features performed below 80%. On average, the 25 features yielded an accuracy of 85.78%.

2.5. Ablation Study

We conducted a preliminary ablation study to evaluate the impact of fine-tuning different numbers of encoder layers of backbone networks (ViT and W2V2) on the performance. As shown in Table 4, our observations revealed that all fine-tuned backbone models delivered unsatisfactory results despite having a large number of trainable parameters, making them less parameter-efficient. To address this issue, we proposed Parameter-Efficient Crossmodal Learning (PECL), which involves training only a few additional modules during fine-tuning. Our experimental results demonstrated that the proposed method outperformed the fine-tuning of the original backbone networks (see Section 5.5 in the main paper).

Next, we report the full ablation study on UT-Adapter positions in Table 7 and kernel size in Table 8. UT-Adapters in parallel with MHSA and FFN with a kernel size of 3 achieved the best accuracy and overall performance. As shown in Table 8, the UT-Adapter with a Kernel size of 3 delivered the best results with fewer parameters and a lower computation budget. For larger kernels, the performance

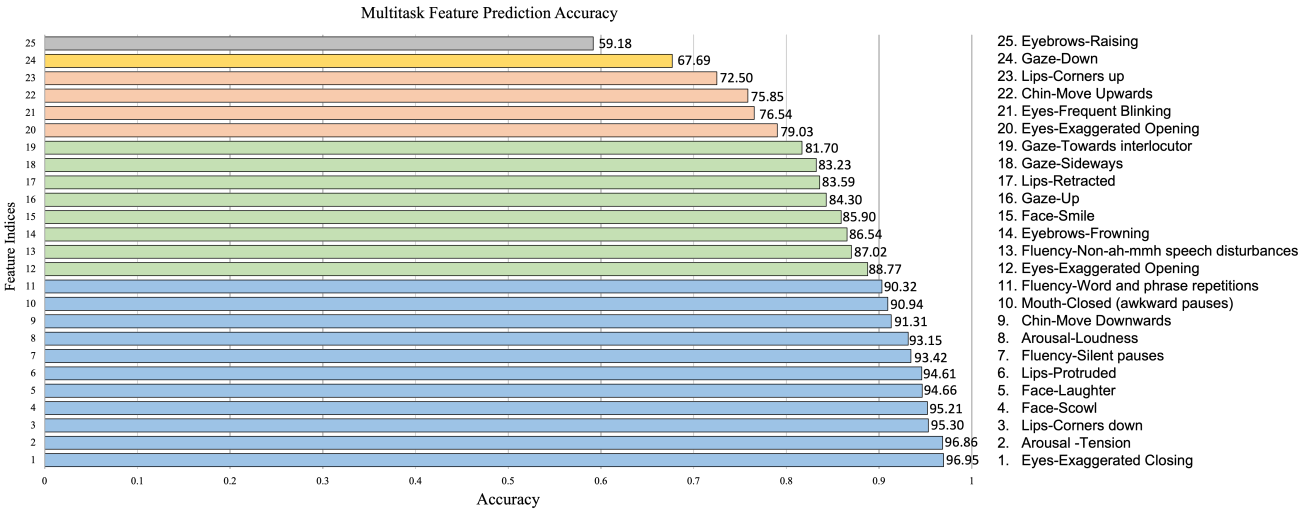


Figure 2. Multitask Learning accuracies on 25 visual-audio features.

Fusion Method	3-Fold Average			Duration Protocol			Gender Protocol		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Scaled Dot-Product Attention [8]	60.40	70.83	56.40	59.72	**71.14**	55.17	55.13	60.37	52.47
Fusion (PAVF)	**64.75**	**71.20**	**62.71**	**62.43**	70.04	**59.92**	**58.28**	**65.41**	**53.31**

Table 5. Comparsion of PAVF with Scaled Dot-Product Attention for fusion. The metrics are ACC (%), F1 (%), and AUC(%).

	w/o multitask	5A	20V	25(A+V)
ACC (%)	64.75	64.05	64.90	**66.84**
F1 (%)	71.20	71.61	70.68	**73.35**
AUC (%)	62.71	61.37	61.44	**64.58**

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

Position	‖ MHSA ‖ FFN	‖ MHSA	‖ FFN	MHSA △ FFN
ACC (%)	**64.75**	63.88	63.37	64.39
F1 (%)	**71.20**	68.72	71.14	70.44
AUC (%)	**62.71**	61.89	60.12	61.93

Table 7. Ablation study on UT-Adapter positions. ‖ indicates "in parallel with" and △ indicates "between".

