

# Improving Generalization in Speech Deepfake Detection via Orthogonality-Constrained Common-Specific Feature Decorrelation

Donghee Kim<sup>1</sup> and Wooil Kim<sup>1,2,3</sup>

<sup>1</sup> Department of Artificial Intelligence, <sup>2</sup> Department of Computer Science and Engineering,

<sup>3</sup> Human-centered Artificial Intelligence Research Center,

Incheon National University, Incheon 22012, Republic of Korea

{d00h,wikim}@inu.ac.kr

## Abstract

Advances in deep learning-based Text-to-Speech (TTS) and Voice Conversion (VC) have enabled highly realistic speech deepfakes, posing serious threats such as fraud and misinformation. Although end-to-end deepfake detection models like RawNet2, RawGAT-ST, and AASIST have been proposed, they often overfit to dataset-specific artifacts under empirical risk minimization. To improve generalization, we propose an orthogonality-based feature decorrelation method for SSL-AASIST, which learns domain-invariant common features that capture consistent authenticity cues and attack-specific specialized features, together with an orthogonality regularization that minimizes their correlation. This encourages the model to rely on consistent authenticity cues rather than domain-dependent shortcuts. Experimental results demonstrate that our method achieves lower EERs than the baseline across unseen corpora such as 21DF, ASV5, and ITW, confirming improved robustness to unseen domains and spoofing attacks.

**Index Terms:** speech deepfake detection, anti-spoofing, domain generalization, feature decorrelation

## 1. Introduction

With the rapid advancement of deep learning-based speech synthesis technologies such as Text-to-Speech (TTS) and Voice Conversion (VC), it has become increasingly easy to generate high-quality deepfake speech that is perceptually indistinguishable from genuine speech. Although these technologies have driven remarkable innovation in applications such as voice assistants and digital media production, they also raise serious security and societal concerns, including voice phishing, financial fraud, and the dissemination of misinformation.

In response, extensive research has been devoted to detecting synthetic speech, with the ASVspoof Challenge [1, 2, 3] serving as a major international benchmark that has continuously advanced detection performance. Early approaches [4, 5, 6] relied on hand-crafted acoustic features or spectrogram-based representations, while recent studies have shifted toward end-to-end architectures that operate directly on raw waveforms [7, 8, 9]. A notable example is RawNet2 [7], which employs a sinc-convolutional front-end to extract meaningful features directly from raw waveforms for deepfake detection. Building on this, RawGAT-ST [8] incorporated a Graph Attention Network [10] to explicitly model intricate temporal-spectral relationships within the feature space, effectively capturing dependencies among local nodes. This line of work culminated in AASIST [9], which integrates an advanced graph-based module to jointly exploit temporal and spectral cues and is now widely used as a strong benchmark model in speech deepfake detection.

Despite impressive in-domain performance with exceptionally low Equal Error Rates (EERs), these models are typically trained under the empirical risk minimization (ERM) paradigm, causing them to overfit to dataset-specific artifacts and biases [11, 12]. Consequently, their detection capabilities significantly degrade when faced with distributional shifts—such as unseen attack algorithms, environmental noise, or codec-induced distortions—that occur in real-world conditions. For instance, [13] reported that detectors trained on the ASVspoof 2019 LA corpus can experience up to a tenfold increase in EER when evaluated on in-the-wild deepfake recordings. This drastic performance drop indicates an overreliance on superficial artifacts or dataset-specific patterns rather than on intrinsic cues of authenticity, contradicting the essential goal of reliable deepfake detection across diverse conditions.

To address this issue, we approach the problem from a domain generalization perspective and propose a method that minimizes correlations between domain-invariant and domain-specific representations through an orthogonality constraint [14, 15]. Orthogonality-based approaches in domain generalization encourage the separation of task-relevant and domain-related information by enforcing low correlation between feature subspaces used for main task prediction and for domain alignment. By regularizing these two representations to be orthogonal, the model learns disentangled features that capture independent sources of variation. This prevents the detector from relying on spurious, domain-specific shortcuts and instead promotes learning of more fundamental cues for authenticity. Specifically, we design a framework that explicitly formulates domain-common (authenticity-related) and domain-specific (attack-type-related) features and enforces orthogonality between them, thereby facilitating domain-agnostic representation learning for robust speech deepfake detection.

