Exploring Audio-Visual Fusion Methods in Foundation Model-Based Deception Detection

Jiaxiang Meng*, Hardik B. Sailor†, Qiongqiong Wang†, Tianchi Liu†‡, Kong Aik Lee§, and Xingmei Wang¶
*College of Software, Taiyuan University of Technology, Taiyuan, China
†Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore
‡Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore
§Dept. of Electrical and Electronic Engineering, Hong Kong Polytechnic University, Hong Kong
¶College of Computer Science and Technology, Harbin Engineering University, Harbin, China
mengjiaxiang@tyut.edu.cn, {sailor_hardik_bhupendra, wang_qiongqiong, liu_tianchi}@i2r.a-star.edu.sg, kong-aik.lee@polyu.edu.hk, wangxingmei@hrbeu.edu.cn

***Abstract*—The deception detection task aims to identify if a speaker is speaking truth or lie. It is a challenging problem due to limited training data and hence this poses a restriction for representation learning models to learn better features. Audio-visual multi-modal detection has gained significant attention for its superior performance compared to single-modality approaches. In practical scenarios, multi-modal integration can be challenging due to the distinct characteristics of each modality, making the fusion process difficult. In this paper, we employ the Querying Transformer (Q-former) to temporally align audio and visual features from foundation models and explore various methods to fuse these two modalities. Comprehensive experiments on the DOLOS dataset show that our proposed fusion technique outperforms systems based on individual audio and visual modalities. The experiments also indicate that early fusion using layer-by-layer alignment between the two modalities in the foundation model is not required. Instead, integrating over all layers of the two modalities yields the best performance.**

I. Introduction

Deception detection is to evaluate truthfulness and identify deceptive behaviors. It is a widespread and intricate phenomenon that permeates various facets of human existence [1]. Traditionally, law enforcement agencies have relied on polygraph tests as a conventional method for detecting deceptive behavior. However, this technique can be impractical, requiring direct skin contact and the expertise of trained professionals for accurate measurement and interpretation [2], [3]. Moreover, the outcomes suffer from inaccuracies and biases, and informed individuals can manipulate their physiological responses to deceive both the device and the human examiner [4].

In response to the limitations of traditional polygraph methodologies, researchers have turned to automatic deception detection through deep learning algorithms [5], [6]. The potential in this domain is vast, given the success in various complex tasks [7]–[9]. However, identifying deceptive behavior presents unique and significantly more complex challenges, mainly due to the multi-modal nature and the subtleties involved in deception. Deception detection typically incorporates audio cues (verbal, etc.) and visual cues (facial expressions, etc.). However, one significant challenge in multi-modal task is the limited availability of training data for state-of-the-art models. Foundation models are a key choice to address this issue, as they reduce the need to train models from scratch. By fine-tuning a foundation model, the required amount of data for specific tasks is significantly less than training a model from scratch [10]–[12]. Additionally, because these models are trained on diverse and extensive datasets, they can generalize well across various tasks with minimal additional training. Therefore, in this work, we explore deception detection using foundation models.

Efficient fusion methods to process and integrate these diverse audio-visual data streams to detect deception remain challenging. One is about the varying data lengths, thus resulting in the temporal length mismatch between different modalities. However, the data fusion needs the same temporal length between different modalities. Thus, in this paper, we explore various fusion techniques for audio-visual deception detection, particularly addressing the issue of temporal length mismatch between audio and visual modalities. Our contributions are as follows:

- We employ the Querying Transformer (Q-former) [13] to align audio and visual features, handling the unique characteristics of each modality from their respective foundation models.
- We explore the effectiveness of four fusion techniques and two layer aggregation methods using two audio encoders on the state-of-the-art DOLOS dataset [14]. Additionally, we demonstrate the impact of layer alignment and features from hidden layers on the performance of the system.

