Uncertainty-driven Multi-View Feature Fusion for Spoofed Speech Detection

Xinxin Luo[1], Liangbo Zhu[2], Zhongyu He[3], Xiaolong Wu[4,1*]
[1]Xinjiang University, China
[2]Beijing University of Technology, China
[3]Shandong Jiaotong University, China
[4]Naval Aviation University, China
luoxx@stu.xju.edu.cn, zhuliangbo@emails.bjut.edu.cn,
24221007@stu.sdjtu.edu.cn, wuxl@stu.xju.edu.cn

***Abstract*—Robust spoofed speech detection hinges on generalization to unseen spoofing attacks and adverse acoustic conditions (e.g., noise and channel mismatch), where the quality of different representations can vary substantially across utterances. We propose an uncertainty-driven multi-view framework that fuses raw-waveform and LFCC-based spectral representations by estimating per-view evidential uncertainty with a Beta distribution, converting it into utterance-level credibility weights, and using these weights to condition a lightweight multi-head self-attention fusion module; we further introduce a conflict-aware objective to explicitly discourage inconsistent view opinions. Experiments on ASVspoof 2019 LA and cross-dataset evaluation on ASVspoof 2021 LA show consistent improvements over strong single-view baselines and fixed fusion strategies, and controlled SNR corruption tests further confirm that uncertainty-guided credibility weighting yields more stable performance under severe noise.**

***Index Terms*—spoofed speech detection, multi-view feature fusion, uncertainty estimation, evidential deep learning, Beta distribution**

I. Introduction

Voice has become one of the most widely adopted interfaces for human–computer interaction in the era of artificial intelligence. Automatic Speaker Verification (ASV) systems rely on speaker-specific voiceprint characteristics to secure access to smart devices and online services. However, recent advances in text-to-speech (TTS), voice conversion (VC), and replay technologies have enabled attackers to generate highly realistic spoofed speech, posing a serious threat to the security of ASV systems [1]. Spoofed speech detection aims to distinguish bona fide speech from spoofed speech, thereby preventing unauthorized access [2].

A key challenge in practical deployments is generalization: models trained on limited attack types often degrade when faced with unseen spoofing algorithms or adverse acoustic conditions. Many existing approaches extract features from a single representation of the speech signal and train a detector on top of that representation [3]–[5]. Although effective in controlled settings, single-view detectors may miss complementary spoofing cues that are present in other representations of the same utterance. In this work, we consider two complementary views that are widely used in anti-spoofing systems: (i) raw waveforms, which preserve fine-grained temporal structure, and (ii) frequency-domain features, which highlight spectral artifacts introduced by generation algorithms [6].

Recent end-to-end anti-spoofing systems further confirm the value of exploiting complementary cues from multiple representations, including dual-branch and raw-waveform-oriented architectures [7]–[9].

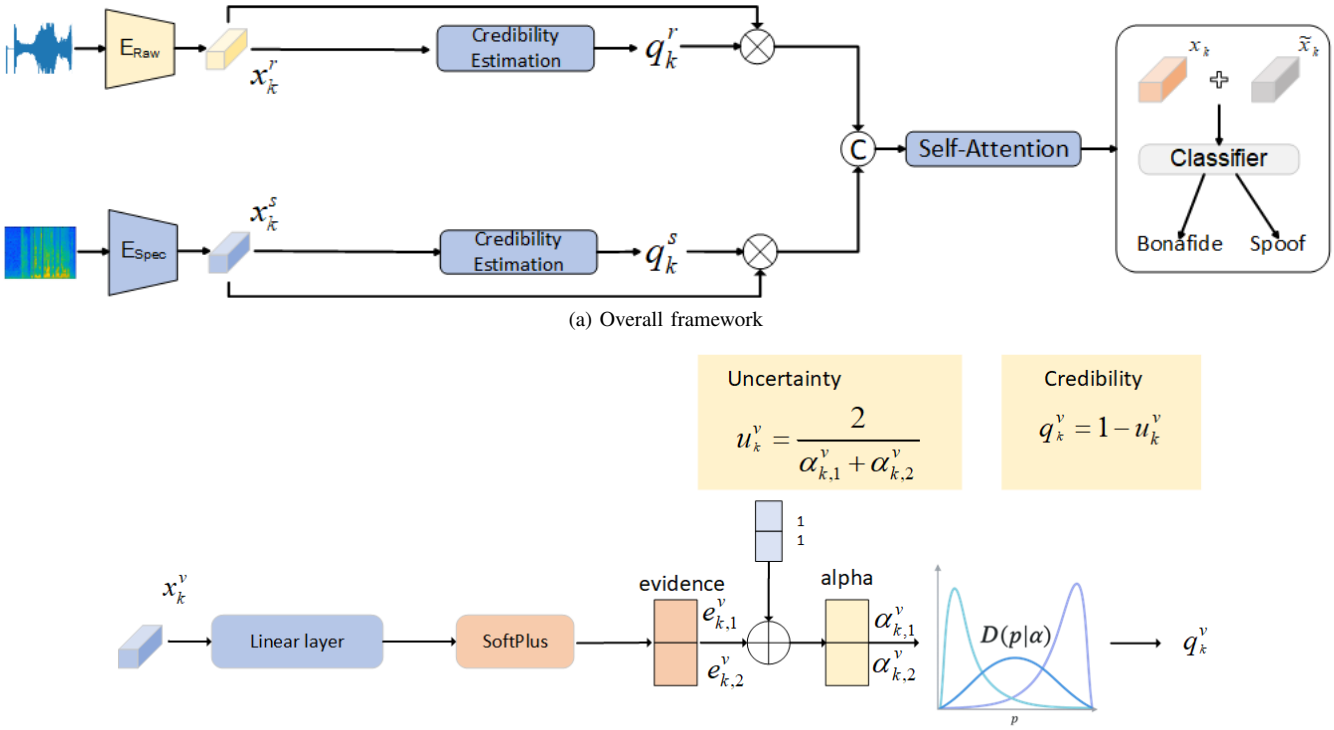
Multi-view fusion can exploit complementary cues, but fixed aggregation often treats all views equally and may amplify corrupted representations [10], [11]. We therefore propose an uncertainty-driven multi-view spoofed speech detector that estimates the credibility of raw-waveform and LFCC views, uses these credibility scores to reweight the views, and performs lightweight self-attention fusion. Experiments on ASVspoof 2019 LA and cross-dataset evaluation on ASVspoof 2021 LA show improved accuracy and better robustness under additive noise [12].