## 2. Related Work

### 2.1. Baseline system

For our baseline, we adopt SSL-AASIST [16], which combines self-supervised speech representations [17] with an AASIST back-end for speech deepfake detection. Raw waveforms are encoded into time-frequency feature maps and then transformed by a graph module into spectral and temporal graphs,  $\mathcal{G}_s$  and  $\mathcal{G}_t$ . These graphs are fused by a heterogeneous stacking graph attention layer and subsequent pooling to produce the final bonafide/spoof prediction. Further architectural details can be found in [16].

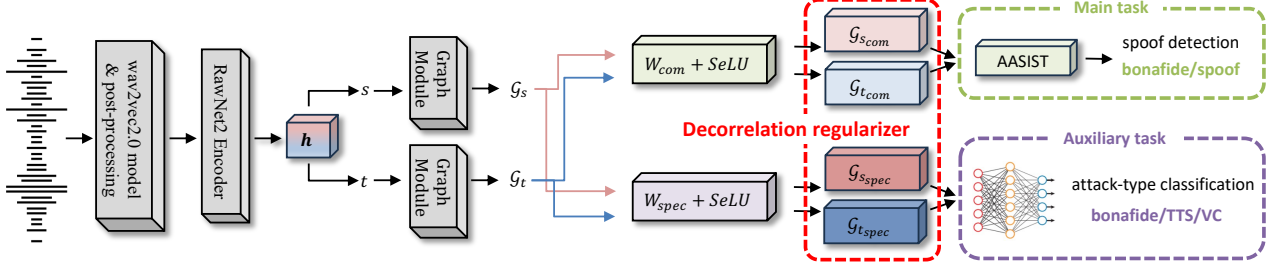


Figure 1: Overall architecture of the proposed method. The waveform is encoded by a wav2vec 2.0 front-end and RawNet2 into spectral-temporal features, which are aggregated and transformed into spectral and temporal graphs  $\mathcal{G}_s$  and  $\mathcal{G}_t$ . The resulting graph representations are projected into common and specific features, with the common branch used for the main spooof detection task and the specific branch for the auxiliary attack-type classification task.

## 2.2. Orthogonality constraint

Domain generalization aims to train on multiple source domains with diverse styles and conditions (e.g., PACS[18], VLCS[19]) and to obtain a classifier that maintains performance on unseen target domains, i.e., achieves out-of-distribution (OOD) generalization. A common approach decomposes the feature space into domain-invariant common features and domain-specific features, and improves generalization by enforcing orthogonality between them. The key assumption is that common features are consistently correlated with labels across domains, whereas domain-specific features have domain-dependent correlations and may cause performance drops under domain shifts if the classifier overrelies on them. Minimizing the correlation between common and specific components promotes their independence and encourages the model to focus on intrinsic, domain-invariant evidence.

Motivated by this, we incorporate an orthogonality-based common/specific feature formulation into the SSL-AASIST backbone. We introduce a common branch to capture authenticity cues and a specific branch to model attack-type-dependent variations, and impose an orthogonality regularization between them to enhance generalization to unseen domains and novel spoofing attacks.

## 3. Proposed Method

### 3.1. Definition of common-specific feature

The overall architecture of the proposed method is illustrated in Fig. 1. Our method is built upon the SSL-AASIST model described in Section 2.1. From the self-attentive aggregation module, SSL-AASIST derives spectral and temporal features  $s$  and  $t$ , which are then converted into graph representations by graph modules. Formally, let

$$\mathcal{G}_s = \text{GraphModule}_s(s) \in \mathbb{R}^{N_s \times d}, \quad (1)$$

$$\mathcal{G}_t = \text{GraphModule}_t(t) \in \mathbb{R}^{N_t \times d}, \quad (2)$$

denote the spectral and temporal graph representations, respectively, where  $N_s$  and  $N_t$  are the numbers of nodes in the spectral and temporal graphs and  $d$  is the feature dimension. Although  $\mathcal{G}_s$  and  $\mathcal{G}_t$  already encode rich time-frequency information, when trained under standard empirical risk minimization (ERM) they tend to entangle bonafide/spoof-related cues with domain-specific factors. To address this, we branch these representations into two separate feature streams and design each branch to learn a distinct type of representation.