Kernel Size	3	5	7	9
ACC (%)	**64.75**	64.4	64.31	64.14
F1 (%)	**71.20**	69.91	70.88	71.05
AUC (%)	62.71	**62.83**	62.22	61.51

Table 8. Ablation study on UT-Adapter kernel size.

Dimension of UT-Adapter	64	128	256
ACC (%)	60.36	**64.75**	61.29
F1 (%)	70.95	**71.20**	70.73
AUC (%)	53.83	**62.71**	55.73

Table 9. Ablation study on UT-Adapter dimensions.

dropped marginally and also at the expense of more parameters. We also conducted an ablation study on the embedding dimensions of UT-Adapter shown in Table 9, where the dimension of 128 achieved the best performance.

No. of PAVF modules	1	2	3	4
ACC (%)	60.66	61.89	63.48	**64.75**
F1 (%)	70.48	71.90	71.12	**71.20**
AUC (%)	57.03	57.39	62.22	**62.71**

Table 10. Ablation on numbers of PAVF modules.

Dimension of PAVF	128	256	512
ACC (%)	62.82	**64.75**	61.08
F1 (%)	66.32	**71.20**	69.53
AUC (%)	59.72	**62.71**	58.33

Table 11. Ablation study on PAVF correlation dimensions.

2.6. Ablation Study on Fusion

Table 5 shows the performance of the proposed PAVF module and the scaled dot-product attention proposed by Vaswani *et al.* [8] on DOLOS. We only replaced the PAVF module with the scaled dot-product attention. Specifically, the scaled dot-product attention was conducted on the concatenation of visual and audio features from each encoder layer. PAVF performed better than scaled dot-product attention in terms of accuracy and AUC on all the protocols.

We reported the full results on ablations of the number of PAVF modules in Table 10. The results demonstrated that with four PAVF modules, our model achieved the best performance. Note that the dimension of the UT-Adapter was 128.

We performed an ablation study on the fusion head in PAVF. PAVF module without the fusion head resulted in

63.48%, 71.12%, and 62.22% on the accuracy, F1, and AUC metrics, respectively. In comparison, the PAVF module with a fusion head achieved better performance. Learning the correlation between multiple modalities was one of the commonly used methods. However, PAVF was effective because it learned the crossmodal attention in a lower dimension space, which reduced the computational cost. The fusion head fused multimodal features and further reduced the dimension. PAVF was easy to be applied at any stage between unimodal learning.

3. Analysis and Discussion

We discuss a few points related to our established dataset and experiments.

Dataset collection. Deceptive gameshow provides a rich resource for deceptive samples. However, it is challenging to collect clean and high-quality data due to cinematography (camera viewpoint frequently shifts between the speakers and the host in order to captivate the viewers) and background effects (audience laughs, funny meme sounds). In comparison, these problems can be controlled and eliminated in the lab-based setting. However, the incentive for speakers may not be high. Experiences from establishing gameshow datasets can be useful for future deception detection dataset construction.

Impact of duration and gender. DOLOS provides a duration protocol and a gender protocol. The lower performance in these protocols reveals that duration and gender factors affect the multimodal deception detection accuracy. It is crucial to develop multimodal AI models that are robust to different impact factors like age, gender, ethnicity, spoken language, etc. We hope that our dataset opens up new venues for investigating these issues.

Natural language. The proposed method captured visual and audio cues to perform deception detection. It is challenging to learn useful text information and fuse multiple modalities. In the future, we will consider text modalities for the deception detection task.

References

[1] 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. 3

[2] 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. 2

[3] Charles F Bond Jr and Bella M DePaulo. Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234, 2006. 2

[4] 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*, MM '10, page 1459–1462, New York, NY, USA, 2010. Association for Computing Machinery. 3

[5] 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. 3

[6] 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. 3

[7] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015. 2

[8] 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. 4

[9] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. Elan: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006)*, pages 1556–1559, 2006. 2

[10] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi. Torchaudio: Building blocks for audio and speech processing. *arXiv preprint arXiv:2110.15018*, 2021. 2, 3

[11] 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. 2