The key motivation is that view quality is inherently utterance-dependent. Under some conditions, waveform features retain useful temporal cues, whereas under others LFCC features reveal synthesis artifacts more clearly. A practical fusion model should therefore decide which view deserves more trust for each input rather than applying one fixed aggregation rule to all samples.

In summary, the main contributions of our paper are as follows:

- We propose a multi-view spoofed speech detection framework that jointly leverages raw-waveform and frequency-domain representations.
- We introduce an uncertainty-driven dynamic fusion strategy that estimates view credibility and adaptively reweights different views to mitigate noise accumulation.
- We validate the proposed method on public benchmarks and demonstrate improved performance over strong baselines and existing fusion approaches.

* Corresponding authors.



(a) Overall framework

(b) Credibility estimation block

Fig. 1. Overall structure of proposed method: (a) overall framework; (b) credibility estimation block.

II. Related Work

A. Single-view spoofed speech detection

Many spoofed speech detectors rely on a single representation of the input utterance. Spectral-feature-based systems typically operate on cepstral or spectrogram-like inputs and use CNN backbones to capture spoofing artifacts [6], [14], [31], while waveform-based systems directly model temporal cues from the raw signal using architectures such as RawNet2 and its variants [9], [25], [30]. These approaches have achieved strong performance on standard benchmarks, but their robustness often depends on how well the selected view matches the acoustic condition and spoofing pattern at test time.

B. Fusion and uncertainty-aware learning

To exploit complementary information, recent fusion-based systems combine spectral, prosodic, and waveform cues through concatenation or heuristic aggregation [7], [8], [13]. Such fusion strategies improve robustness, but fixed weighting may over-emphasize noisy views and underuse cleaner ones. This limitation becomes especially visible when one representation is more vulnerable to channel mismatch or additive noise than another.

Uncertainty-aware multi-view learning addresses this issue by estimating evidence reliability and dynamically weighting each view [19]–[21]. In spoofed or deepfake speech detection, uncertainty estimation and robust end-to-end modeling have also been explored in anti-spoofing backbones such as AASIST and RawNet-style systems [22]–[25]. Related work has also examined robustness from other perspectives, such as manipulation attacks, adversarial threats, privacy-preserving detection, and cross-bona-fide generalization [26]–[29]. Our work combines these directions by using uncertainty explicitly as a view-credibility signal for fusion.

III. Proposed Method

Fig. 1 illustrates the proposed uncertainty-driven multi-view spoofed speech detection framework. As shown in Fig. 1(a), the overall pipeline consists of two view-specific encoders and a fusion/classification head. The raw-waveform encoder E_{raw} extracts a temporal representation from the waveform, while the frequency-domain encoder E_{spec} extracts a spectral representation from LFCC features; together they provide complementary cues.

