AFSS: Artifact-Focused Self-Synthesis for Mitigating Bias in Audio Deepfake Detection

Hai-Son Nguyen-Le[⋆†] Hung-Cuong Nguyen-Thanh[⋆†] Nhien-An Le-Khac[‡]
Dinh-Thuc Nguyen[†] Hong-Hanh Nguyen-Le[‡§]

[⋆] Equal contribution. [§] Corresponding author: hong-hanh.nguyen-le@ucdconnect.ie
[†] University of Science, Ho Chi Minh City, Vietnam
[‡] University College Dublin, School of Computer Science, Dublin, Ireland

***Abstract*—The rapid advancement of generative models has enabled highly realistic audio deepfakes, yet current detectors suffer from a critical bias problem, leading to poor generalization across unseen datasets. This paper proposes Artifact-Focused Self-Synthesis (AFSS), a method designed to mitigate this bias by generating pseudo-fake samples from real audio via two mechanisms: self-conversion and self-reconstruction. The core insight of AFSS lies in enforcing same-speaker constraints, ensuring that real and pseudo-fake samples share identical speaker identity and semantic content. This forces the detector to focus exclusively on generation artifacts rather than irrelevant confounding factors. Furthermore, we introduce a learnable reweighting loss to dynamically emphasize synthetic samples during training. Extensive experiments across 7 datasets demonstrate that AFSS achieves state-of-the-art performance with an average EER of 5.45%, including a significant reduction to 1.23% on WaveFake and 2.70% on In-the-Wild, all while eliminating the dependency on pre-collected fake datasets. Our code is publicly available at https://github.com/NguyenLeHaiSonGit/AFSS.**

***Index Terms*—Audio deepfake detection, bias mitigation, generalization**

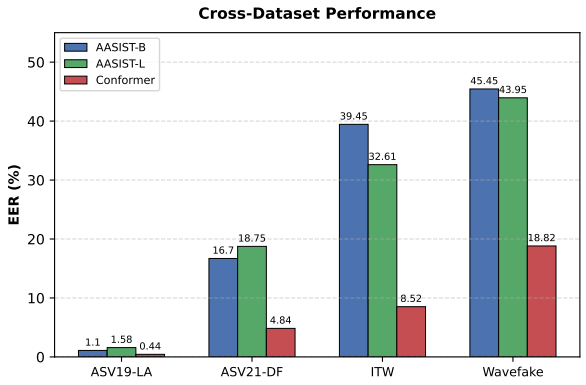

Fig. 1: Cross-dataset performance of audio deepfake detection methods (EER%) when trained on ASV19-LA and tested on different datasets. Lower values indicate better performance.

I. Introduction

The rapid development of generative models (GMs) has enabled the creation of highly realistic synthetic speech. Notable examples include Generative Adversarial Networks (GANs) [1], Variational Autoencoders (VAEs) [2], and Diffusion Models (DMs) [3]. Voice conversion (VC) and text-to-speech (TTS) systems built upon these architectures can now generate audio that is perceptually indistinguishable from genuine human speech [4]. While these technologies offer legitimate applications in entertainment and human-computer interaction, they have been increasingly exploited for malicious purposes, such as financial fraud, identity impersonation, and large-scale misinformation campaigns [5]–[7]. The sophistication and accessibility of these synthesis tools underscore the urgent need for effective audio deepfake (DF) detection methods capable of identifying manipulated audio across diverse real-world scenarios.

Recent detection approaches can be broadly categorized into two architectural paradigms. The first employs specialized end-to-end (E2E) architectures that learn discriminative features directly from spectral representations. Notable examples include RawNet [8], [9], which performs time-domain convolution directly on raw audio, and AASIST [10], which utilizes graph neural networks to model the non-Euclidean relationships within speech signals to capture complex short-range and long-range features. These methods design task-specific architectures optimized for capturing synthesis artifacts. The second paradigm adopts self-supervised learning (SSL) models as front-end feature extractors. Models such as Wav2vec 2.0 [11], XLS-R [12], and HuBERT [13] are first pre-trained on large-scale generic speech corpora to learn rich acoustic representations, then fine-tuned on downstream DF detection tasks. This paradigm has gained popularity due to its ability to leverage powerful pre-trained representations that capture subtle acoustic characteristics directly from raw waveforms, eliminating the need for handcrafted features such as MFCCs or spectrograms.