The first branch models the common features. Using a linear projection matrix  $W_{com} \in \mathbb{R}^{d \times d_{com}}$  and a SeLU activation function, we map  $\mathcal{G}_s$  and  $\mathcal{G}_t$  into common representations as follows:

$$\mathcal{G}_{s_{com}} = \text{SeLU}(\mathcal{G}_s \cdot W_{com}) \in \mathbb{R}^{N_s \times d_{com}}, \quad (3)$$

$$\mathcal{G}_{t_{com}} = \text{SeLU}(\mathcal{G}_t \cdot W_{com}) \in \mathbb{R}^{N_t \times d_{com}}. \quad (4)$$

We follow the SSL-AASIST backbone in using SeLU and share  $W_{com}$  between the spectral and temporal graphs so that both are mapped into the same common feature space. These common features are used for the bonafide/spoof decision and are trained to capture generic cues that indicate whether an utterance is a genuine or spoofed speech signal, so we refer to them as common features. The final bonafide/spoof classifier makes its decisions based on this common representation.

The second branch models the specific features. Using another linear projection matrix  $W_{spec} \in \mathbb{R}^{d \times d_{spec}}$  and the same SeLU activation, we map the graph module outputs  $\mathcal{G}_s$  and  $\mathcal{G}_t$  into specific representations as

$$\mathcal{G}_{s_{spec}} = \text{SeLU}(\mathcal{G}_s \cdot W_{spec}) \in \mathbb{R}^{N_s \times d_{spec}}, \quad (5)$$

$$\mathcal{G}_{t_{spec}} = \text{SeLU}(\mathcal{G}_t \cdot W_{spec}) \in \mathbb{R}^{N_t \times d_{spec}}. \quad (6)$$

Here,  $W_{spec}$  is similarly shared between the spectral and temporal graphs. This branch is designed to primarily encode domain- and attack-specific factors, such as attack type and algorithm, dataset characteristics, and channel conditions. We introduce an auxiliary attack-type classification task (bonafide/TTS/VC) that is applied only to the specific representations, so that the specific features are naturally encouraged to capture variations across different attack types.

The choice of bonafide/TTS/VC as the auxiliary labels is motivated by the analysis in [2], which reports that, in the LA source data, bonafide utterances often contain long silence segments at the beginning and end, whereas many TTS attacks are generated without such non-speech regions. In this case, the duration of silence can become a strong, dataset-specific cue for distinguishing bonafide from spoofed speech and even TTS from VC. A detector that overly relies on such corpus-specific artifacts may fail to provide reliable detection in real-world scenarios. Therefore, we jointly perform the main bonafide/spoof detection and the auxiliary bonafide/TTS/VC classification to guide the specific branch toward modeling attack-related variations, while the common branch focuses on authenticity cues.

Each feature branch is then processed as illustrated in Fig. 1: the common features ( $\mathcal{G}_{s_{com}}$ ,  $\mathcal{G}_{t_{com}}$ ) are fed into the

AASIST back end, while the specific features ( $\mathcal{G}_{spec}, \mathcal{G}_{t_{spec}}$ ) are passed to an MLP classifier. With this role separation, the encoder is encouraged to encode shared cues about “whether the speech is bonafide or spoofed” into the common features, and attack-type-specific cues about “what kind of attack it is” into the specific features, even though both branches originate from the same input. This contrasts with the original architecture, where both kinds of information are entangled within a single feature space.

