Variational Regularization for End-to-end Speech Deepfake Detection

Siqing QIN*, Kong Aik LEE*, Man-Wai MAK*, Pasquale LISENA†, Massimiliano TODISCO†
* Dept. of Electrical and Electronic Engineering
The Hong Kong Polytechnic University, Hong Kong SAR, China
† EURECOM, Sophia Antipolis, France
Email: siqing.qin@connect.polyu.hk

***Abstract*—Current research in end-to-end speech deepfake detection predominantly centers around inputting "raw" waveforms to a deep architecture, such as RawNet2, and training the deep neural network to predict if the waveforms are fake. However, direct processing of waveforms could cause over-parameterization in the network, reducing its generalizability. To overcome this limitation, we propose a multi-level variational regularization framework integrating a modified Variational Autoencoder (VAE) with discriminative constraints. Specifically, we adopt an VAE with a deepfake discrimination constraint to regularize a RawNet2-based high-level feature map (HFM) extractor. Experimental results show that the proposed variational regularization leads to HFM features that improve the performance of AASIST, SE-Rawformer, and RawBMamba by 36.01%, 10.07%, and 6.35%, respectively.**

I. Introduction

Speech deepfakes are a type of AI-generated content created by manipulating existing or synthesizing new speech data from scratch to deceive humans. Recent advancements in speech deepfakes have made it nearly impossible for the naked ear to distinguish between fake and genuine speech. Consequently, speech deepfake detection—the process of determining whether a speech signal is naturally uttered or artificially generated—has gained significant attention [1]–[3]. Existing solutions predominantly adopt one of two main approaches. The first approach employs a front-end feature extractor paired with a back-end classifier [4], [5]. The second approach utilizes end-to-end models, simultaneously optimizing feature extraction and classification by directly processing raw audio waveforms [6], [7].

Research on end-to-end models for speech deepfake detection is progressing rapidly. A pioneering approach, RawNet2 [6], employs a time-domain convolution on raw audio to capture subtle artifacts overlooked by traditional spectral methods. Building upon the RawNet2's encoder architecture, subsequent studies have explored innovative designs. For example, AASIST [8], [9] leverages graph attention networks to model the non-Euclidean relationships between time-frequency nodes. Similarly, SE-Rawformer [10] enhances RawNet2 capability by integrating squeeze-and-excitation blocks with Transformer layers, enabling dynamic channel recalibration for raw signal processing. Another variant, RawBMamba [11], replaces the transformer layer in Rawformer with a bidirectional state space model (SSM) to improve sequential modeling efficiency. These end-to-end methods show better performance than the classical approaches [4], [5]. Consequently, we adopt an end-to-end framework in our paper.

Current research on detecting speech deepfakes shows promise but still struggles with poor generalization, limiting its practical applications due to the variations in speech quality and style [12]–[14]. While data augmentation, self-supervised learning, and domain-adaptation could improve cross-domain performance [15]–[17], the use of generative models for domain adaptation in anti-spoofing systems is still underexplored. This gap prompts an investigation into integrating traditional generative models with end-to-end classification frameworks.

Generative models, such as variational encoders (VAEs) [18], have proven to be effective for feature regularization [19], [20], particularly in addressing domain mismatches between training and evaluation data and in enhancing overall generalization capability. Prior studies [21] have assessed the Gaussianization effects of deep normalization flow (GM-DNF) [22], maximum likelihood deep normalization flow (ML-DNF), and VAE normalization methods in out-of-domain contexts. The study highlights the significance of normalization and regularization in the extraction of speaker embeddings.

DNF and VAE are both generative models but utilize different regularization mechanisms. While DNF employs reversible transformation chains to achieve Gaussianization, VAE explicitly regularizes via the evidence lower bound (ELBO), which is more interpretable and can be adapted across different models. Recent findings suggest that incorporating discriminative information through the class-related structures of the learned latent space can improve the regularization performance in DNF [21], [22]. However, the impact of discriminative information on the latent space learned through VAE regularization remains understudied. Moreover, these investigations have predominantly concentrated on the probabilistic linear discriminative analysis (PLDA) back-end instead of end-to-end models. Hence, exploring whether the Gaussianization can improve the generalization capacity of end-to-end anti-spoofing models remains an open question.