As shown in Fig. 1(b), we introduce a dedicated credibility estimation block to quantify the reliability of each view. Following [21], we model the evidence of each view with a Beta distribution and derive an uncertainty score, which is then converted into a credibility weight for feature fusion. Using these credibility weights, we perform uncertainty-guided fusion to suppress unreliable view representations. We further employ a lightweight self-attention module to facilitate information interaction across views and enhance integration.

Finally, the fused representation is fed into a classifier to determine whether an utterance is bona fide or spoofed.

A. Multi-view modeling layer

Raw waveform encoder: RawNet2 has demonstrated excellent performance in processing raw speech waveform signals [30]. Given an input utterance (sample) indexed by k, we denote the raw waveform as a_k^r and the extracted LFCC feature as a_k^s. We use RawNet2 to extract features from the raw waveform:

$$x_k^r = E_{raw}(a_k^r). \tag{1}$$

Speech frequency-domain encoder: ResNet-18 has shown good performance in processing speech frequency-domain features [31]. In this paper, we use ResNet-18 to encode the frequency-domain features of speech. Specifically, we first extract the Linear Frequency Cepstral Coefficients (LFCC) [32] features a_k^s from the raw speech waveform signal a_k^r. LFCC compresses the information in the spectrogram and removes redundancies, effectively reducing dimensionality and improving computational efficiency. Therefore, in this paper, we treat LFCC as a low-dimensional representation of the spectrogram and feed it into ResNet-18 to extract spoofing cues from the frequency domain perspective, i.e.,

$$x_k^s = E_{spec}(a_k^s). \tag{2}$$

B. Single view credibility calculation

In various spoofed speech detection scenarios, the credibility of different feature views (waveform features and spectral features) can vary considerably. By estimating and combining the credibility of each view, detection performance can be improved.

TMC [33] shows that evidential distributions can quantify view reliability. Following [21], and grounded in subjective logic [34], we adapt this idea to binary spoofed speech detection. For sample k and view $v \in \{r, s\}$, let $x_{k,c}^v$ denote the logit for class $c \in \{1, 2\}$. We map logits to non-negative evidence via

$$e_{k,c}^v = \mathrm{Softplus}(x_{k,c}^v), \tag{3}$$

and parameterize a Beta distribution by

$$\alpha_{k,c}^v = 1 + e_{k,c}^v, \tag{4}$$

Let $S_k^v = \alpha_{k,1}^v + \alpha_{k,2}^v$ denote the strength of the Beta distribution. Under the subjective logic constraint for binary classification, we define the belief masses and uncertainty as

$$b_{k,c}^v = \frac{e_{k,c}^v}{S_k^v}, \qquad u_k^v = \frac{2}{S_k^v}, \tag{5}$$

which satisfies $b_{k,1}^v + b_{k,2}^v + u_k^v = 1$ for each view v.

Based on the estimated uncertainty, we first define the unnormalized credibility score of each view as

$$q_k^v = 1 - u_k^v, \quad v \in \{r, s\}. \tag{6}$$

This formulation assigns lower credibility to views with higher uncertainty.

We convert uncertainty into normalized reliability weights by

$$\bar{q}_k^v = \frac{q_k^v}{q_k^r + q_k^s}, \quad v \in \{r, s\}. \tag{7}$$

where $\bar{q}_k^r + \bar{q}_k^s = 1$ for each sample k.

C. Multi-view Fusion

Similarly, for multi-view fusion, we refer to the method proposed in [21], with the key distinction that our approach explicitly considers the impact of opinion conflicts between different views.

First, after obtaining per-view credibility estimates, we employ a Multi-Head Self-Attention (MHSA) module to fuse the encoder features conditioned on their credibility scores. This design enables information exchange across views while suppressing unreliable representations. Given the view-reliability weights, we first perform view-wise scaling:

$$\hat{x}_k^v = \bar{q}_k^v \, x_k^v, \quad v \in \{r, s\}. \tag{8}$$

Then, we treat the two scaled view features as a length-2 token sequence and apply a Multi-Head Self-Attention (MHSA) module. We use mean pooling over the two output tokens to obtain an utterance-level fused vector:

$$x_k = \mathrm{MeanPool}(\mathrm{MHSA}([\hat{x}_k^r, \hat{x}_k^s])). \tag{9}$$

To preserve structural interaction information between the two views, we additionally keep the token-level MHSA output matrix

$$M_k = \mathrm{MHSA}([\hat{x}_k^r, \hat{x}_k^s]), \tag{10}$$

and convert it into a compact auxiliary feature by flattening and dimension reduction:

$$\tilde{x}_k = DownSample(Flatten(M_k)). \tag{11}$$

The pooled vector x_k captures the global fused representation, while $\tilde{x}_k$ preserves finer cross-view interaction patterns produced by the attention module.