### 3.2. Orthogonality between common and specific features

Simply introducing two feature branches does not guarantee that the common and specific features encode different information; they may still redundantly capture similar cues. To avoid this, we introduce an orthogonality loss that suppresses the correlation between the two feature types. Given the input graph features  $\mathcal{G}_s$  and  $\mathcal{G}_t$ , we obtain the common representations ( $\mathcal{G}_{s_{com}}, \mathcal{G}_{t_{com}}$ ) and the specific representations ( $\mathcal{G}_{s_{spec}}, \mathcal{G}_{t_{spec}}$ ), and then compute a cosine-similarity-based penalty between each pair of common and specific features.

Formally, the orthogonality loss is defined as

$$\mathcal{L}_{ortho} = \frac{1}{N_s} \sum_{n_s=1}^{N_s} \left( \frac{\mathcal{G}_{s_{com}}^{(n_s)} \cdot (\mathcal{G}_{s_{spec}}^{(n_s)})^\top}{\|\mathcal{G}_{s_{com}}^{(n_s)}\| \|\mathcal{G}_{s_{spec}}^{(n_s)}\|} \right)^2 + \frac{1}{N_t} \sum_{n_t=1}^{N_t} \left( \frac{\mathcal{G}_{t_{com}}^{(n_t)} \cdot (\mathcal{G}_{t_{spec}}^{(n_t)})^\top}{\|\mathcal{G}_{t_{com}}^{(n_t)}\| \|\mathcal{G}_{t_{spec}}^{(n_t)}\|} \right)^2 \quad (7)$$

and is minimized during training. This loss reduces the correlation between the common and specific representations, encouraging them to become more independent and thereby minimizing the overlap between domain-invariant information and attack-specific information.

### 3.3. Overall training objective

The proposed model is trained with a combination of three loss terms.

First, the main loss (green region in Fig. 1,  $\mathcal{L}_{main}$ ) is a cross-entropy loss for binary bonafide/spoof classification computed from the common representations ( $\mathcal{G}_{s_{com}}, \mathcal{G}_{t_{com}}$ ), which directly optimizes the speech deepfake detection performance of interest.

Second, the auxiliary loss (purple region in Fig. 1,  $\mathcal{L}_{aux}$ ) is a cross-entropy loss for three-way bonafide/TTS/VC classification applied to the specific representations ( $\mathcal{G}_{s_{spec}}, \mathcal{G}_{t_{spec}}$ ). This auxiliary task encourages the specific branch to encode richer attack- and domain-dependent information.

Third, the decorrelation regularizer term (red region in Fig. 1) corresponds to the orthogonality loss ( $\mathcal{L}_{ortho}$ ) defined in Section 3.2, which reduces the cosine similarity between the common and specific features so that the two branches capture complementary and more independent information. The final training objective is defined as the sum of these three losses as follows:

$$\mathcal{L} = \mathcal{L}_{main} + \mathcal{L}_{aux} + \mathcal{L}_{ortho} \quad (8)$$

This allows both branches to retain their representational capacity while becoming more independent. In the next section, we evaluate the impact of the proposed method on generalization by reporting EERs on multiple evaluation datasets.

## 4. Experiments and Results

### 4.1. Databases

To evaluate the generalization capability of the proposed method, we use five databases with diverse characteristics. ASVspoof 2019 LA is used for both training and evaluation following its official partitions, while ASVspoof 2021 LA, ASVspoof 2021 DF, ASVspoof 5, and In-the-Wild (ITW) are used only for evaluation. ASVspoof 2019 LA and ASVspoof 2021 LA are based on VCTK [22], whereas the remaining datasets either include additional source corpora or are built from entirely different data.