II. Fusion method

In practical scenarios, integrating audio-visual modality can be challenging due to their distinct characteristics, making the fusion process difficult. The variable lengths in the temporal dimension of speech and video features can result in a mismatch in the temporal dimension between the audio and visual modalities. To effectively leverage the knowledge from the audio and visual modalities, we explore different fusion techniques for audio-visual deception detection.

979-8-3315-7206-8/25/$31.00

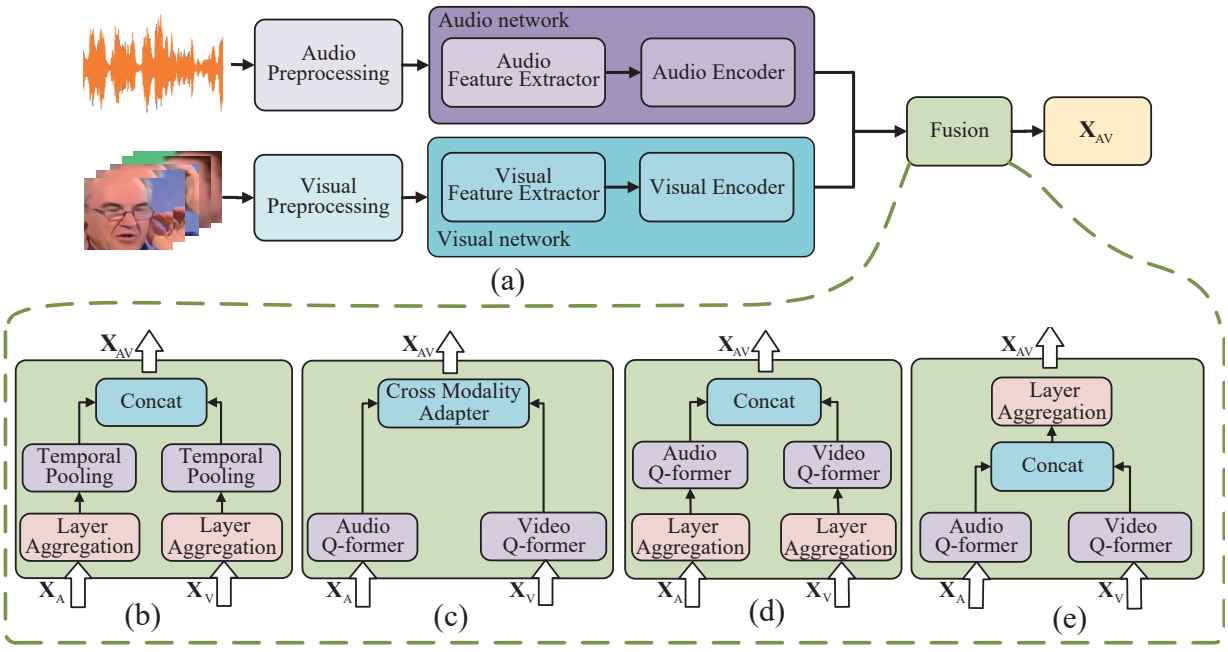

Fig. 1. The proposed audio-visual deception detection model. (a) indicates the model detail. (b)-(e) are different fusion methods in (a). $\mathbf{X}_{\mathrm{A}}$ and $\mathbf{X}_{\mathrm{V}}$ are the audio and visual embeddings from all encoder layers, respectively. $\mathbf{X}_{\mathrm{AV}}$ is fusion embeddings.

In an audio-visual deception detection system using two foundation model encoders, as shown in Fig. 1 (a), the embedding $\mathbf{X}_{\mathrm{A}}^{i} \in \mathbb{R}^{L_A \times D_A}$ and $\mathbf{X}_{\mathrm{V}}^{j} \in \mathbb{R}^{L_V \times \mathrm{D}_V}$ are extracted by the layer i of the transformer-based encoder, where $i \in [1, N_{\mathrm{A}}]$ and $j \in [1, N_{\mathrm{V}}]$. Here, N_A and N_{V} represent the number of layers in the audio encoder and visual encoder. L_A and L_V indicate the temporal dimensions of the audio and visual embeddings, while D_A and D_V denote embedding dimensions. Therefore, audio embeddings $\mathbf{X}_A \in \mathbb{R}^{N_A \times L_A \times D_A}$ and visual embeddings $\mathbf{X}_V \in \mathbb{R}^{N_V \times L_V \times D_V}$ are extracted for the fusion.