In [23], an VAE is used in conjunction with a Wav2Vec frontend to retain the most informative feature for spoofing detection. In contrast to the variational information bottleneck approach in [23], we advocate the integration of variational regularization alongside an auxiliary discriminative decision

boundary. This paper introduces a novel form of variational regularization, grounded in VAE, to enhance the generalization of end-to-end speech deepfake detection. The primary contributions of our proposed method are summarized as follows:

- We undertake a comprehensive investigation into the benefits of using VAE divergence for model regularization to enhance generalization capability. We show that, during the High-level Feature Maps (HFM) extraction process, the Kullback-Leibler (KL) divergence in VAE can facilitate the Gaussianization of latent vectors.
- To ensure class separation in latent space while regularizing the distributions of individual classes, we integrate discriminative information. By establishing appropriate decision boundaries, the model can effectively balance the processes of generation and classification, alleviating issues related to posterior collapse.
- We provide experimental evidence for effectively integrating the VAE regularization into mainstream end-to-end models in deepfake detection systems. Our results show that the proposed method significantly outperforms leading raw audio-input models, thereby enhancing overall accuracy.

II. RawNet2 Encoder Backbone

There is a growing trend in using end-to-end models, such as RawNet2 and its variants [6], to detect spoofed speech. RawNet2 has three core innovations. First, it uses a parameter-sharing sinc-convolution layer to reduce computational complexity while maintaining performance; second, it leverages the frequency-channel interdependency through the squeeze-and-excitation (SE) operations in the ResNet blocks to enhance the network's robustness to acoustic noise; and, third, it uses hierarchical feature learning to derive HFM directly from waveforms. This paper considers several advanced models, including AASIST [8], RawFormer [10], and RawBmamba [11]. All of these models are equipped with a RawNet2 encoder.

The RawNet2 encoder employs a sinc layer to function as a band-pass filter, as shown in Figure 1. This layer generates Low-level short-range Feature Map (LFM), denoted as $\boldsymbol{F}_{\text{LFM}} \in \mathbb{R}^{F \times T}$, where F is the number of frequency bins and T represents the temporal duration. Then, the LFM is considered as an image and a sequence of ResNet blocks is applied to extract High-level Feature Maps (HFM), denoted as $\boldsymbol{F}_{\text{HFM}} \in \mathbb{R}^{C \times F' \times T'}$. Here, C is the number of output channels, $F' < F$ reflects the reduced frequency resolution after the Conv2D and pooling operations, and $T' < T$ indicates the reduced temporal resolution. Each ResNet block incorporates squeeze-and-excitation operations to enhance feature discrimination. Finally, the HFM undergoes flattening along both the time and frequency axes. This process yields a two-dimensional feature sequence $\boldsymbol{F}_s \in \mathbb{R}^{C \times F'T'}$, which is passed to the downstream modules.

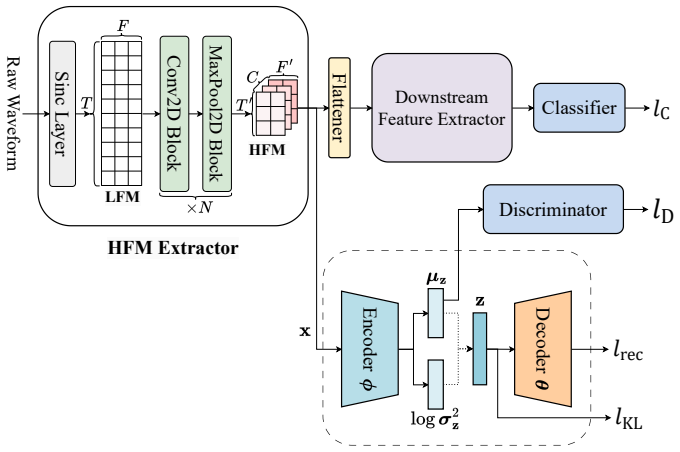

Fig. 1. Diagram illustrating the idea of multi-level VAE regularization. T and F denote the input's temporal duration and number of frequency bins; C, F', and T' represent the number of channels, reduced number of frequency bins, and reduced temporal length. The encoder $\boldsymbol{\phi}$ outputs the latent variable $\mathbf{z}$'s mean $\boldsymbol{\mu}_{\text{z}}$ and log-variance $\log \boldsymbol{\sigma}_{\text{z}}^2$ of the Gaussian distribution from which the latent vector $\mathbf{z}$ is sampled, and the dotted arrow represents the sampling processing (in (1)). The decoder reconstructs the input $\mathbf{x} \in \mathbb{R}^{C \times F' \times T'}$. The model optimizes four losses: reconstruction loss l_{rec}, KL divergence loss l_{KL}, and binary classification losses l_{D} and l_{C} for spoofed/bonafide detection. Frame-level regularization includes all modules, whereas class-conditional regularization excludes the decoder and nullifies l_{rec}.