Despite achieving excellent in-domain accuracy, both paradigms exhibit significant performance degradation when evaluated on unseen datasets, as shown in Fig. 1. This considerable cross-domain performance gap reveals a fundamental limitation of current detection approaches that transcends architectural choices. The primary cause of this failure is the ***bias problem***: detectors inadvertently learn to distinguish real from fake audio based on irrelevant information rather than authentic forgery artifacts [14]. For SSL-based approaches, models like Wav2vec 2.0 inherently encode semantic content

and speaker identity since they were optimized for speech recognition and speaker identification tasks. For E2E architectures, the bias emerges from dataset-specific characteristics such as speaker populations or recording conditions. When these irrelevant features vary across datasets, detection performance degrades substantially [15]. In addition to the bias problem, collecting large and diverse fake datasets for training remains a major practical challenge. Although recent studies, such as the ALDEN framework [16], have attempted to disentangle synthetic-irrelevant information to improve generalization, these methods often still require extensive fake data from the training dataset. This dependency makes it difficult for detection systems to keep pace with the continuous emergence of new generative technologies.

In this work, we propose ***Artifact-Focused Self-Synthesis (AFSS)*** method specifically designed to mitigate the bias problem in audio DF detection. The core insight of AFSS is to generate pseudo-fake samples that contain authentic forgery artifacts while explicitly controlling for confounding factors. AFSS employs two artifact-focused self-generation mechanisms: **self-conversion**, which transforms audio using the same speaker as both source and target, and **self-reconstruction**, which passes audio through neural vocoders to introduce characteristic synthesis artifacts. By ensuring that real and pseudo-fake samples share identical speaker identity and semantic content, AFSS eliminates the possibility of the detector exploiting irrelevant features (i.e., speakers' identity) as classification shortcuts. The detector is thus forced to focus exclusively on the generation artifacts introduced by VC systems and neural vocoders. Furthermore, we introduce a learnable reweighting loss that dynamically adjusts the emphasis on synthetic samples during training, encouraging the model to actively explore and learn universal generation artifacts for enhanced cross-domain generalization. Importantly, AFSS operates without requiring any pre-collected fake datasets, generating all training samples from real audio alone.

Our main contributions are summarized as follows:

- We propose Artifact-Focused Self-Synthesis method to mitigate the bias problem in audio DF detection by generating artifact-focused pseudo-fake samples through self-conversion and self-reconstruction mechanisms with same-speaker constraints. Our method also eliminates dependency on pre-collected fake datasets by generating all training samples from real audio alone.
- We introduce a learnable reweighting loss that dynamically emphasizes synthetic samples, encouraging the model to focus on universal generation artifacts.
- Extensive experiments across seven datasets demonstrate state-of-the-art cross-domain performance with an average EER 5.45%, including 1.23% on WaveFake and 2.70% on In-the-Wild.

II. Related Work

A. Audio Deepfake Detection Methods

Recent advancements in audio DF detection are primarily characterized by two architectural paradigms: the pipeline-based and the E2E detection approaches.

Pipeline-based approaches employ a two-stage architecture where the front-end feature extractor operates independently of the back-end classifier. The front-end computes handcrafted acoustic features such as MFCCs, LFCCs, or spectrograms through signal processing algorithms without learnable parameters. These fixed representations are then fed to a separately trained back-end classifier. Since no gradient flows from the classifier to the feature extractor, the two stages are optimized in isolation. In contrast, E2E approaches integrate feature extraction and classification into a unified network where backpropagation flows from the classification output through the entire architecture [8]–[10], [17], [18]. This enables joint optimization of both components for the detection objective. Furthermore, SSL-based methods [19]–[21] using Wav2vec 2.0 or XLS-R as fine-tuned front-ends also fall within this paradigm, as gradient updates propagate through the pre-trained encoder during task-specific training.

Beyond architectural design, several strategies address generalization: one-class learning models bonafide distributions to detect anomalies [22]; synthetic data training exposes detectors to diverse TTS and VC outputs [23], [20]; augmentation techniques such as RawBoost [24] simulate acoustic distortions; and parameter-efficient methods including LoRA enable adaptation to novel generators [25].

