OTM-TITANET: LEVERAGING PRE-TRAINED SPEAKER EMBEDDINGS WITH OPTIMAL TRANSPORT MEMORY FOR AUDIO ANTI-SPOOFING

Fei-Yueh Chen
Department of Electrical and Computer Engineering, University of Rochester
fchen27@ur.rochester.edu

ABSTRACT

Existing one-class methods for audio deepfake detection often assume that all bona fide (real) speech is similar, failing to capture the natural diversity of different speakers. To address this, we propose **OTM-TitaNet**, a framework that combines a TitaNet backbone with a Speaker-Agnostic Dual Memory Network. We finetune the TitaNet encoder to detect spoofing artifacts and use Sinkhorn Optimal Transport (OT) to ensure the memory learns diverse acoustic patterns without supervision. Experiments on the ASVspoof 2019 LA dataset show our method achieves an Equal Error Rate (EER) of **1.37%**. Surprisingly, our analysis reveals that this simple architecture outperforms complex models using additional clustering losses (like OC-Softmax), suggesting that a well-regularized memory network is sufficient for effective detection.

1. INTRODUCTION

Recent advances in Text-to-Speech (TTS) and Voice Conversion (VC) allow attackers to synthesize high-quality speech with ease, posing a significant threat to Automatic Speaker Verification (ASV) systems. These Logical Access (LA) attacks require robust countermeasures that can generalize to unseen synthesis algorithms. Prior research has actively explored advanced architectures to capture subtle spoofing artifacts. A prevailing trend involves adapting large-scale Self-Supervised Learning (SSL) models, such as Wav2Vec 2.0 or HuBERT, often coupled with heavy backend classifiers like Graph Neural Networks [1], Conformers [2], or specialized attention mechanisms [3]. While effective, these Transformer-based approaches often incur high computational costs and focus primarily on general speech representations.

In this work, we propose a novel perspective: instead of general speech encoders, we hypothesize that an encoder optimized for *speaker verification*, like the TitaNet architecture [4], is uniquely suited for this task. TitaNet's 1D depth-wise separable convolutions efficiently capture the spectral-temporal dependencies that define a speaker's voice. To our best knowledge, we are the first to investigate fine-tuning TitaNet as a backbone for anti-spoofing, aiming to expose synthesis artifacts as deviations in the robust speaker embedding space.

Furthermore, a critical limitation in current loss function design (e.g., OC-Softmax [5]) is the assumption that all bona fide speech clusters into a single center. Recent findings by Kwok et al. [6] highlight a weak spot in current systems: the lack of diversity in bona fide testing data leads to overestimated performance. They argue that real speech varies significantly (e.g., different environments, speaking styles), and single-center models fail to capture this natural variance. While SAMO [7] attempts to solve this with multi-center learning, it relies on speaker enrollment, which is often unavailable.

To address these gaps, we propose a Speaker-Agnostic Dual Memory Network. We combine the TitaNet backbone with a dual memory bank trained via Sinkhorn Optimal Transport. This allows the model to unsupervisedly learn multiple prototypes for both diverse bona fide speech (addressing the issue raised in [6]) and varied spoofing patterns, without requiring complex Transformers or speaker enrollment.

2. RELATED WORK

2.1 Model Architecture

Early anti-spoofing methods primarily relied on hand-crafted features, such as LFCC and CQCC, combined with Gaussian Mixture Models (GMMs). The advent of deep learning shifted the focus towards end-to-end modeling directly from raw waveforms, with RawNet2 [8] demonstrating the viability of this approach. More recently, attention-based mechanisms have dominated the leaderboard. Rosello et al. [2] proposed a Conformer-based classifier to handle variable-length utterances, leveraging the global context capabilities of self-attention. Similarly, Truong et al. [3] introduced a Temporal-Channel Modeling module to enhance Multi-Head Self-Attention, targeting artifacts in specific time-frequency regions.

Despite their success in detecting unseen attacks, these transformer-based models often incur significant computational costs, as they frequently utilize large-scale pre-trained encoders like Wav2Vec 2.0 (300M parameters) [9]. To address efficiency concerns in the broader speaker recognition domain, TitaNet [4] was introduced as a scalable architecture utilizing 1D depth-wise separable convolutions and Squeeze-and-Excitation layers. With model sizes ranging from Small (10M) to Large (25M), TitaNet offers a significantly more lightweight alternative to massive self-supervised models, though its application has traditionally focused on verification rather than spoofing detection.

2.2 Loss Function Design

To improve generalization against unseen attacks, One-Class Learning has become a standard paradigm. The OC-Softmax loss [5] operates by compacting bona fide embeddings into a single center while pushing spoofed samples away. However, this approach relies on the strong assumption that bona fide speech is homogeneous. Kwok et al. [6] recently challenged this premise, demonstrating that single-center models struggle when bona fide speech deviates from standard "clean read" distributions (e.g., in cross-dataset testing), thereby highlighting the necessity for modeling the diversity within real speech.

To address the issue of intra-class diversity, SAMO [7] introduced a Speaker Attractor Multi-Center One-Class learning framework. By clustering bona fide speech around multiple centers based on speaker identity, SAMO achieves finer-grained decision boundaries. However, a major limitation of this approach is its dependency on speaker enrollment data to define these centers, which restricts its applicability in scenarios where speaker identity is unknown or enrollment data is unavailable.

3. METHODOLOGY

Our framework consists of three integrated components: (1) a fine-tuned TitaNet encoder for speaker-discriminative feature extraction, (2) a Speaker-Agnostic Dual Memory Network inspired by non-parametric instance discrimination [10], and (3) an Optimal Transport regularization mechanism adapting the Sinkhorn-Knopp algorithm from SwAV [11] to ensure prototype diversity.