III. Multi-Level Regularization

This section presents a variational regularization method for the RawNet2 encoder by adding a KL divergence term into the loss function. Inspired by the variational autoencoder (VAE) [18], we propose a multi-level variational regularization framework that synergizes latent space alignment with discriminative objectives. For regularization at the frame-level, we leverage the VAE's reconstruction term to ensure that the latent variables follow a distribution capable of reconstructing the HFM, preserving the fine-grained acoustic details essential for capturing subtle deepfake artifacts. For regularization at the class-level, we remove the reconstruction constraint and optimize the latent space to generate embeddings that are compact within the genuine and spoof classes while maximizing their separation through KL-driven alignment.

A. VAE Regularization

VAE regularization is an effective approach to avoid overfitting in the front-end feature extractor, especially when the extractor is used with a PLDA back-end [20]. In essence, an VAE defines the latent space through its encoder and decoder. The encoder parameterized by $\boldsymbol{\phi}$ maps an input $\mathbf{x}$ to a Gaussian distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ that approximates the posterior distribution $p(\mathbf{z}|\mathbf{x})$. The decoder maps a simple distribution $p(\mathbf{z})$ to a complex distribution $p(\mathbf{x})$. These mappings can be expressed as follows:

$$p(\mathbf{z}|\mathbf{x}) \approx q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\text{z}}(\mathbf{x}), \boldsymbol{\sigma}_{\text{z}}^2(\mathbf{x})), \tag{1}$$

$$p(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \tag{2}$$

with ϕ and $\boldsymbol{\theta}$ representing the encoder and decoder parameters, respectively, and $\boldsymbol{\mu}_{\mathrm{z}}(\mathbf{x})$ and $\boldsymbol{\sigma}^2_{\mathrm{z}}(\mathbf{x})$ represent the mean and variance functions provided by the encoder. $p(\mathbf{z})$ is the prior distribution of $\mathbf{z}$.

Previous work on VAE regularization for speaker verification [24], [25] aligns the latent distribution with a Gaussian distribution because PLDA requires that the i-vectors follow a Gaussian distribution. However, the unsupervised nature of VAE may compromise class separability. In particular, the absence of label-aware constraints could lead to latent space congestion [21], thereby degrading discriminative performance.

To regularize the distributions of individual classes and maintain their separation, we propose a novel approach that allows the HFM extractor to enforce class separation via discriminative constraints. This key distinction facilitates better differentiation between classes in the latent space. Specifically, by utilizing the encoder output $\boldsymbol{\mu}_{\mathrm{z}}(\mathbf{x})$ as the input to a discriminator (binary classifier), we explicitly regularize the first moment of the latent distribution. This design enforces discriminative structure in the latent space, ensuring that class centroids are maximally separated while maintaining a well-regularized learning. Such learning not only stabilizes end-to-end training by preventing latent space collapse but also aligns the feature extractor's outputs with the downstream classifier's inductive biases, thereby harmonizing generative and discriminative objectives within a unified framework.

As shown in Figure 1, the discriminator is expressed as follows:

$$p(y|\mathbf{x}) = \mathrm{softmax}(\mathbf{W}\boldsymbol{\mu}_{\mathrm{z}}(\mathbf{x}) + \mathbf{b}), \tag{3}$$

where y denotes the class label (bonafide/spoofed), and $\mathbf{W}$ and $\mathbf{b}$ represent the parameters of the fully-connected (FC) layer. We utilize cross-entropy in the loss function, as follows:

$$l_{\mathrm{D}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2} y_{i,c}\log p(y_i|\mathbf{x}_i), \tag{4}$$

where N is the number of samples in a minibatch and $y_{i,c} = 1$ when $\mathbf{x}_i \in$ Class c; otherwise, $y_{i,c} = 0$.. The minimization of l_{D} will cause the encoder to produce latent means $\boldsymbol{\mu}_{\mathrm{z}}(\mathbf{x})$ that form two distinct groups according to the class labels.