B. Disentanglement methods for audio DF detection

Disentanglement methods for audio DF detection seeks to separate forgery-relevant from forgery-irrelevant factors to mitigate bias in audio deepfake detection. DSVAE [26] introduces an interpretable framework using a Variational Autoencoder to decompose speech into recording environments and fundamental speech characteristics, enhancing robustness against unseen synthesizers. Similarly, SafeEar [27] proposes a content-privacy-preserving mechanism that disentangles audio into "forgery-related" acoustic features and "identity-related" semantic features, utilizing adversarial training to focus the detector solely on acoustic artifacts. More recently, the ALDEN framework [16] has emerged as one of the leading methods in the field by expanding this concept into a dual-level architecture. By isolating vocoder-specific signals from high-level semantics, ALDEN achieves high generalization performance in real-world scenarios.

While effective, these approaches share two fundamental limitations: (1) they require pre-collected fake datasets to learn the disentanglement, creating dependency on existing synthetic corpora, and (2) they rely on the network's capacity to implicitly separate confounding factors during training, which may be dataset-specific. AFSS takes a fundamentally different approach by eliminating confounding factors by construction rather than learning to disentangle them. Through same-speaker constraints, real and pseudo-fake samples share identical speaker identity and semantic content by design, guaranteeing that these factors cannot serve as classification shortcuts. This explicit constraint removes dependency on pre-

collected fake datasets while ensuring complete bias elimination.

III. Proposed Method

This section details our proposed **Artifact-Focused Self-Synthesis (AFSS)** method. The core insight of AFSS is that the bias problem can be systematically eliminated by ensuring that real and pseudo-fake samples share identical speaker identity and semantic content. Under this constraint, the detector cannot exploit speaker-related or content-related features as classification shortcuts, and is thus forced to focus exclusively on the generation artifacts introduced by VC systems and neural vocoders.

AFSS operates through two complementary branches: self-conversion and self-reconstruction, with each branch targets distinct categories of forgery artifacts. Fig. 2 illustrates our proposed method, while Fig. 3 presents the overview of our model architecture.

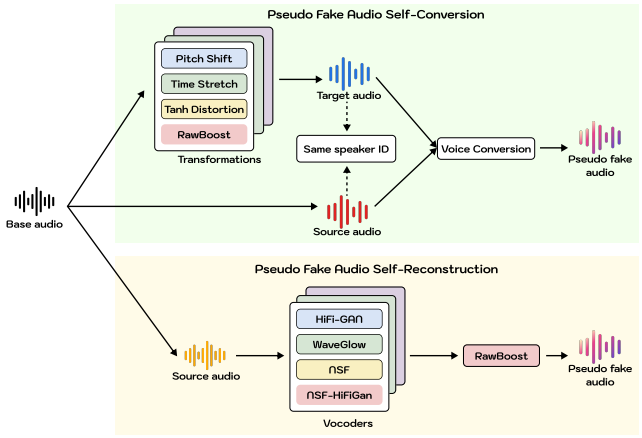

Fig. 2: Overview of the proposed AFSS method.

A. Pseudo-Fake Audio Self-Conversion

In conventional training paradigms, VC systems transform audio from a source speaker to a target speaker, creating a systematic correlation between speaker identity changes and the fake label [28]. Detectors trained on such data may learn to identify *speaker inconsistency* rather than actual VC artifacts, leading to poor generalization when encountering new speaker populations. Our self-conversion branch addresses this bias by converting from the same speaker identity.

Given a base audio sample x_b from speaker s, we first generate an acoustically modified variant $x_t = \alpha_i(x_b)$ using transformation techniques $\mathcal{A} = \{\alpha_i\}_{i=1}^{m}$, where each α_i represents a specific acoustic modification (i.e., pitch shifting, time stretching, tanh distortion, or RawBoost [24]). Crucially, α_i preserves speaker identity while altering acoustic characteristics sufficient to drive the VC process. The VC module $\mathcal{V}$ then transforms the source audio to match the target: $\hat{x}_{\text{vc}} = \mathcal{V}(x_b, x_t)$. Since both x_b and x_t originate from the same speaker s, the resulting pseudo-fake audio $\hat{x}_{\text{vc}}$ contains authentic VC artifacts without any speaker identity confounders.

This approach is compatible with any VC system. In our implementation, we employ kNN-VC (*k*-Nearest-Neighbor Voice Conversion) [29] for its computational efficiency and lightweight architecture, though the framework generalizes to arbitrary VC models.

Transformation Strategy. The placement of transformation techniques before the VC process is critical for artifact preservation. For each sample, we apply one of the four transformation techniques, including PitchShift, TimeStretch, TanhDistortion, or RawBoost, which are selected randomly with equal probability. These applied techniques ensure sufficient acoustic diversity to drive the conversion module while avoiding excessive signal distortion that could compromise speaker identity.