When addressing multi-view opinion conflicts, the uncertainty of multi-view learning results should not automatically decrease with an increase in the number of views. Instead, it should depend on the quality of the fused views, particularly when conflicts arise among the outputs of different views. To tackle this issue, we introduce a conflict opinion aggregation method [20].

In [20], the conflictive degree is defined to quantify disagreement between two views. In our notation, for sample k and view $v \in \{r, s\}$, we define the view opinion as $\mathbf{w}_k^v = (\mathbf{b}_k^v, u_k^v)$, where $\mathbf{b}_k^v = [b_{k,1}^v, b_{k,2}^v]$ is the belief mass vector and u_k^v is the corresponding uncertainty. In our two-view setting, we compute the conflictive degree between the raw view (r) and the spectral view (s) as:

$$c(\mathbf{w}_k^r, \mathbf{w}_k^s) = c_p(\mathbf{w}_k^r, \mathbf{w}_k^s) \cdot c_c(\mathbf{w}_k^r, \mathbf{w}_k^s), \tag{12}$$

where $c_p(\mathbf{w}_k^r, \mathbf{w}_k^s)$ represents the projected distance between $\mathbf{w}_k^r$ and $\mathbf{w}_k^s$, and $c_c(\mathbf{w}_k^r, \mathbf{w}_k^s)$ represents their conjunctive certainty.

The formulas for these components are as follows:

$$c_p(\mathbf{w}_k^r, \mathbf{w}_k^s) = \frac{1}{2} \sum_{c\in\{1,2\}} \left| b_{k,c}^r - b_{k,c}^s \right|, \tag{13}$$

$$c_c(\mathbf{w}_k^r, \mathbf{w}_k^s) = (1 - u_k^r)(1 - u_k^s). \tag{14}$$

By quantifying the degree of divergence between two opinions, the conflictive degree effectively resolves conflicts during multi-view feature fusion, thus enhancing the accuracy and robustness of the detection system.

D. Loss Function

After fusion, we apply a linear layer to obtain the logits z_k:

$$z_k = W_o \cdot [x_k, \tilde{x}_k] + b_o, \tag{15}$$

The predicted posterior is computed by

$$p_k = \mathrm{Softmax}(z_k), \tag{16}$$

and we use the standard cross-entropy loss $\mathcal{L}_{ce}(z_k, \hat{y}_k)$.

During training, we follow evidential learning [35] and introduce a credibility (uncertainty) loss $\mathcal{L}_u$ for each view to encourage calibrated evidence estimation:

$$\mathcal{L}_u(\mathbf{x}_k^v) = \sum_{c\in\{1,2\}} \hat{y}_{k,c} \cdot \Big(\psi\left(S_k^v\right) - \psi\left(\alpha_{k,c}^v\right) \Big), \tag{17}$$

where $\mathbf{x}_k^v = [x_{k,1}^v, x_{k,2}^v]$ collects the two class logits for view v, $\hat{y}_{k,c}$ is the one-hot label for class c, and $\psi(\cdot)$ denotes the digamma function.

To ensure consistency among the results of different views during training, we minimize the conflict between views. In our two-view setting ($v \in \{r, s\}$), the conflict loss for sample k is defined as follows:

$$\mathcal{L}_{c,k} = c(\mathbf{w}_k^r, \mathbf{w}_k^s). \tag{18}$$

The overall loss function can be expressed as:

$$\begin{aligned} \mathcal{L} = &\sum_{k\in\mathcal{Y}} \mathcal{L}_{ce}(z_k, \hat{y}_k) + \lambda_1 \sum_{k\in\mathcal{Y}} \sum_{v\in\{r,s\}} \mathcal{L}_u(\mathbf{x}_k^v) \\ &+ \lambda_2 \sum_{k\in\mathcal{Y}} \mathcal{L}_{c,k}, \end{aligned} \tag{19}$$

where λ_1 and λ_2 are hyperparameters used to balance these components, and $\mathcal{Y}$ denotes the index set of training samples (e.g., a mini-batch) used to compute the loss. The three terms play complementary roles during optimization: $\mathcal{L}_{ce}$ supervises spoof/bona fide discrimination in the fused space, $\mathcal{L}_u$ encourages each view to produce calibrated evidence, and $\mathcal{L}_{c,k}$ discourages confident but inconsistent opinions across views. Jointly optimizing these losses enables the model to learn not only discriminative representations, but also a more reliable and stable fusion behavior.

TABLE I
THE DATASET PARTITION OF ASVSPOOF 2019 LA.