B. Frame-Level VAE Regularization

The training algorithm of VAE optimizes the model parameters ($\boldsymbol{\theta}$ and ϕ) by maximizing the evidence lower bound (ELBO), which is accomplished by minimizing the loss function

$$\mathcal{L}(\boldsymbol{\theta}, \phi) = -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] + \mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \tag{5}$$

where $p(\mathbf{z})$ is the prior distribution of $\mathbf{z}$. The first and second term in (5) encompasses the negative reconstruction error and the KL divergence, respectively. Due to the constraints imposed by the reconstruction error, the regularizer has the potential to reconstruct the temporal and frequency structure of the input $\mathbf{x} \in \mathcal{R}^{C\times F'\times T'}$. Consequently, the model will learn the mappings at the frame level.

The reparameterization trick is used to sample the latent $\mathbf{z} \in \mathcal{R}^L$ as follows:

$$\mathbf{z} = \boldsymbol{\mu}_{\mathrm{z}}(\mathbf{x}) + \boldsymbol{\sigma}_{\mathrm{z}}(\mathbf{x}) \odot \boldsymbol{\epsilon}, \tag{6}$$

where $\boldsymbol{\epsilon}$ is a random vector drawn from a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\odot$ is the elementwise multiplication, and L is the dimension of the latent space. Assuming Gaussian prior for $\mathbf{z}$, the KL divergence in (5) can be evaluated as

$$\begin{aligned} l_{\mathrm{KL}} &= \mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \\ &= -\frac{1}{2}\sum_{j=1}^{L}(1 + \log(\sigma^2_{\mathrm{z},j}(\mathbf{x})) - \mu^2_{\mathrm{z},j}(\mathbf{x}) - \sigma^2_{\mathrm{z},j}(\mathbf{x})). \end{aligned} \tag{7}$$

The reconstruction term l_{rec} in Figure 1 can be implemented by mean square error (MSE). The downstream classification loss l_{C} is the cross-entropy (CE) loss on detecting fake/bonafide.

The loss function with frame-level VAE regularization is

$$\mathcal{L}_{\text{frame}} = \alpha \times l_{\mathrm{C}} + \frac{1-\alpha}{2} \times [(l_{\mathrm{rec}} + \beta \times l_{\mathrm{KL}}) + l_{\mathrm{D}}], \tag{8}$$

where $\alpha \in [0, 1]$ controls task focus (regularization vs. classification), and $\beta > 0$ adjusts the regularization intensity. This formulation enables simultaneous learning of discriminative features while maintaining reconstruction capability at the frame level.

C. Class-Conditional VAE Regularization

We propose a class-conditional variant of VAE regularization that focuses on the class discriminability of the latent variables $\mathbf{z}$'s. To this end, we remove the decoder in Figure 1 from the VAE. By eliminating the reconstruction constraint, the HFM extractor is more discriminative but not over-trained on a specific dataset, as the KL divergence term ensures the penalization of excessive separation of the two classes.

The removal of the decoder alters the regularization mechanism. Specifically, the cross-entropy loss l_{D} explicitly maximizes decision margins to separate the distributions of the two classes. Meanwhile, the KL divergence term l_{KL} compresses each class's latent distribution toward a shared prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. This trick realizes discriminability and avoids over-fitting at the same time. So, the loss function is

$$\mathcal{L}_{\text{class}} = \alpha \times l_{\mathrm{C}} + \frac{1-\alpha}{2} \times (\beta \times l_{\mathrm{KL}} + l_{\mathrm{D}}). \tag{9}$$

We postulate that the absence of a reconstruction process empowers the regularizer to better optimize the HFM extractor in generating embeddings while effectively distinguishing between different classes. Meanwhile, the KL divergence regularization avoids intra-class compactness and reduces inter-class separation, which helps avoid over-fitting. Therefore, the learned latent space prioritizes class discrimination, rather than the individual reconstruction of each inputted vector $\mathbf{x}$.

TABLE I
VAE ENCODER MODEL ARCHITECTURE. CONV DENOTES A CONVOLUTIONAL OPERATION. T: THE NUMBER OF TIME FRAMES. F: THE NUMBER OF FREQUENCY BINS. $M = 32$.