B. Pseudo-Fake Audio Self-Reconstruction

The self-reconstruction branch targets a complementary source of bias: the characteristic fingerprints left by neural vocoders used in the final stage of most TTS pipelines. To expose the detector to diverse vocoder artifacts while maintaining speaker and content consistency, we process base audio through various vocoders $\mathcal{G} = \{g_i\}_{i=1}^{n}$ for reconstruction: $\hat{x}_{\text{rec}} = g_i(\text{encode}(x_b))$.

This simulates the encoding-decoding pipeline typical of DF generation, introducing characteristic vocoder artifacts including over-smoothed spectral details, phase inconsistencies, and loss of natural micro-variations. Since the reconstructed audio $\hat{x}_{\text{rec}}$ derives from the original x_b, speaker identity and semantic content remain unchanged. The framework supports any reconstruction vocoder, but in our experiment, we select four widely-used non-autoregressive models: HiFiGAN [1], WaveGlow [30], Hn-NSF [31], and NSF-HiFiGAN [32].

Transformation Strategy. Unlike the self-conversion branch, transformations in this branch are applied after the vocoder reconstruction using exclusively RawBoost. This placement is physically motivated: neural vocoder fingerprints reside in delicate micro-variations of spectral and phase detail. Aggressive transformations such as PitchShift or TimeStretch applied before or during reconstruction would mask these subtle traces. This design ensures the detector learns robust vocoder fingerprints that generalize across transmission conditions.

C. Learnable Reweighting Loss Function

The standard Binary Cross-Entropy loss treats all authentic and synthetic audio with equal importance, which prevents the model from adequately focusing on synthetic samples generated by our AFSS mechanism. In this work, we expect the model to explore artifacts left by VC models and vocoders by emphasizing synthetic samples, helping enhance cross-domain generalization. To achieve this, we propose a Learnable Reweighting Loss:

$$\mathcal{L} = -[w_{\text{fake}} \cdot y \log(p) + w_{\text{real}} \cdot (1-y) \log(1-p)], \quad (1)$$

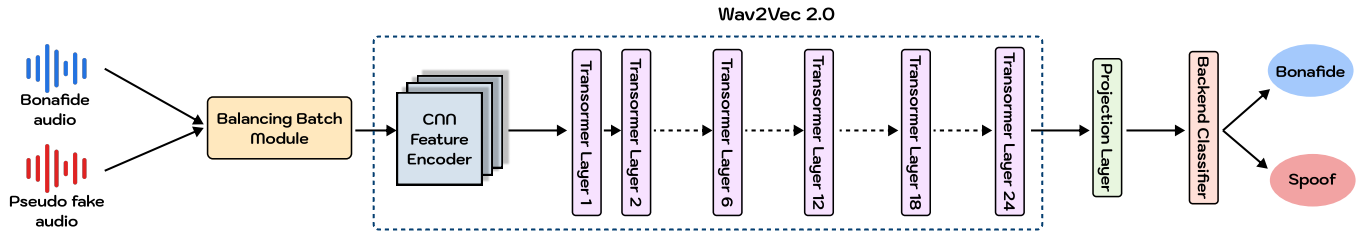

Fig. 3: Overview of model architecture.

where $y \in \{0, 1\}$ is the ground-truths and p is the predicted priori probability. Unlike fixed-weight approaches, w_{fake} and w_{real} are derived from learnable parameters $\tilde{w}_{\text{real}}$ and $\tilde{w}_{\text{fake}}$ optimized via backpropagation. Sigmoid transformations ensure stable weight ranges:

$$w_{\text{fake}} = 1.0 + \sigma(\tilde{w}_{\text{fake}}), \quad w_{\text{real}} = \sigma(\tilde{w}_{\text{real}}) \tag{2}$$

where σ denotes the sigmoid function. This formulation strictly constrains $w_{\text{fake}} \in (1, 2)$ and $w_{\text{real}} \in (0, 1)$. All learnable parameters, including $\tilde{w}_{\text{real}}$ and $\tilde{w}_{\text{fake}}$, are jointly optimized to minimize $\mathcal{L}$. The **1.0 bias** for w_{fake} guarantees that self-generated samples always receive a higher loss weight than real samples, while the learnable sigmoid components allow the weights to dynamically adapt during training, compelling the model to focus on the synthetic class.