To effectively leverage the knowledge from the audio and visual modalities, we investigate: (1) the use of Q-former [13], [15], [16] for the alignment between the audio and visual modalities in the temporal dimension, (2) layer aggregation to merge information from these diverse layers, and (3) Cross modality adapter for layer-wise modality fusion.

A. *Temporal alignment by Q-former*

The variable-length embeddings always cause the unalignment of temporal dimension, resulting in the fusion difficulty. Thus, temporal pooling is traditionally adopted to integrate and maintain the formation over the temporal dimension i.e. convert the variable-length embeddings $\mathbf{X} \in \mathbb{R}^{N \times L \times D}$ to into the fixed-length embeddings $\mathbf{X} \in \mathbb{R}^{N \times 1 \times D}$. However, it can also lead to a loss of specific details, especially fine-grained temporal details. This can be a challenge for tasks requiring timing information such as deception detection.

The Q-former, originally proposed for visual-text modality alignment, typically employs a Transformer-based module to convert the encoder output into a fixed number of textual input tokens [13]. With appropriate modifications, it can also process variable-length audio signals. The key is to employ trainable queries $\mathbf{Q} \in \mathbb{R}^{L \times D}$ in each block to interact with the variable-length signal by cross-attention layer. In this paper, we try to transform the variable-length video and audio data into the fixed query representations that fit the input of the fusion module like SALMONN[15], [16]. They can also extract the most relevant information from audio and video modalities and align them with the fusion input space.

In this work, the Q-Former in each modality, named Video Q-Former and Audio Q-Former in Fig. 1, is used to achieve temporal alignment between the audio and visual modalities. Based on the locations where the Q-former is placed in the network, our techniques is summarized as follows:

- Q-former before layer aggregation. The embeddings of two modalities are fixed into the same temporal dimension by the Q-former, and layer aggregation is adopted to fuse the embedding of audio and visual modalities from all encoder layers.
- Q-former after layer aggregation. The stack of all the embeddings in each modality is first fused by layer aggregation, and then the final embedding is projected as the fixed temporal dimension by layer aggregation.
- Cross-modality adapter. This technique employs the cross-correlation to fuse the audio and visual modalities as the fusion embedding. Before this, the embedding is only aligned between audio and visual modalities.

B. *Layer aggregation techniques*

Both the audio and visual foundation models can capture varying levels of abstraction at the different layers. We aim to implement layer aggregation to merge information from

these diverse layers. For a foundation model that consists of N layers, embeddings $\mathbf{X} \in \mathbb{R}^{N \times L \times D}$ are extracted. We assume that the N layers can be aggregated by

$$\mathbf{Y} = \mathbf{A} \otimes \mathbf{X} \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{L \times D}$ represents the layer-aggregation representation, $\mathbf{A}$ is a row vector of dimension $[1, N]$, and $\otimes$ denotes the dot product operation. We investigate two approaches to derive the weights $\mathbf{A}$ for the layer aggregation: (1) Weighted Sum, and (2) Attentive Pooling.

Weighted Sum (WS) is a straightforward layer-aggregation technique that initiates a set of learnable weights to each layer's embedding and then sums them up. It allows the model to assign different importance to the embeddings of different layers based on their relevance or contribution to the task. It preserves the global information across the encoder layers. These weights $\mathbf{A}_{\text{WS}}$ are learnable parameters that are optimized during the training process.