ASVspoof 2019 LA (19LA) [1] is a large-scale corpus of bonafide and spoofed speech generated by various TTS and VC systems in controlled studio conditions, released with official training and evaluation sets; the training set contains 2,580 bonafide and 22,800 spoofed utterances from six spoofing algorithms, and the evaluation set has 7,355 bonafide and 63,882 spoofed utterances from 13 algorithms. ASVspoof 2021 LA (21LA) [2] extends the 19LA scenario by introducing a wider range of codecs and transmission conditions, while ASVspoof 2021 DF (21DF) [2] is constructed partly from the 19LA evaluation set along with additional speech data from VCC 2018 and 2020, and applies various lossy codecs to both bonafide and spoofed signals to emulate real-world media compression conditions. ASVspoof 5 (ASV5) [3] is a recent dataset that covers more diverse recording environments and devices than 19LA and includes not only deepfakes but also adversarial attacks in its evaluation set. The In-the-Wild (ITW) [13] database contains bonafide and deepfake speech from 58 politicians and other public figures collected from social media and online platforms, reflecting real-world dissemination scenarios.

### 4.2. Experimental Setup

For all experiments, we feed single utterances of approximately 4 s (64,600 samples) to a Wav2Vec 2.0 XLS-R 300M front-end, which outputs 1024-dimensional representations and is fine-tuned jointly with the SSL-AASIST back end. All hyperparameters, including the batch size (14) and learning rate ( $10^{-6}$ ), are kept identical to those used in SSL-AASIST [16]. The graph module outputs have dimension  $d = 64$ , and the projection matrices  $W_{com}$  and  $W_{spec}$  map them to 32-dimensional common and specific spaces ( $d_{com} = d_{spec} = 32$ ). We assign equal weights to the main, auxiliary, and orthogonality losses.

We further employ RawBoost [23], a waveform-level augmentation technique designed to mimic nuisance variability such as encoding, transmission, microphone, and channel distortions common in logical access scenarios. We adopt the best-performing configuration reported in the original paper, namely the combination of linear and non-linear convolutive noise (LnL) and impulsive signal-dependent additive noise (ISD).

### 4.3. Performance evaluation

Table 1 reports an ablation study on different loss combinations under the RawBoost setting. Using only the auxiliary loss with an AASIST back-end degrades performance on all evaluation datasets, confirming that the auxiliary task alone is not sufficient for reliable detection. Adding the auxiliary loss to the main loss slightly improves in-domain performance but provides limited gains on cross-corpus sets. In contrast, introducing the orthogonality loss on top of the main and auxiliary losses consistently reduces EERs on 21DF, ASV5, and ITW, demonstrating that explicitly decorrelating common and specific features plays an

Table 1: EER (%) of ablation systems using different loss combinations on each evaluation dataset with RawBoost (LnL & ISD) augmentation. A check mark indicates which loss terms are used.

| System                           | Loss                        |                            |                              | Evaluation Dataset |              |              |               |              |
|----------------------------------|-----------------------------|----------------------------|------------------------------|--------------------|--------------|--------------|---------------|--------------|
|                                  | $\mathcal{L}_{\text{main}}$ | $\mathcal{L}_{\text{aux}}$ | $\mathcal{L}_{\text{ortho}}$ | 19LA               | 21LA         | 21DF         | ASV5          | ITW          |
| Baseline (main) [16]             | ✓                           |                            |                              | 0.22%              | <b>0.82%</b> | 6.64%        | 16.25%        | 11.22%       |
| Baseline (aux)                   |                             | ✓                          |                              | 0.24%              | 1.28%        | 13.65%       | 42.32%        | 14.70%       |
| Baseline (main+aux)              | ✓                           | ✓                          |                              | <b>0.17%</b>       | 1.11%        | 6.46%        | 16.90%        | 12.69%       |
| <b>Proposed (main+aux+ortho)</b> | ✓                           | ✓                          | ✓                            | 0.32%              | 0.92%        | <b>3.67%</b> | <b>14.39%</b> | <b>8.84%</b> |

Table 2: EER (%) of the baseline and proposed system on each evaluation dataset without data augmentation.

| System   | 19LA         | 21LA         | 21DF         | ASV5          | ITW           |
|----------|--------------|--------------|--------------|---------------|---------------|
| Baseline | <b>0.29%</b> | <b>4.78%</b> | 9.89%        | 26.81%        | 22.40%        |
| Proposed | 0.37%        | 5.66%        | <b>8.10%</b> | <b>23.15%</b> | <b>15.72%</b> |