Set	#Samples	Attacks
Train	25,380	A01–A06
Development	24,844	A01–A06
Evaluation	71,237	A07–A19

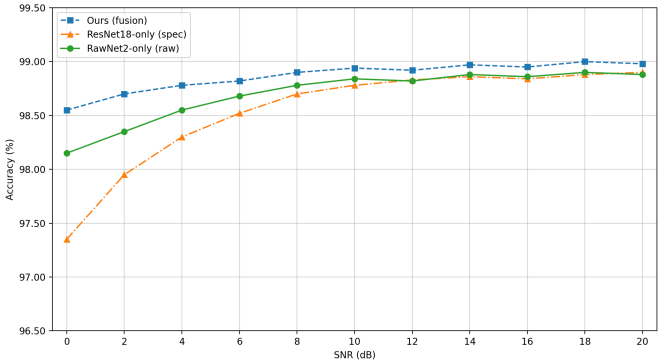


Fig. 2. Accuracy (%) under different signal-to-noise ratio (SNR) conditions for single-view baselines (ResNet18-only and RawNet2-only) and our fusion model.

IV. EXPERIMENTS

A. Experimental settings

We use the ASVspoof 2019 LA dataset for in-domain evaluation and train on the official train and development sets while testing on the evaluation set, where attacks A07–A19 are unseen during training [12], [40]. Table I summarizes the dataset partition. Cross-dataset generalization is evaluated on ASVspoof 2021 LA using the same model trained on ASVspoof 2019 LA. EER is the primary metric. For robustness analysis, we additionally report classification accuracy under additive Gaussian noise at multiple SNR levels.

Each utterance is standardized to 64,600 waveform samples (about 4 seconds). The spectral branch uses 60-dimensional LFCC features extracted with the official ASVspoof 2019 implementation. We train all models jointly with Adam (learning rate 10^{-4}, weight decay 10^{-4}, batch size 14) for up to 100 epochs and select checkpoints on the development set. We compare with two reproduced single-view baselines, Wav+RawNet2 and LFCC+ResNet-18, and report external baselines from the corresponding papers. For noisy-condition analysis, additive Gaussian noise is injected into the evaluation utterances at multiple SNR levels for robustness comparison only.

B. Results

Table II reports the main ASVspoof 2019 LA results. Our method achieves the lowest EER (1.02%), outperforming the reproduced single-view baselines Wav+RawNet2 (3.14%) and LFCC+ResNet-18 (4.75%) as well as representative fusion systems. These results indicate that uncertainty-guided credibility weighting can better suppress unreliable views than fixed fusion.

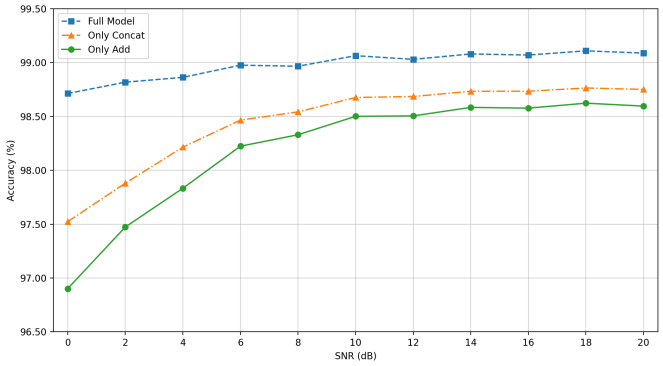


Fig. 3. Accuracy (%) under different signal-to-noise ratio (SNR) conditions for the full model and fusion ablations (Only Concat and Only Add).

TABLE II
EERs (%) ON THE ASVSPOOF 2019 LA EVALUATION SET.

Method	EER (%)
Baseline Wav+RawNet2 [30]	3.14
Baseline LFCC+ResNet-18 [31]	4.75
Spec+CQCC+ResNet+SE [14]	5.56
8 Features+MLP [36]	4.25
MFCC+Mel-Spec+ECAPA-TDNN+Prosody Encoder [13]	5.39
Wav+Spec-Res-TSSDNet [6]	3.39
Ours	**1.02**

Cross-dataset results on ASVspoof 2021 LA are shown in Table III. Training only on ASVspoof 2019 LA, our method achieves 4.85% EER, better than the reproduced single-view baselines and competitive external systems, indicating improved generalization to unseen attacks and conditions.

This cross-dataset result suggests that the proposed fusion does not merely overfit the in-domain benchmark. By estimating the reliability of each view for each utterance, the model can better preserve useful spoofing cues under mismatched conditions and suppress less informative representations when the test distribution shifts.

The ablation study in Table IV confirms that each component contributes to performance. Replacing the proposed fusion with concatenation or addition degrades EER to 1.24% and 1.40%, respectively, while removing the uncertainty or conflict loss further increases EER to 1.78% and 1.66%. Fig. 3 shows the same trend under noise, where the full model remains consistently stronger than the two ablations across SNR levels.