Attentive Pooling (AP) has been widely adopted in speaker verification to obtain utterance-level representations by aggregating information in the temporal domain [17]–[19]. Inspired by this, we introduce it as a layer aggregation method to capture information from various layers. It uses an attention mechanism to dynamically assign weights to the embeddings of different layers based on their importance or relevance to the task. It is done as follows:

$$\mathbf{A}^{\text{AP}} = \text{Softmax}\left(\text{Conv}\left(\text{Tanh}\left(\text{Conv}\left(\mathbf{X}\right)\right)\right)\right) \tag{2}$$

The above-mentioned two approaches all assign weights to aggregate the output embeddings from different layers based on a similar underlying concept. The key difference lies in how the weights are determined. Specifically, Weighted Sum uses fixed weights, while Attentive Pooling and Attentive Merging employ different attention mechanisms to dynamically assign weights based on the input and task conditions.

We investigate the use of layer aggregation within each modality and across both modalities. For within-modality aggregation, the embeddings $\mathbf{X}_{\text{A}}$ from the audio modality and $\mathbf{X}_{\text{V}}$ from the video modality are individually aggregated into $\mathbf{Y}_{\text{A}} \in \mathbb{R}^{L_{\text{A}} \times D_{\text{A}}}$ and $\mathbf{Y}_{\text{V}} \in \mathbb{R}^{L_{\text{V}} \times D_{\text{V}}}$, respectively. This process is illustrated in Fig. 1 (b) and (d). For cross-modality aggregation, audio and visual Q-formers are applied to ensure that both modalities have a fixed temporal dimension L. In the scenario, where the encoder outputs from the foundation models have the same feature dimensions, layer aggregation can be further applied over $N_{\text{A}} + N_{\text{V}}$ layers from both modalities, as shown in Fig. 1 (e). The resulting aggregated layer representation has dimensions $\mathbb{R}^{L \times D}$.

C. Cross modality adapter

In this work, we also investigate the cross-modality adapter (CMA) which is applied to the intermediate layers of both audio and video foundation models for layer-by-layer alignment, as done in [14]. This requires the embeddings from both modalities to have the same temporal length L. Therefore, we apply CMA following Q-formers at each layer, as shown in Fig. 1 (c), where the projected audio embeddings $\hat{\mathbf{X}}^i_{\text{A}} \in \mathbb{R}^{L \times D_{\text{A}}}$ and visual embeddings $\hat{\mathbf{X}}^i_{\text{V}} \in \mathbb{R}^{L \times D_{\text{V}}}$ are then used for further fusion and understanding of audio-visual content. The final classification head includes one linear layer followed by Layer-Normalization and ReLU non-linearity. The output scores are computed by averaging all frame-level scores.

III. Experiments

A. Dataset

Inspired by the conclusion from [14], DOLOS is chosen for audio-visual deception detection over other datasets like BOL to validate the efficacy of the proposed fusion techniques [20]–[24]. At the time of our experiments, 1628 videos are available, including 885 deceptive samples and 743 truthful samples. In all experiments, we employed 5-fold cross-validation to ensure a robust evaluation. Each fold included different speakers in the training and testing sets, guaranteeing that the evaluation data exclusively consisted of unseen speakers. For training, randomly selected 4 seconds of the video are used dynamically. During evaluation, videos are segmented into 4-second intervals with a 2-second overlap. The final score for each video is obtained by averaging the scores across all segments. For the audio modality, the raw speech audio is resampled to a 16k sampling rate.

B. Model training and evaluation

For generalization, both Wav2Vec2 [10] and WavLM [11] foundation models are utilized for the audio modality. The raw audios are tokenized using a 1D-CNN module. For the visual modality, 25 images per second are sampled, with face areas cropped by MTCNN face detector [25]. These images are further normalized and resized to 160×160 pixels. We utilize the pre-trained ViT [12] as the visual backbone network. Face images are tokenized by a 2D-CNN module. Then the visual and audio backbone networks project the above visual and audio tokens as $L_{\text{V}} \times 768$ and $L_{\text{A}} \times 768$. The UT-Adapter [14] are also inserted into Wav2Vec2 and WavLM. Then, the audio and video Q-former project the variable temporal dimension L_{A} and L_{V} into fixed one L=64, thus capturing meaningful audio-visual events (such as lip movements) without introducing excessive noise or computational overhead. The batch size is 4, and the step of Gradient Accumulation Optimization is adopted as 4. Accuracy, F1-score, and Area Under the Curve (AUC) are the metrics used for model evaluation.