Table 3: EER (%) on each dataset with RawBoost augmentation for the baseline and proposed systems, together with reported results of other methods.

| System         | 19LA         | 21LA         | 21DF         | ASV5          | ITW          |
|----------------|--------------|--------------|--------------|---------------|--------------|
| Baseline [16]  | 0.22%        | <b>0.82%</b> | 6.64%        | 16.25%        | 11.22%       |
| ProtoMAML [20] | 1.35%        | 6.39%        | 6.05%        | —             | 16.29%       |
| SLIM [21]      | <b>0.20%</b> | —            | 4.40%        | —             | 12.50%       |
| Proposed       | 0.32%        | 0.92%        | <b>3.67%</b> | <b>14.39%</b> | <b>8.84%</b> |

important role in out-of-distribution generalization.

Table 2 and Table 3 compare the baseline and proposed systems without and with RawBoost augmentation. Without augmentation, the proposed method yields slightly higher EERs than the baseline on the VCTK-based corpora 19LA and 21LA, but achieves clear improvements on the more challenging cross-corpus evaluation sets 21DF, ASV5, and ITW, reducing the average EER from 12.83% to 10.60%. With RawBoost (LnL & ISD) augmentation, the proposed system maintains almost the same EER as the baseline on 19LA and 21LA, while further reducing EERs on 21DF (6.64% to 3.67%), ASV5 (16.25% to 14.39%), and ITW (11.22% to 8.84%), indicating improved robustness to unseen attacks and out-of-distribution domains. For comparison with other approaches designed to improve robustness to unseen attacks and cross-corpus conditions, Table 3 also includes ProtoMAML [20], which keeps the wav2vec 2.0 encoder frozen and uses meta-learning for few-shot adaptation on each corpus with a small support set, and SLIM [21], which exploits SSL-based speech representations with a style-linguistics mismatch framework using a two-stage pipeline and additional external real-speech corpora. In contrast, our approach keeps the SSL-AASIST backbone and training setup largely unchanged and adds only a lightweight orthogonality regularization, yet still reduces EERs compared to SLIM and ProtoMAML, indicating that our proposed method is a competitive strategy for out-of-distribution robustness.

Fig. 2 presents t-SNE [24] visualizations of the common and specific embeddings extracted from the spectral and temporal graphs,  $\mathcal{G}_s$  and  $\mathcal{G}_t$ , on the 21DF, ASV5, and ITW evaluation sets (19LA and 21LA are omitted due to page limitation). For 21DF and ASV5, up to 300 utterances per spoofing algorithm,

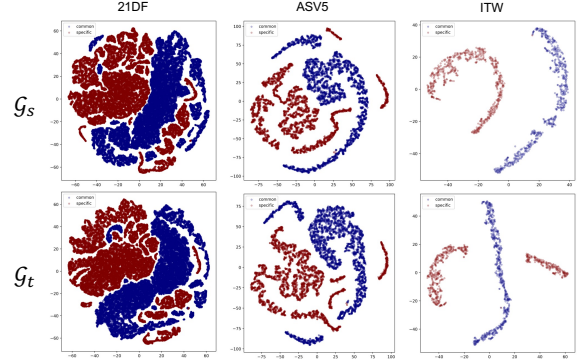


Figure 2: t-SNE visualization of common (blue) and specific (red) embeddings from the spectral graph  $\mathcal{G}_s$  (top) and temporal graph  $\mathcal{G}_t$  (bottom) on 21DF, ASV5, and ITW. Similar separation patterns are also observed on 19LA and 21LA.

together with bonafide samples, are randomly selected, while for ITW we randomly sample up to 300 bonafide and 300 spoof utterances since attack-type labels are not available. Blue markers indicate common embeddings and red markers indicate specific embeddings. Although both embeddings are derived from the same input utterances, they form clearly separated clusters in the two-dimensional space. This observation indicates that the orthogonality constraint effectively drives the two representations toward different directions in the feature space, resulting in distinct regions.