Fig. 4 visualizes the fused embedding space and shows clearer bona fide/spoof separation for the full model than for the two simplified fusion variants. Fig. 5 further indicates that attack conditions with larger estimated uncertainty tend to have higher EER, supporting uncertainty as a meaningful reliability signal.

Noise robustness. Fig. 2 compares our model with the two single-view branches under additive Gaussian noise. All

TABLE III
EERs (%) ON THE ASVSPOOF 2021 LA EVALUATION SET, WITH SYSTEMS TRAINED ON THE ASVSPOOF 2019 LA TRAIN AND DEVELOPMENT SETS.

Method	EER (%)
Baseline Wav+RawNet2 [30]	9.50
Baseline LFCC+ResNet-18 [31]	5.75
AASIST [37]	8.08
AASIST + FADEL [37]	5.60
UR-AIR [38]	5.46
DeepLASD [39]	12.76
Ours	**4.85**

TABLE IV
ABLATION STUDY RESULTS ON THE ASVSPOOF 2019 LA EVALUATION SET (EER, %).

Category	Ablation Settings	EER (%)
Fusion Method	Only Concat	1.24
	Only Add	1.40
Loss	w/o $\mathcal{L}_u$	1.78
	w/o $\mathcal{L}_c$	1.66
Full Model	**Ours**	**1.02**

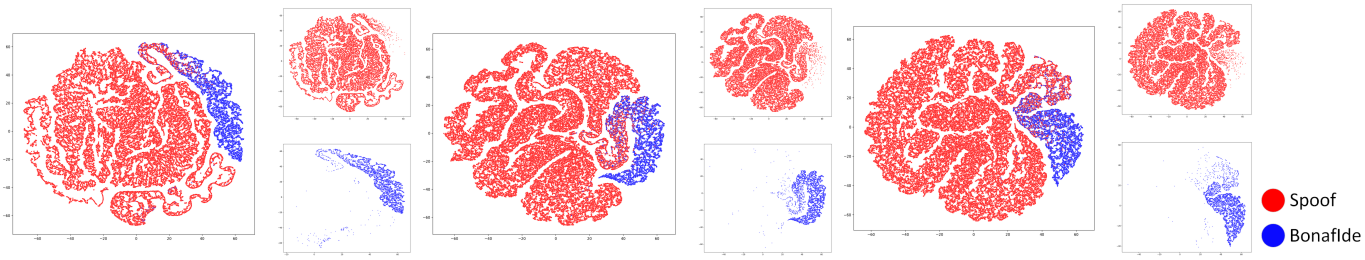
systems improve as SNR increases, but the spectral-only model degrades more sharply at low SNR. Across the entire range, the proposed fusion model remains the most accurate, showing that uncertainty-guided view reweighting improves robustness when one representation is heavily corrupted.

The visualization results support the same conclusion from a representation-learning perspective. Compared with concatenation and addition baselines, the full model forms a more compact bona fide cluster and a clearer spoof boundary in the embedding space, indicating that uncertainty-aware fusion improves not only the final classifier score but also the geometry of the fused representation.

C. Discussion

The experimental results show that the proposed method benefits from both adaptive weighting and explicit disagreement modeling. Credibility weights help the model avoid blindly trusting a corrupted view, while the conflict term encourages the two branches to produce compatible opinions when reliable complementary evidence exists. These components therefore address different failure modes and work best together, which is consistent with the ablation results in Table IV.

Another useful observation is that uncertainty plays a dual role in the framework. On the one hand, it serves as a direct reliability indicator that modulates the contribution of each view during fusion. On the other hand, it provides an interpretable signal for understanding difficult attacks and mismatch conditions. The positive relationship between uncertainty and EER in Fig. 5 suggests that the model is not only accurate, but also reasonably aware of when the decision problem becomes harder.



(a) Full Model (b) Only Concat (c) Only Add

Fig. 4. Visualization of the embedding space using t-SNE on the ASVspoof 2019 LA evaluation dataset.

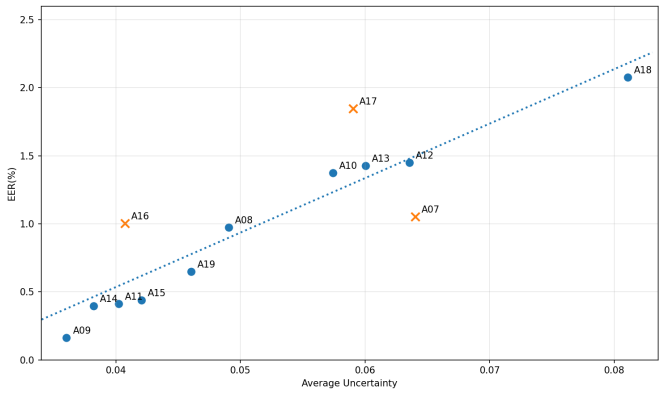


Fig. 5. Scatter plot of attack-wise uncertainty versus EER on the ASVspoof 2019 LA evaluation set.