D. Balanced Batch Sampling

Random batch sampling can introduce class imbalance that destabilizes training and interferes with our learnable loss function. Inspired by the work [17], we implement a balanced batch strategy that ensures each mini-batch contains equal numbers of real and fake samples. This strategy prevents the learnable weights from being skewed by random class distribution fluctuations, maintaining stable gradient updates and consistent learning dynamics throughout training.

IV. Experimental Setup

A. Dataset

1) Training Dataset: Since our AFSS method automatically generates pseudo-fake samples from real audio, we only use authentic audio samples from the training and development partitions of the ASVspoof 2019 LA dataset [33] for training.

2) Evaluation Datasets: To evaluate our method on cross-domain generation scenarios, we utilize 7 different datasets: ASVspoof 2019 LA-Eval [33], ASVspoof 2021 LA evaluation and hidden sets [34], ASVspoof-2021 DF evaluation and hidden sets [34], WaveFake [35], and In-the-Wild [36].

B. Baseline

We compare our method with 5 baselines, including AASIST-B [10], AASIST-L [10], SCL [17], Conformer-based detector [18], and SSL [20]. We use the code provided by these studies for implementation, excluding the work [20] which we report the results from the authors' publication. We also compare with disentanglement methods: ALDEN [16], DSVAE [26], and SafeEar [27].

C. Evaluation Metric

We use Equal Error Rate (EER), Accuracy (ACC), Area Under the ROC Curve (AUC), and Average Precision (AP) as the primary evaluation metrics for this paper.

D. Implementation Details

The framework utilizes XLS-R as the front-end feature extractor, projecting 1024-dimensional features to 128 dimensions via a linear layer. The back-end classifier comprises ReLU activation, Dropout (p = 0.5), Mean Pooling, and a Dense layer for final classification.

Training employs AdamW optimizer with weight decay of 10^{-4} for 30 epochs maximum. We use differential learning rates: 5×10^{-6} for the SSL front-end, 10^{-4} for back-end layers and learnable loss parameters. A 5-epoch linear warm-up precedes linear decay to 10^{-6}. BalancedBatchSampler creates mini-batches of 12 samples (6 bona fide, 6 fake). Early stopping with patience of 10 epochs ensures training stability.

E. Transformation Intensities

The intensity levels are defined by the aggressiveness of their parameter ranges:

- Intensity 0 (Baseline): Applies minimal transformation.
- Intensity 1 (Optimal): Our proposed setting that balances diversity and realism. It includes TimeStretch ($[0.9, 1.1]$), TanhDistortion ($[0.15, 0.6]$), PitchShift ($[-0.5, 0.5]$ semitones), and RawBoost (utilizing both convolutive noise with $nBands = 5$ and additive noise).
- Intensity 2 & 3: Use increasingly aggressive ranges to simulate highly distorted environments.

V. Experimental Results

A. Performance Across Datasets

We report results on EER and AP metrics across 7 audio DF datasets. From Tab. I and II, some insights can be observed:

- **Superior Cross-Domain Generalization**. AFSS achieves the lowest average EER of 5.45% and highest average AUC of 98.15% across all evaluation datasets. This validates our core hypothesis: eliminating speaker

Method	Venue	ASVspoof 19	ASVspoof 21				Real-World		Avg
		eval	DF eval	DF hidden	LA eval	LA hidden	In-the-Wild	WaveFake	
AASIST-B [10]	ICASSP '22	1.10	16.70	21.16	7.40	32.81	39.45	45.45	23.44
AASIST-L [10]	ICASSP '22	1.58	18.75	20.11	8.57	26.47	32.61	43.95	21.72
SCL [17]	Interspeech '24	2.88	**2.17**	**5.76**	13.90	15.15	4.52	2.99	6.77
Conformer [18]	Interspeech '23	**0.44**	4.84	8.37	**1.67**	**12.19**	8.52	18.82	7.84
SSL [20]	ICASSP '24	1.91	5.67	8.84	15.92	14.97	6.10	1.30	7.82
AFSS (Ours)		2.72	2.19	6.92	10.02	12.35	**2.70**	**1.23**	**5.45**

TABLE I: Cross-dataset generalization performance comparison (EER%). Lower is better. Best results are in **bold**.