## 5. Conclusions

In this work, we improved the robustness of speech deep-fake detection to unseen attacks and domain shifts by integrating an orthogonality-based domain generalization scheme into the SSL-AASIST backbone, where a common branch encodes domain-invariant authenticity cues and a specific branch captures attack-type information under a cosine-similarity-based loss. Beyond a simple architectural modification, our contribution lies in systematically validating that incorporating domain generalization into a representative SSL-AASIST framework effectively enhances cross-corpus robustness while preserving the original model structure. Experiments on ASVspoof 2019 LA, 2021 LA, 2021 DF, ASVspoof 5, and In-the-Wild showed consistent EER reductions on cross-corpus evaluation sets (21DF, ASV5, ITW) compared to the baseline. This improvement can be attributed to orthogonality-based feature decorrelation, which promotes domain-invariant authenticity representations and suppresses domain-dependent shortcuts, leading to more reliable decisions under unseen conditions.

## 6. Acknowledgements

This work was supported by KOITA grant funded by MSIT(KOITA20250002-38) and research funding from the Incheon National University Artificial Intelligence & Big Data Center in 2025.

## 7. Generative AI Use Disclosure

Generative AI tools were used for language editing assistance under the authors' supervision.

## 8. References

- [1] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.
- [2] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. 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.
- [3] X. Wang, H. Delgado, H. Tak, J.-W. Jung, H.-J. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. Kinnunen *et al.*, "Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale," *arXiv preprint arXiv:2408.08739*, 2024.
- [4] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Proc. Odyssey*, vol. 2016, 2016, pp. 283–290.
- [5] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.
- [6] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech*, 2015, pp. 2087–2091.
- [7] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *Proc. ICASSP*. IEEE, 2021, pp. 6369–6373.
- [8] H. Tak, J.-W. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *arXiv preprint arXiv:2107.12710*, 2021.
- [9] 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 *Proc. ICASSP*. IEEE, 2022, pp. 6367–6371.
- [10] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *Proc. ICLR*, 2018.
- [11] M. Ilse, J. M. Tomczak, and P. Forré, "Selecting data augmentation for simulating interventions," in *Proc. ICML*. PMLR, 2021, pp. 4555–4562.
- [12] E. Rosenfeld, P. Ravikumar, and A. Risteski, "The risks of invariant risk minimization," in *Proc. ICLR*, vol. 9, 2021.
- [13] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does Audio Deepfake Detection Generalize?" in *Proc. Interspeech*, 2022, pp. 2783–2787.
- [14] V. Piratla, P. Netrapalli, and S. Sarawagi, "Efficient domain generalization via common-specific low-rank decomposition," in *Proc. ICML*. PMLR, 2020, pp. 7728–7738.
- [15] S.-Y. Jo and S. W. Yoon, "Poem: polarization of embeddings for domain-invariant representations," in *Proc. AAAI*, vol. 37, no. 7, 2023, pp. 8150–8158.
- [16] H. Tak, M. Todisco, X. Wang, J.-W. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *arXiv preprint arXiv:2202.12233*, 2022.
- [17] 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.
- [18] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proc. ICCV*, 2017, pp. 5542–5550.
- [19] C. Fang, Y. Xu, and D. N. Rockmore, "Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias," in *Proc. ICCV*, 2013, pp. 1657–1664.
- [20] I. Kukanov, J. Laakkonen, T. Kinnunen, and V. Hautamäki, "Meta-learning approaches for improving detection of unseen speech deepfakes," in *Proc. SLT*. IEEE, 2024, pp. 1173–1178.
- [21] Y. Zhu, S. Koppiseti, T. Tran, and G. Bharaj, "Slim: Style-linguistics mismatch model for generalized audio deepfake detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 67 901–67 928, 2024.
- [22] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.
- [23] 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 *Proc. ICASSP*. IEEE, 2022, pp. 6382–6386.
- [24] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 11, 2008.