C. Experimental results analysis

We first compare single-modality (audio or visual) and audio-visual deception detection, as presented in Table I. The results demonstrate that the proposed audio-visual fusion techniques enhance performance by utilizing the information from both modalities. The performance improvements are consistent for both Wav2Vec2 and WavLM models. Furthermore, compared to using only the last layer, systems that aggregate embeddings from all layers like WS and AP achieve better performance in most cases. For example, there are 3.78% ACC

TABLE I. Performance of deception detection using single audio, single visual, and audio-visual multimodal fusion approaches. 'WS' and 'AP' following 12 or 24 indicate the operations of weighted sum and attentive pooling applied to 12 or 24 layers, respectively. Underscore indicates the second-best result.

		Feature extraction	ACC	F1	AUC
Audio	Wav2Vec2	Last Layer	$49.08_{\pm 9.01}$	$48.00_{\pm 24.32}$	$52.15_{\pm 2.84}$
		WS12	$52.86_{\pm 7.26}$	$59.20_{\pm 10.14}$	$51.90_{\pm 3.88}$
		AP12	$48.59_{\pm 8.96}$	$43.73_{\pm 25.22}$	$51.91_{\pm 2.91}$
	WavLM	Last Layer	$\underline{60.30}_{\pm 2.85}$	$54.64_{\pm 20.37}$	$56.33_{\pm 2.12}$
		WS12	$58.90_{\pm 4.25}$	$58.10_{\pm 14.64}$	$56.01_{\pm 2.40}$
		AP12	$59.88_{\pm 1.71}$	$59.09_{\pm 16.03}$	$56.46_{\pm 2.25}$
Visual	ViT	Last Layer	$57.91_{\pm 5.85}$	$56.17_{\pm 19.48}$	$53.81_{\pm 2.83}$
		WS12	$59.63_{\pm 4.79}$	$54.72_{\pm 28.97}$	$55.45_{\pm 3.71}$
		AP12	$59.56_{\pm 4.80}$	$57.20_{\pm 18.89}$	$56.63_{\pm 4.25}$
Fusion	Wav2Vec2&ViT	Q-former+WS24	$58.86_{\pm 3.01}$	$\mathbf{63.21}_{\pm 9.56}$	$56.38_{\pm 3.82}$
		Q-former+AP24	$59.10_{\pm 4.33}$	$\underline{62.12}_{\pm 10.92}$	$56.69_{\pm 3.97}$
	WavLM&ViT	Q-former+WS24	$\mathbf{60.93}_{\pm 2.87}$	$59.32_{\pm 16.31}$	$\mathbf{59.62}_{\pm 2.60}$
		Q-former+AP24	$59.47_{\pm 1.25}$	$56.70_{\pm 12.58}$	$\underline{57.62}_{\pm 2.87}$

TABLE II. Performance of audio-visual deception detection on different fusion techniques. Underscore indicates the second-best result.