Layer	Input shape	Stride size
Conv1	$C \times F \times T$	2×2
Conv2	$M \times F/2 \times T/2$	2×2
Conv3	$2M \times F/4 \times T/4$	2×2

TABLE II
IMPACT OF THE VALUES OF THE HYPERPARAMETER β ON THE PERFORMANCE ON THE 19LA, 21LA, AND 21DF DATASETS. HERE, † IS THE REPRODUCED RESULTS IN [11]. THE CLASS-CONDITIONAL (IN (9)) VAE REGULARIZATION WAS APPLIED IN THIS EXPERIMENT.

Models	Beta β	19LA		21LA		21DF
		EER(%)	t-DCF	EER(%)	t-DCF	EER(%)
AASIST [8]	-	0.93	0.0285	10.51	0.4884	-
	2	**0.90**	**0.02522**	7.61	0.3967	21.85
	4	1.35	0.0347	**6.40**	**0.3632**	21.64
	6	0.94	0.0260	6.80	0.3753	**20.89**
SE-Rawformer† [10]	-	1.15	0.0314	4.31	0.2851	20.26
	2	0.91	0.0251	**3.28**	0.2673	18.93
	4	0.99	0.0315	3.81	0.2673	19.44
	6	**0.76**	**0.0232**	4.48	0.3102	**17.63**
RawBMamba [11]	-	1.19	0.0360	3.28	0.2709	15.85
	2	**1.05**	**0.0276**	3.37	0.2726	**14.59**
	4	1.10	0.0342	6.67	0.3444	19.67
	6	1.14	0.0343	**3.06**	0.2632	18.04

IV. EXPERIMENTS AND ANALYSIS

A. Experimental Settings

*a) **Dataset**:* We evaluate the effectiveness and generalizability of the proposed VAE regularization on the ASVspoof datasets [1], [2], specifically, ASVspoof2019 LA (19LA), ASVspoof2021 LA (21LA), and ASVspoof2021 DF (21DF). The 19LA dataset comprises two types of spoofing attacks, namely Text-to-Speech (TTS) and Voice Conversion (VC), implemented across 19 distinct algorithms (A01-A19). The 21LA dataset contains both genuine and artificially generated speech transmitted through telephony systems, including voice over IP and the public switched telephone network. Meanwhile, the 21DF dataset features bonafide and spoofed audio samples that have been modified by various media codecs, which introduce distortion during the processes of encoding, compressing, and decoding.

*b) **Metrics**:* Performance metrics used in this study include the equal error rate (EER) and the minimum tandem detection cost function (min t-DCF) [26].

*c) **Models and architectures**:* The model structure of the HFM extractor and the downstream feature extractor adheres to the original configurations of AASIST (**Baseline 1**) [8], SE-Rawformer (**Baseline 2**) [10], and RawBMamba (**Baseline 3**)[11]. These models serve as the baselines in this paper. The settings for the VAE variants are detailed in Table I. The latent dimension was set to 64.

*d) **Training configuration**:* During the training phase, we utilized input waveform that comprises $64,000$ time points, roughly equivalent to 4 seconds. And α in (8) and (9) was set to 0.7. The Adam optimizer [27] was employed, with a training batch size of 32. Models were trained on the combination of ASVspoof 2019 LA training and development sets, following the training settings in [10], [11], using a single RTX 4090 GPU.

TABLE III
PERFORMANCE ON THE 19LA, 21LA, AND 21DF DATASETS. HERE, † IS THE REPRODUCED RESULTS IN [11]. "FRAME-LEVEL" AND "CLASS-CONDITIONAL" MEAN THAT THE FRAME-LEVEL (IN (8)) AND CLASS-CONDITIONAL (IN (9)) VAE REGULARIZATIONS WERE APPLIED, RESPECTIVELY. - MEANS THIS METHOD IS NON-APPLICABLE (FOR BASELINES ONLY).