From a deployment perspective, this property is valuable for practical spoofed speech detection systems. In real applications, the detector may encounter channel distortion, novel synthesis algorithms, or recording conditions that differ from those seen during training. A fusion framework that can adapt its reliance on different views and expose uncertainty information is more suitable for such open-world settings than a fixed fusion strategy that assumes all views are equally trustworthy.

An additional practical advantage is that the proposed system does not rely on a large ensemble of heterogeneous branches. Instead, it combines two complementary views with an explicit credibility mechanism, which keeps the model comparatively simple while still improving robustness and interpretability. This balance is useful for real deployment scenarios where computational budget and decision transparency matter alongside raw benchmark performance.

More broadly, the results suggest that the fusion stage in spoofed speech detection should be treated as a reliability-allocation problem rather than a simple feature-combination step. Even with only two views, uncertainty-aware weighting yields consistent gains over stronger fixed aggregation baselines, which indicates that modeling when a representation should dominate the final decision is itself an important source of improvement.

This sample-level adaptability is particularly relevant for benchmark conditions in which spoofing algorithms, channels, and recording quality vary substantially across trials. A fusion rule that changes with the input is therefore better matched to the task than one global weight configuration learned once and applied uniformly to all samples.

V. Conclusion

This paper proposes an uncertainty-guided multi-view feature fusion framework for spoofed speech detection. By combining raw-waveform and LFCC representations with credibility-weighted fusion, the proposed method achieves lower EER than strong single-view and fusion baselines on ASVspoof 2019 LA and improves cross-dataset generalization on ASVspoof 2021 LA. The results confirm that uncertainty-aware view reweighting is effective for robust spoofed speech detection. The framework is also practically attractive because it uses only two complementary views and a lightweight credibility mechanism, providing a reasonable balance between robustness, interpretability, and implementation complexity. Overall, the results support reliability-aware fusion as a principled and effective design choice for building more dependable anti-spoofing systems, especially when acoustic conditions vary across utterances and the usefulness of each view is no longer constant. These findings also show that improving anti-spoofing performance is not only a matter of stronger encoders, but also of learning how much each representation should be trusted for each input sample.

References

[1] C. Stupp, "Fraudsters used AI to mimic a CEO's voice in unusual cybercrime case," The Wall Street Journal, 2019, aug. 30, 2019.

[2] A. Mittal and M. Dua, "Automatic speaker verification systems and spoof detection techniques: review and analysis," *International Journal of Speech Technology*, vol. 25, no. 1, pp. 105–134, 2022.

[3] H. Ma, J. Yi, J. Tao, Y. Bai, Z. Tian, and C. Wang, "Continual learning for fake audio detection," *arXiv preprint arXiv:2104.07286*, 2021.

[4] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," *arXiv preprint arXiv:2009.09637*, 2020.

[5] A. K. S. Yadav, K. Bhagtani, Z. Xiang, P. Bestagini, S. Tubaro, and E. J. Delp, "Dsvae: Interpretable disentangled representation for synthetic speech detection," *arXiv preprint arXiv:2304.03323*, 2023.

[6] N. Yu, L. Chen, T. Leng, Z. Chen, and X. Yi, "An explainable deepfake of speech detection method with spectrograms and waveforms," *Journal of Information Security and Applications*, vol. 81, p. 103720, 2024.

[7] K. Ma, Y. Feng, B. Chen, and G. Zhao, "End-to-end dual-branch network towards synthetic speech detection," *IEEE Signal Processing Letters*, vol. 30, pp. 359–363, 2023.

[8] F. Hassan and A. Javed, "Voice spoofing countermeasure for synthetic speech detection," in *2021 International conference on artificial intelligence (ICAI)*. IEEE, 2021, pp. 209–212.

[9] Z. Teng, Q. Fu, J. White, M. E. Powell, and D. C. Schmidt, "Arawnet: A lightweight solution for leveraging raw waveforms in spoof speech detection," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 692–698.

[10] C. Wang, J. Yi, J. Tao, C. Zhang, S. Zhang, and X. Chen, "Detection of cross-dataset fake audio based on prosodic and pronunciation features," *arXiv preprint arXiv:2305.13700*, 2023.

[11] Q. Zhang, H. Wu, C. Zhang, Q. Hu, H. Fu, J. T. Zhou, and X. Peng, "Provable dynamic fusion for low-quality multimodal data," in *International conference on machine learning*. PMLR, 2023, pp. 41 753–41 769.