Method	Venue	ASVspoof 19	ASVspoof 21				Real-World		Avg
		eval	DF eval	DF hidden	LA eval	LA hidden	In-the-Wild	WaveFake	
AASIST-B [10]	ICASSP '22	0.9984	0.9059	0.8389	0.9811	0.7229	0.6531	0.9165	0.8595
AASIST-L [10]	ICASSP '22	0.9982	0.8989	0.8652	0.9702	0.7985	0.7372	0.9250	0.8847
SCL [17]	Interspeech '24	0.9984	**0.9968**	**0.9851**	0.9343	0.9200	0.9910	0.9901	0.9737
Conformer [18]	Interspeech '23	**0.9995**	0.9903	0.9720	**0.9942**	0.9418	0.9712	0.8868	0.9651
AFSS (Ours)		0.9939	0.9965	0.9794	0.9610	**0.9440**	**0.9964**	**0.9990**	**0.9815**

TABLE II: Cross-dataset generalization performance comparison (AUC). Higher is better. Best results are in **bold**.

Method	Venue	ASVspoof 19	ASVspoof 21				Real-World	
		LA eval	DF eval	DF hidden	LA eval	LA hidden	In-the-Wild	WaveFake
ALDEN [16]	ACM MM '25	1.22	4.39	-	-	-	11.27	4.66
DSVAE [26]		**2.16**	-	-	-	-	-	-
SafeEar [27]	CCS '24	3.10	-	-	**7.22**	-	-	-
AFSS (Ours)		2.72	**2.19**	6.92	10.02	12.35	**2.70**	**1.23**

TABLE III: Comparison with disentanglement methods (EER%). Results are taken from original publications; '-' indicates the metric was not reported by the original authors. Best results are in **bold**.

identity and semantic content as confounding factors forces the detector to learn universal generation artifacts.

- **High Real-World Performance**. The most pronounced advantages appear on challenging real-world scenarios. For example, on In-the-Wild dataset, AFSS achieves 2.70% EER, substantially outperforming AASIST-L (32.61%) and Conformer (8.52%). These results demonstrate effective transfer to unseen synthesis methods.
- **Generalization-Specialization Trade-off**. While Conformer achieves lower EER on ASVspoof 2019-eval, this advantage does not transfer to out-of-domain scenarios. AFSS sacrifices marginal in-domain accuracy for substantially better generalization, which is a more desirable property for real-world deployment where attack methods are unknown a priori.

B. Comparison with Disentanglement Methods

Tab.III compares AFSS against state-of-the-art disentanglement approaches. While ALDEN achieves lower EER on ASVspoof 2019-eval (1.22% vs. 2.72%), AFSS demonstrates superior performance on challenging real-world datasets: $4.2\times$ improvement on In-the-Wild and $3.8\times$ improvement on WaveFake. The key distinction lies in methodology. Disentanglement methods such as ALDEN, DSVAE, and SafeEar require pre-collected fake datasets to learn representations that separate forgery-relevant from forgery-irrelevant features. In contrast, AFSS eliminates this dependency entirely by generating pseudo-fake samples from real audio alone.

C. Ablation Study and Analysis

1) Components Contribution: Tab.IV shows that combining self-conversion and self-reconstruction (S-VC + S-Rec) achieves 8.57% EER on ASVspoof 2019-eval and 4.08% on In-the-Wild. This result validates that same-speaker constraints alone provide effective bias mitigation. Combining the learnable reweighting loss with balanced batch sampling yields improvements, reducing EER from 8.57% to 2.72% on ASVspoof 2019-eval and from 4.08% to 2.70% on In-the-Wild.

2) Effects of Transformation Intensity Levels: Since transformations in the self-conversion branch must preserve speaker identity while providing sufficient acoustic variation to drive the VC process, we investigate the optimal intensity level. As shown in Tab.V, Intensity 1 achieves the best generalization, with EERs of 1.23% on WaveFake and 2.70% on In-the-Wild. Insufficient transformation (Intensity 0) yields higher EER on WaveFake (6.46%), indicating inadequate acoustic diversity for robust artifact learning. Conversely, excessive transformation (Intensity 2–3) causes performance degradation, as the

Components	ASV19-eval				In-the-Wild			
	EER	AUC	ACC	AP	EER	AUC	ACC	AP
S-VC + S-Rec	8.57	97.81	91.44	99.75	4.08	99.14	95.92	98.57
+ Alpha	6.61	97.37	93.39	99.71	3.19	99.52	96.81	99.19
+ Bal	6.84	97.31	93.16	99.69	3.74	99.29	96.26	98.81
+ Alpha + Bal	**2.72**	**99.39**	**97.26**	**99.93**	**2.70**	**99.64**	**97.30**	**99.39**

TABLE IV: Effectiveness of the proposed components (%). S-VC: Self-conversion, S-Rec: Self-reconstruction, Alpha: Learnable Reweighting Loss, Bal: Balanced Batch Sampling.