Models	VAE Regularization	Beta β	19LA		21LA		21DF
			EER(%)	t-DCF	EER(%)	t-DCF	EER(%)
AASIST [8]	-	-	0.93	0.0285	10.51	0.4884	-
	Frame-level	6	0.94	0.0293	**6.38**	**0.3630**	23.01
	Class-conditional	6	0.94	**0.0260**	6.80	0.3753	**20.89**
SE-Rawformer†	-	-	1.15	0.0314	4.31	0.2851	20.26
	Frame-level	2	0.92	0.0263	**2.72**	**0.2572**	19.64
	Class-conditional	2	**0.91**	**0.0251**	3.28	0.2673	**18.93**
RawBMamba [11]	-	-	1.19	0.0360	3.28	0.2709	15.85
	Frame-level	2	1.18	0.0354	**2.97**	**0.2585**	14.93
	Class-conditional	2	**1.05**	**0.0276**	3.37	0.2726	**14.59**

*e) **Evaluation configuration**:* In previous research, various datasets have been considered as representing different domains because of the varying compression and transmission conditions [17], [28], [29]. To assess the generalization capacity of models, we define 19LA as an **in-domain setting** due to its consistent acoustic conditions and speech codecs. In contrast, 21LA involves changes in the acoustic environment and transmission conditions, representing a shift in acoustic conditions. Furthermore, 21DF alters the speech codecs, categorizing it as a codec shift. Consequently, these two datasets serve as our **cross-domain settings**, where domain shifts occur due to alterations in acoustic conditions or codecs. To fairly assess the model's generalization performance, we selected the epoch in which the models performed the best on the ASVspoof 2019 LA evaluation set to do the cross-domain evaluation. This approach ensures that we are evaluating the models at their optimal performance level of the source domain.

B. Analysis of the Hyperparameter Settings

This section discusses the selection of the hyperparameter β in (8) and (9), as detailed in Table II. Each framework exhibits an optimal value of β for different evaluation sets. Specifically, in the remaining experiments, we selected β values of **6**, **2**, and **2** for **AASIST**, **SE-Rawformer**, and **RawBMamba**, respectively, to achieve the best generalization capacity while maintaining strong in-domain performance.

C. Results of Frame-Level Regularization

This section evaluates the effectiveness of the proposed frame-level regularization method by applying it to various baselines, as detailed in Table III. In this experiment, the VAE regularization contains the reconstruction loss.

The frame-level regularization exhibits enhanced performance on unseen datasets. Notably, when assessed on the

TABLE IV
ABLATION STUDY OF THE PROPOSED FRAME-LEVEL (**FL**) VAE REGULARIZATION AND CLASS-CONDITIONAL (**CC**) VAE REGULARIZATION ON THE 19LA, 21LA, AND 21DF DATASETS. HERE, † IS THE REPRODUCED RESULTS IN [11]. "**DI**" IS SHORT FOR DISCRIMINATIVE INFORMATION. ✓ AND × REPRESENT WHETHER THE CORRESPONDING METHOD IS IMPLEMENTED OR NOT.

Models	VAE Regularization			19LA		21LA		21DF
	FL	CC	DI	EER(%)	t-DCF	EER(%)	t-DCF	EER(%)
SE-Rawformer†	✓	×	✓	0.92	0.0263	2.72	0.2572	19.64
	×	✓	✓	0.91	0.0251	3.28	0.2673	18.93
	✓	×	×	1.09	0.0326	6.86	0.2957	21.27
	×	✓	×	0.63	0.0179	3.29	0.2730	19.21
	×	×	✓	0.87	0.0258	4.82	0.3091	19.83
RawBMamba[11]	✓	×	✓	1.18	0.0354	2.97	0.2585	14.93
	×	✓	✓	1.05	0.0276	3.37	0.2726	14.59
	✓	×	×	1.11	0.0372	3.96	0.2957	16.46
	×	✓	×	0.94	0.0301	3.35	0.2630	16.91
	×	×	✓	1.07	0.0316	3.84	0.2786	18.24

19LA dataset, which belongs to the same domain as the training set, it performs nearly identically to **Baseline 1**. However, it demonstrates superior generalizability when tested on the 21LA dataset, achieving a 36.01% relative reduction in average EER.

In the case of **Baseline 2**, the proposed frame-level regularization consistently displays better generalizability across all evaluation sets. It results in an overall 9.45% relative reduction in average EER. When compared to **Baseline 3** in prior studies, our method also shows performance improvement across all evaluation sets, with a 6.06% relative reduction in average EER.