Fusion method	Wav2Vec2&ViT			WavLM&ViT		
	ACC	F1	AUC	ACC	F1	AUC
Q-former+CMA Fig. 1(c)	$\underline{61.47}_{\pm 3.00}$	$55.59_{\pm 19.21}$	$\underline{58.02}_{\pm 3.77}$	$60.34_{\pm 3.31}$	$58.09_{\pm 10.36}$	$58.53_{\pm 1.93}$
Q-former+WS24 Fig. 1(e)	$58.86_{\pm 3.01}$	$\mathbf{63.21}_{\pm 9.56}$	$56.38_{\pm 3.82}$	$\mathbf{60.93}_{\pm 2.87}$	$59.32_{\pm 16.31}$	$\mathbf{59.62}_{\pm 2.60}$
Q-former+AP24 Fig. 1(e)	$59.10_{\pm 4.33}$	$62.12_{\pm 10.92}$	$56.69_{\pm 3.97}$	$59.47_{\pm 1.25}$	$56.70_{\pm 12.58}$	$57.62_{\pm 2.87}$
WS12+Pooling Fig. 1(b)	$58.41_{\pm 3.27}$	$58.46_{\pm 10.19}$	$55.40_{\pm 4.11}$	$\underline{60.58}_{\pm 2.95}$	$60.41_{\pm 10.91}$	$\underline{58.61}_{\pm 2.14}$
WS12+Q-former Fig. 1(d)	$\mathbf{61.52}_{\pm 2.87}$	$\underline{60.46}_{\pm 9.49}$	$\mathbf{59.70}_{\pm 2.20}$	$60.04_{\pm 2.87}$	$\underline{60.60}_{\pm 12.45}$	$57.06_{\pm 1.65}$
AP12+Q-former Fig. 1(d)	$58.80_{\pm 3.86}$	$49.45_{\pm 28.97}$	$53.98_{\pm 3.32}$	$59.62_{\pm 3.35}$	$55.18_{\pm 14.41}$	$57.07_{\pm 1.97}$

improvement for the WS12 of Wav2Vec2. This demonstrates the benefits of incorporating information from all layers.

Furthermore, Table. II conducts the experiments on temporal pooling with layer aggregation, Q-former before layer aggregation, Q-former after layer aggregation, and cross-modality adapter. It is first observed that the fusion techniques perform differently with different audio encoders. Although CMA performs second-best in ACC, our layer aggregation techniques after Q-former perform better in F1 and AUC. For example, WS outperforms CMA by 1.68% AUC. The better average values demonstrate that WS can benefit deception detection. In other words, WS has the lowest standard deviation of 9.75% in Wav2Vec2. Meanwhile, according to the average results, especially F1, the layer aggregation techniques after Q-former are more suitable for Wav2Vec2. Overall, WavLM performs better as a whole, which has higher average values. Besides, Q-former is generally better than temporal pooling.

To better understand why WS-based approaches perform better, Fig. 2 visualizes the weights obtained from WS. The hidden layers within the audio encoder emphasize auditory features such as pitch, tone, and spectral features, while the hidden layers of the video encoder focus on visual cues like facial expressions, gestures, and body language. Weight analysis in Fig. 2 shows that audio and visual encoder layers contribute differently. Given the inherently different types of information captured by the audio and video encoders, it is unnecessary to enforce a layer-by-layer alignment between these modalities for effective multi-modal fusion. Instead, we leverage the complementary nature of the encoded features in all layers, independently of each modality, allowing for a more flexible and holistic integration. This approach enhances the robustness and accuracy of our deception detection system by utilizing the full spectrum of available information.

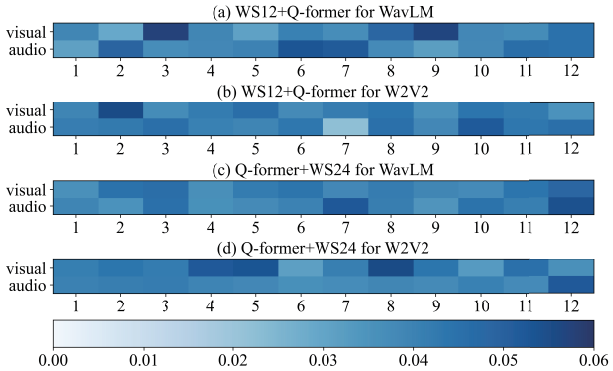

Fig. 2. The visualisation of weights in WS12+Q-former and Q-former+WS24. The weights in (a) and (b) are multiplied by 0.5 for better visualization since they are normalized with 12 layers not 24 in (c) and (d).