Dataset	Intensity 0	Intensity 1	Intensity 2	Intensity 3
ASV19 eval	**1.71%**	2.72%	9.01%	7.45%
DF21 eval	2.65%	**2.19%**	3.15%	5.74%
DF21 hidden	9.04%	**6.92%**	11.82%	13.64%
LA21 eval	**8.63%**	10.02%	13.96%	16.23%
LA21 hidden	12.55%	**12.35%**	15.05%	15.97%
WaveFake	6.46%	**1.23%**	14.10%	5.63%
In The Wild	3.34%	**2.70%**	6.02%	6.85%

TABLE V: Effects of data transformation intensity (EER%)

model learns transformation-specific distortions rather than universal forgery artifacts.

This analysis confirms that moderate transformation intensity optimally balances acoustic diversity with artifact preservation, which is essential for bias mitigation and cross-domain generalization.

LR	Alpha value	EER	ACC	AUC	AP
10^{-4}	$w_f = 1.73, w_r = 0.63$	1.63%	98.37%	99.81%	99.98%
10^{-5}	$w_f = 1.77, w_r = 0.68$	2.97%	97.02%	99.31%	99.93%
$\mathbf{10^{-6}}$	$w_f = 1.77, w_r = 0.69$	**1.23%**	**98.77%**	**99.90%**	**99.99%**

TABLE VI: Optimization stability analysis: learning rate effects on reweighting parameters. Results are on WaveFake dataset.

3) Sensitivity Analysis of Learnable Reweighting Parameters: The learnable reweighting parameters require careful optimization to maintain training stability. We analyze their sensitivity to learning rate on the WaveFake dataset. As shown in Tab. VI, a learning rate of 10^{-6} yields optimal performance (1.23% EER), while higher rates (10^{-4}, 10^{-5}) lead to degraded results despite converging to similar weight values. This indicates that rapid weight updates destabilize the balance between real and synthetic sample gradients.

These findings validate our design choice of constraining $w_{\text{fake}} \in (1, 2)$ with a 1.0 bias: a conservative learning rate combined with bounded weight ranges ensures that synthetic samples consistently receive higher emphasis while preventing model divergence.

4) Comparison with Conversion between Different Identities: To validate that our same-speaker constraint effectively mitigates bias, we compare AFSS against a standard cross-speaker voice conversion baseline. In cross-speaker VC, the source and target speakers differ, creating a systematic correlation between speaker identity change and the fake label.

Method	ASV19-eval	ASV21-DF	WaveFake	In-the-Wild
Cross-Speaker VC	5.68	21.25	33.11	55.98
AFSS (Ours)	**2.72**	**2.19**	**1.23**	**2.70**

TABLE VII: Comparison of Equal Error Rate (EER, %) between our proposed AFSS and a standard cross-speaker voice conversion (VC) baseline.

As shown in Tab. VII, cross-speaker VC exhibits severe degradation on out-of-domain datasets (33.11% on WaveFake, 55.98% on In-the-Wild), while AFSS achieves 1.23% and 2.70%, respectively. This performance gap confirms that cross-speaker VC introduces speaker-related bias that fails to generalize across datasets. By enforcing identical speaker identity between real and pseudo-fake samples, AFSS eliminates this confounding factor, compelling the detector to learn universal generation artifacts rather than spurious speaker-dependent correlations.

VI. Conclusion

In this paper, we presented the Artifact-Focused Self-Synthesis (AFSS) method to address the bias problem and enhance the generalization capability of audio DF detection systems. By enforcing same-speaker constraints through self-conversion and self-reconstruction mechanisms, AFSS effectively eliminates confounding factors related to speaker identity and semantic content, compelling the detector to focus exclusively on authentic forgery artifacts. Integrated with a learnable reweighting loss and balanced batch sampling, our approach achieved state-of-the-art cross-domain performance with an average EER of 5.45% across seven evaluation datasets. The high reliability of the proposed framework is evidenced by significant error reductions on WaveFake (1.23% EER) and In-the-Wild (2.70% EER). Most importantly, AFSS removes the dependency on pre-collected fake datasets, providing a practical and scalable solution for addressing the threats posed by rapidly evolving speech synthesis technologies.

REFERENCES

[1] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17022–17033, 2020.