The proposed frame-level variational regularization shows improved performance across all three baselines. Notably, generalization is enhanced to a lesser extent when there is a change in codec compared to shifts in acoustic conditions. This suggests that frame-level variational regularization is particularly effective in managing varying acoustic conditions. We reason that the reconstruction loss in frame-level VAE helps the model become more robust to acoustic perturbations by preserving finer temporal-frequency structure.

D. Results of Class-Conditional Regularization

This section discusses the benefits of the proposed class-conditional regularization method. In this experiment, the reconstruction process is eliminated by invalidating the decoder in VAE.

As indicated in Table III, there is a slight improvement in t-DCF on the 19LA dataset compared to **Baseline 1**. However, the generalizability is significantly enhanced on the 21LA and 21DF datasets in relation to both **Baseline 1** and the frame-level regularized model.

In terms of **Baseline 2**, the model performance shows a relative reduction of 10.07% in average EER. Interestingly, there is a decrease in EER on the 21DF dataset compared to the frame-level regularization, despite a slight performance decline on the 21LA dataset. The class-conditional method on RawBMamba witnesses the same trend with a relative EER reduction of 6.35%. This indicates an improved generalization ability of the proposed model when evaluated on data from a completely different domain.

Compared to frame-level regularization, the proposed approach demonstrates better performance in addressing codec shifts across three different models. This suggests that the class-conditional variant emphasizes utterance-level generalization, which may be more helpful under codec variability.

E. Ablation Study

This section examines the effectiveness of frame-level VAE regularization, class-conditional VAE regularization, and the discriminative branch.

*a) **The Effectiveness of Discriminative Information**:* We evaluated the impact of discriminative information by removing the discriminative term l_{D} from the frame-level (FL) and class-conditional (CC) regularization approaches. As shown in Table IV, excluding discriminative information in VAE regularization resulted in reduced cross-dataset performance, highlighting an overfitting issue, evidenced by the increased EER on the ASVspoof 2019 LA evaluation set, representing an in-domain scenario. The findings confirm the effectiveness of integrating discriminative information into VAE regularization. It is noteworthy that this ablation experiment led to posterior collapse within the VAE, indicating that the synergy between discriminative and generative learning can help mitigate posterior collapse.

*b) **The Effectiveness of Multi-Level Regularization**:* To validate the effectiveness of the proposed FL regularization, we disabled the KL divergence l_{KL} and reconstruction term l_{rec} in (8). The outcomes, presented in Table IV, reveal a decline in generalization across cross-domain settings without FL regularization. This suggests that without FL regularization, the discriminative term applied to the HFM extractor fails to enhance the model's generalizability. Besides, disabling the KL divergence l_{KL} in (9) resulted in a noticeable decline in generalization performance on the 21LA and 21DF datasets. Through these comprehensive experiments, we demonstrate that the proposed methods effectively address the generalization problem. Discriminative information strengthens VAE regularization by preventing posterior collapse and facilitating class separation within the latent space.

V. CONCLUSION

This paper proposes a novel variational regularization framework to enhance the generalization capability of end-to-end models for speech deepfake detection. Experimental results demonstrate that frame-level regularization achieves great performance on acoustic conditions shifts, while class-conditional regularization significantly improves cross-codec generalization. We also investigated the relationship between the weighting factor of KL divergence and the HFM extractor's depth. Ablation studies and analysis confirm the critical roles of discriminative information in model regularization and the effectiveness of the proposed multi-level variational regularization. Future work will explore discriminative Gaussian priors

for latent variables and visualize the regularization's impact on gradients.

REFERENCES

[1] A. Nautsch, X. Wang, N. Evans, *et al.*, "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. DOI: 10.1109/TBIOM.2021.3059479.

[2] J. Yamagishi, X. Wang, M. Todisco, *et al.*, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

[3] X. Wang, H. Delgado, H. Tak, *et al.*, "ASVspoof 5: Crowd-sourced speech data, deepfakes, and adversarial attacks at scale," in *Proc. INTERSPEECH 2024*, 2024, pp. 1–8.

[4] C. Wang, J. Yi, J. Tao, C. Y. Zhang, S. Zhang, and X. Chen, "Detection of cross-dataset fake audio based on prosodic and pronunciation," in *Proc. INTERSPEECH 2023*, 2023, pp. 3844–3848.