[12] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[13] L. Attorresi, D. Salvi, C. Borrelli, P. Bestagini, and S. Tubaro, "Combining automatic speaker verification and prosody analysis for synthetic speech detection," in *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*. Springer Nature Switzerland, 2023, pp. 247–263.

[14] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "Assert: Anti-spoofing with squeeze-excitation and residual networks," *arXiv preprint arXiv:1904.01120*, 2019.

[15] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan, S. Liang, S. Wang, S. Zhang, Z. Wen, X. Yan, L. Xu, and H. Li, "Add 2022: The first audio deep synthesis detection challenge," 2022, iCASSP 2022 presentation / challenge description, DOI: 10.17023/wp17-7k98.

[16] Z. Ji, C. Lin, H. Wang, and C. Shen, "Speech-forensics: Towards comprehensive synthetic speech dataset establishment and analysis," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, pp. 413–421.

[17] Z. Xie, B. Li, X. Xu, Z. Liang, K. Yu, and M. Wu, "Fakesound: Deepfake general audio detection," in *Interspeech 2024*, 2024, preprint: arXiv:2406.08052.

[18] N. A. Chandra, R. Murtfeldt, L. Qiu, A. Karmakar, H. Lee, E. Tanumihardja, K. Farhat, B. Caffee, S. Paik, C. Lee, J. Choi, A. Kim, and O. Etzioni, "Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024," 2025.

[19] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 2, pp. 2551–2566, 2022.

[20] C. Xu, J. Si, Z. Guan, W. Zhao, Y. Wu, and X. Gao, "Reliable conflictive multi-view learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 16 129–16 137.

[21] Z. Ma, M. Luo, H. Guo, Z. Zeng, Y. Hao, and X. Zhao, "Event-radar: Event-driven multi-view learning for multimodal fake news detection," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 5809–5821.

[22] Ç. Süslü, E. Eren, and C. Demiroglu, "Uncertainty assessment for detection of spoofing attacks to speaker verification systems using a bayesian approach," *Speech Communication*, vol. 137, pp. 44–51, 2022.

[23] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371.

[24] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[25] C. Wang, J. Yi, J. Tao, C. Zhang, S. Zhang, R. Fu, and X. Chen, "Torawnet: Improving rawnet with tcn and orthogonal regularization for fake audio detection," 2023.

[26] H. Wu, J. Chen, R. Du, C. Wu, K. He, X. Shang, H. Ren, and G. Xu, "Clad: Robust audio deepfake detection against manipulation attacks with contrastive learning," 2024.

[27] M. Rabhi, S. Bakiras, and R. Di Pietro, "Audio-deepfake detection: Adversarial attacks and countermeasures," *Expert Systems with Applications*, vol. 250, p. 123941, 2024.

[28] X. Li, K. Li, Y. Zheng, C. Yan, X. Ji, and W. Xu, "Safeear: Content privacy-preserving audio deepfake detection," 2024, accepted by ACM CCS 2024.

[29] C. Y. Kwok, J. Q. Yip, Z. Qiu, C. H. Chi, and K. Y. Lam, "Bona fide cross testing reveals weak spot in audio deepfake detection systems," in *Interspeech 2025*, 2025, pp. 2230–2234.

[30] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.

[31] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[32] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, "Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion," in *Interspeech 2018-19th Annual Conference of the International Speech Communication Association*. ISCA, 2018.

[33] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification," in *International Conference on Learning Representations (ICLR)*, 2021.

[34] A. Jøsang, *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer, 2016.

[35] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[36] T.-H. Shih, C.-Y. Yeh, and M.-S. Chen, "Does audio deepfake detection rely on artifacts?" in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 446–12 450.

[37] J. Y. Kang, J. W. Yoon, S. Kim, M. H. Han, and N. S. Kim, "Fadel: Uncertainty-aware fake audio detection with evidential deep learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025.

[38] X. Chen, Y. Zhang, G. Zhu, and Z. Duan, "Ur channel-robust synthetic speech detection system for asvspoof 2021," 2021, aSVspoof 2021 Workshop Paper (arXiv:2107.12018). ISCA Archive: https://www.isca-archive.org/asvspoof_2021/chen21_asvspoof.html.

[39] H. Al-Tairi, A. Javed, T. Khan, and A. K. J. Saudagar, "Deeplasd countermeasure for logical access audio spoofing," *Scientific Reports*, 2025, eER 12.76% on ASVspoof 2021 LA evaluation.

[40] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for the cstr voice cloning toolkit," University of Edinburgh. Datashare, 2016, available at: https://datashare.is.ed.ac.uk/handle/10283/2651.