[2] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7836–7846.

[3] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 13916–13932.

[4] Yogesh Kumar, Apeksha Koul, and Chamkaur Singh, "A deep learning approaches in text-to-speech system: a systematic review and recent research perspective," *Multimedia Tools and Applications*, vol. 82, no. 10, pp. 15171–15197, 2023.

[5] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[6] Catherine Stupp, "Fraudsters used ai to mimic ceo's voice in unusual cybercrime case," https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402, 2019.

[7] Natnicha Surasit, "Criminal exploitation of deepfakes in south east asia," 2024, https://globalinitiative.net/analysis/deepfakes-ai-cyber-scam-south-east-asia-organized-crime/.

[8] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony 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.

[9] Zhenyu Wang and John HL Hansen, "Audio anti-spoofing using a simple attention module and joint optimization based on additive angular margin loss and meta-learning," *arXiv preprint arXiv:2211.09898*, 2022.

[10] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 6367–6371.

[11] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[12] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al., "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[13] 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*, vol. 29, pp. 3451–3460, 2021.

[14] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu, "Ucf: Uncovering common features for generalizable deepfake detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 22412–22423.

[15] Jingze Lu, Yuxiang Zhang, Wenchao Wang, Zengqiang Shang, and Pengyuan Zhang, "One-class knowledge distillation for spoofing speech detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11251–11255.

[16] Yuxiong Xu, Bin Li, Weixiang Li, Sara Mandelli, Viola Negroni, and Sheng Li, "Alden: Dual-level disentanglement with meta-learning for generalizable audio deepfake detection," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 7277–7286.

[17] Thien-Phuc Doan, Long Nguyen-Vu, Kihun Hong, and Souhwan Jung, "Balance, multiple augmentation, and re-synthesis: A triad training strategy for enhanced audio deepfake detection," in *Proc. Interspeech 2024*, 2024, pp. 2105–2109.

[18] Eros Rosello, Alejandro Gomez-Alanis, Angel M Gomez, and Antonio Peinado, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," in *Proc. Interspeech*, 2023, vol. 2023, pp. 5281–5285.

[19] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *arXiv preprint arXiv:2202.12233*, 2022.

[20] Xin Wang and Junichi Yamagishi, "Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end?," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10311–10315.

[21] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[22] Hyun Myung Kim, KANGWOOK JANG, and Hoi-Rin Kim, "One-class learning with adaptive centroid shift for audio deepfake detection," in *Interspeech 2024*. ISCA, 2024, pp. 4853–4857.

[23] Xin Wang and Junichi Yamagishi, "Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6382–6386.

[25] Haochen Wu, Wu Guo, Shengyu Peng, Zhuhai Li, and Jie Zhang, "Adapter learning from pre-trained model for robust spoof speech detection," in *Interspeech*, 2024, vol. 2024, pp. 2095–2099.

[26] Amit Kumar Singh Yadav, Kratika Bhagtani, Ziyue Xiang, Paolo Bestagini, Stefano Tubaro, and Edward J Delp, "Dsvae: Interpretable disentangled representation for synthetic speech detection," *arXiv preprint arXiv:2304.03323*, 2023.

[27] Xinfeng Li, Kai Li, Yifan Zheng, Chen Yan, Xiaoyu Ji, and Wenyuan Xu, "Safeear: Content privacy-preserving audio deepfake detection," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 3585–3599.

[28] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao, "Audio deepfake detection: A survey," *arXiv preprint arXiv:2308.14970*, 2023.

[29] Matthew Baas, Benjamin van Niekerk, and Herman Kamper, "Voice conversion with just nearest neighbors," *arXiv preprint arXiv:2305.18975*, 2023.

[30] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[31] Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 402–415, 2019.

[32] Natalia Tomashenko, Xiaoxiao Miao, Pierre Champion, Sarina Meyer, Xin Wang, Emmanuel Vincent, Michele Panariello, Nicholas Evans, Junichi Yamagishi, and Massimiliano Todisco, "The voiceprivacy 2024 challenge evaluation plan," *arXiv preprint arXiv:2404.02677*, 2024.

[33] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al., "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, pp. 101114, 2020.

[34] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al., "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[35] Joel Frank and Lea Schönherr, "Wavefake: A data set to facilitate audio deepfake detection," *arXiv preprint arXiv:2111.02813*, 2021.

[36] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger, "Does audio deepfake detection generalize?," *arXiv preprint arXiv:2203.16263*, 2022.