[5] J. Xue, C. Fan, J. Yi, *et al.*, "Learning from yourself: A self-distillation method for fake speech detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096837.

[6] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6369–6373. DOI: 10.1109/ICASSP39728.2021.9414234.

[7] 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 *Proc. 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 692–698. DOI: 10.1109/ICPR56361.2022.9956138.

[8] J.-W. Jung, H.-S. Heo, H. Tak, *et al.*, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371. DOI: 10.1109/ICASSP43922.2022.9747766.

[9] K. Borodin, V. Kudryavtsev, D. Korzh, *et al.*, "AASIST3: Kan-enhanced AASIST speech deepfake detection using SSL features and additional regularization for the ASVspoof 2024 challenge," *arXiv preprint arXiv:2408.17352*, 2024.

[10] X. Liu, M. Liu, L. Wang, K. A. Lee, H. Zhang, and J. Dang, "Leveraging positional-related local-global dependency for synthetic speech detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096278.

[11] Y. Chen, J. Yi, J. Xue, *et al.*, "RawBMamba: End-to-end bidirectional state space model for audio deepfake detection," in *Proc. INTERSPEECH 2024*, 2024, pp. 2720–2724.

[12] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Domain generalization via aggregation and separation for audio deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 344–358, 2024. DOI: 10.1109/TIFS.2023.3324724.

[13] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Single domain generalization for audio deepfake detection.," in *Proc. IJCAI 2023*, 2023, pp. 58–63.

[14] M. Yousif, J. J. Mathew, H. Pallan, *et al.*, "Enhancing generalization in audio deepfake detection: A neural collapse based sampling and training approach," *arXiv preprint arXiv:2404.13008*, 2024.

[15] Y. Guo, H. Huang, X. Chen, H. Zhao, and Y. Wang, "Audio deepfake detection with self-supervised WavLM and multi-fusion attentive classifier," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 702–12 706. DOI: 10.1109/ICASSP48485.2024.10447923.

[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," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112–119.

[17] X. Chen, W. Lu, R. Zhang, *et al.*, "Continual unsupervised domain adaptation for audio deepfake detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[19] Y. Zhang, L. Li, and D. Wang, "VAE-based regularization for deep speaker embedding," in *INTERSPEECH 2019*, 2019, pp. 4020–4024.

[20] Z. Benhafid, S. A. Selouani, and A. Amrouche, "Light-SpineNet variational autoencoder for logical access spoof utterances detection in speaker verification systems," in *Proc. 5th International Conference on Bio-engineering for Smart Technologies (BioSMART)*, 2023, pp. 1–4. DOI: 10.1109/BioSMART58455.2023.10162119.

[21] Z. Benhafid, S. A. Selouani, and A. Amrouche, "Deep normalization for light SpineNet speaker anti-spoofing systems," *Multimedia Tools and Applications*, vol. 83, no. 33, pp. 80 261–80 275, 2024.

[22] Y. Cai, L. Li, A. Abel, X. Zhu, and D. Wang, "Maximum gaussianality training for deep speaker vector normalization," *Pattern Recognition*, vol. 145, p. 109 977, 2024. DOI: https://doi.org/10.1016/j.patcog.2023.109977.

[23] T.-P. Doan, H. Dinh-Xuan, T. Ryu, *et al.*, "Trident of poseidon: A generalized approach for detecting deepfake voices," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 2222–2235.

[24] Y. Tu, M.-W. Mak, and J.-T. Chien, "Information maximized variational domain adversarial learning for speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6449–6453. DOI: 10.1109/ICASSP40776.2020.9053735.

[25] Y. Tu, M.-W. Mak, and J.-T. Chien, "Variational domain adversarial learning with mutual information maximization for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2013–2024, 2020. DOI: 10.1109/TASLP.2020.3004760.

[26] T. Kinnunen, K. A. Lee, H. Delgado, *et al.*, "T-dcf: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," *arXiv preprint arXiv:1804.09618*, 2018.

[27] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] H. Zhang, L. Wang, K. A. Lee, M. Liu, J. Dang, and H. Chen, "Learning domain-invariant transformation for speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7177–7181. DOI: 10.1109/ICASSP43922.2022.9747514.

[29] V. Negroni, D. Salvi, A. I. Mezza, P. Bestagini, and S. Tubaro, "Leveraging mixture of experts for improved speech deepfake detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.