IV. Conclusion

We explore various fusion techniques for audio-visual deception detection based on foundation models. We employ Q-former to temporally align audio and visual features, addressing the challenge of mismatched temporal lengths in two-modality fusion. We investigate Weighted Sum, and Attentive Pooling layer aggregation approaches to capture diverse information from different layers in the foundation models. Experimental results confirm the advantage of multi-modality fusion over single modality approaches, with the combination of the Q-former and Weighted Sum layer aggregation achieving the best performance. In the future, more common fusion way is needed for original variable-length data adaptively.

Acknowledgment

This work was supported by the China Scholarship Council program (Project ID:202306680025).

References

[1] A. S. Constâncio, D. F. Tsunoda, H. d. F. N. Silva, J. M. d. Silveira, and D. R. Carvalho, "Deception detection with machine learning: A systematic review and statistical analysis," *Plos one*, vol. 18, no. 2, e0281323, 2023.

[2] T. R. Levine, "Truth-default theory and the psychology of lying and deception detection," *Current Opinion in Psychology*, vol. 47, p. 101 380, 2022.

[3] B. Kleinberg and B. Verschuere, "How humans impair automated deception detection performance," *Acta psychologica*, vol. 213, p. 103 250, 2021.

[4] H. Alaskar, Z. Sbaı¨, W. Khan, A. Hussain, and A. Alrawais, "Intelligent techniques for deception detection: A survey and critical study," *Soft Computing*, vol. 27, no. 7, pp. 3581–3600, 2023.

[5] S. A. Prome, N. A. Ragavan, M. R. Islam, D. Asirvatham, and A. J. Jegathesan, "Deception detection using machine learning (ml) and deep learning (dl) techniques: A systematic review," *Natural Language Processing Journal*, p. 100 057, 2024.

[6] W. Khan, K. Crockett, J. O'Shea, A. Hussain, and B. M. Khan, "Deception in the eyes of deceiver: A computer vision and machine learning based automated deception detection," *Expert Systems with Applications*, vol. 169, p. 114 341, 2021.

[7] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN computer science*, vol. 2, no. 6, p. 420, 2021.

[8] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," pp. 8748–8763, 2021.

[9] Y. Jin, G. Hu, H. Chen, D. Miao, L. Hu, and C. Zhao, "Cross-modal distillation for speaker recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 12 977–12 985.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[11] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[13] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*, PMLR, 2023, pp. 19 730–19 742.

[14] X. Guo, N. M. Selvaraj, Z. Yu, A. W.-K. Kong, B. Shen, and A. Kot, "Audio-visual deception detection: Dolos dataset and parameter-efficient crossmodal learning," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 135–22 145.

[15] G. Sun, W. Yu, C. Tang, *et al.*, "Video-SALMONN: Speech-enhanced audio-visual large language models," in *Forty-first International Conference on Machine Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=nYsh5GFIqX.

[16] C. Tang, W. Yu, G. Sun, *et al.*, "SALMONN: Towards generic hearing abilities for large language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=14rn7HpKVk.

[17] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *INTERSPEECH*, vol. 2018–September, 2018, pp. 2252–2256.

[18] Y. Wu, C. Guo, H. Gao, X. Hou, and J. Xu, "Vector-based attentive pooling for text-independent speaker verification.," in *Interspeech*, 2020, pp. 936–940.

[19] T. Liu, K. A. Lee, Q. Wang, and H. Li, "Disentangling voice and content with self-supervision for speaker recognition," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 50 221–50 236.

[20] F. Soldner, V. Pérez-Rosas, and R. 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)*, 2019, pp. 1768–1777.

[21] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, "Bag-of-lies: A multimodal dataset for deception detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[22] E. P. Lloyd, J. C. Deska, K. Hugenberg, A. R. McConnell, B. T. Humphrey, and J. W. Kunstman, "Miami university deception detection database," *Behavior research methods*, vol. 51, pp. 429–439, 2019.

[23] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 59–66.

[24] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. Linton, and M. Burzo, "Verbal and nonverbal clues for real-life deception detection," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2336–2346.

[25] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.