3.1 TitaNet Encoder

We adopt the pre-trained TitaNet checkpoints [4] as our backbone f_θ. TitaNet utilizes 1D depth-wise separable convolutions with global context attention to efficiently encode variable-length audio into a fixed-dimensional embedding. Given a raw waveform X, the encoder produces a normalized embedding:

$$z = f_\theta(X) \in \mathbb{R}^D, \quad ||z||_2 = 1 \tag{1}$$

where $D = 192$. Unlike standard approaches that freeze pre-trained weights, we fine-tune the entire encoder. This allows the model to shift from a purely speaker-verification space (where spoofing artifacts might be ignored as channel noise) to an anti-spoofing space.

3.2 Dual Memory Banks with Sparse Attention

To capture the multi-modal nature of speech without supervision, we employ external memory modules. Inspired by the memory bank structure proposed by Wu et al. [10] for unsupervised feature learning, we maintain a storage of representations that is decoupled from the mini-batch size. However, distinct from [10] which stores an embedding for every training instance (instance-level), we adapt this to store learnable *prototypes* (cluster-level) that represent canonical acoustic patterns.

We define two distinct memory banks:

$$M_{real} = \{m_{r,1}, ..., m_{r,K}\} \in \mathbb{R}^{K\times D} \tag{2}$$

$$M_{spoof} = \{m_{s,1}, ..., m_{s,K}\} \in \mathbb{R}^{K\times D} \tag{3}$$

where K is the number of prototypes (slots). M_{real} aims to cover the diverse acoustic environments and speaking styles of bona fide speech, while M_{spoof} captures varied attack artifacts.

For an input embedding z, we compute the cosine similarity with all prototypes. To filter out irrelevant prototypes and focus on the most similar acoustic modes, we apply a Sparse Attention mechanism that considers only the Top-k entries:

$$\hat{z}_{real} = \sum_{j\in \text{TopK}(z, M_{real})} \frac{\exp(z \cdot m_{r,j}/\tau)}{\sum_l \exp(z \cdot m_{r,l}/\tau)} m_{r,j} \tag{4}$$

The reconstruction error $E_{real} = ||z - \hat{z}_{real}||^2$ serves as mean square error relative to the bona fide distribution.

3.3 Optimal Transport

A critical challenge in learning discrete prototypes is mode collapse, where the encoder maps all inputs to a small subset of prototypes, leaving the majority of the memory bank unused. This creates a trivial solution where the memory degenerates into a single mean vector.

To resolve this, we adopt the Equipartition Constraint strategy from SwAV [11]. Caron et al. demonstrated that enforcing an equal distribution of data samples across prototypes during training effectively prevents collapse.

We formulate the memory addressing as an Optimal Transport (OT) problem. Let $Z = [z_1, ..., z_B]$ be the batch of embeddings and C be the matrix of prototypes (from either bank). We seek an assignment matrix $Q \in \mathbb{R}^{B\times K}$ that maximizes the similarity between features and prototypes, subject to the constraint that Q is a transport plan distributing samples uniformly. Formally, we optimize Q to maximize $\text{Tr}(Q^T C^T Z) + \varepsilon H(Q)$, where H is the entropy regularization. As proposed in SwAV, the solution is obtained efficiently using the Sinkhorn-Knopp algorithm:

$$Q^* = \text{Diag}(u) \exp(\frac{C^T Z}{\varepsilon}) \text{Diag}(v) \tag{5}$$

where u and v are renormalization vectors computed iteratively.

We then impose an OT Loss that forces the model's predicted softmax probabilities P to match this optimal assignment Q:

$$L_{ot} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{K} Q^*_{ij} \log P_{ij} \tag{6}$$

By integrating this loss, we ensure that both M_{real} and M_{spoof} maintain diverse, active prototypes, enabling the system to generalize across different speakers and attacks without explicit labels.

3.4 Objective Functions

The final training objective combines reconstruction, optimal transport regularization, and an explicit diversity constraint:

$$L_{total} = L_{recon} + \lambda_{ot} L_{ot} + \lambda_{div} L_{div} \tag{7}$$

3.4.1 Dual Reconstruction Loss (L_{recon})

We minimize the reconstruction error for the correct memory bank (Attract) and maximize it for the incorrect one (Repel) using a margin-based hinge loss:

$$\begin{aligned} L_{recon} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} [E_{real}(x) + \max(0, m - E_{spoof}(x))] \\ + \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} [E_{spoof}(x) + \max(0, m - E_{real}(x))] \end{aligned} \tag{8}$$

where $\mathcal{B}$ and $\mathcal{S}$ denote bona fide and spoof samples, and m is the margin.

3.4.2 Diversity Loss (L_{div})

To further ensure that the memory slots are utilized uniformly and to prevent the "dying slot" problem, we maximize the entropy of the averaged attention distribution. Although Sinkhorn OT handles batch-wise distribution, L_{div} acts as a global constraint:

$$L_{div} = -H(\bar{\mathbf{w}}) = \sum_{j=1}^{K} \bar{w}_j \log(\bar{w}_j + \epsilon) \tag{9}$$

where $\bar{\mathbf{w}} = \frac{1}{B} \sum_{i=1}^{B} \mathbf{w}_i$ is the mean attention weight across the batch. Minimizing negative entropy L_{div} encourages a uniform distribution over all K slots.

4. EXPERIMENTAL SETUP

4.1 Dataset and Protocols

We used the official train, dev, and test splits from ASVspoof2019 [12]. The training process was monitored using the development set. specifically, we selected the model checkpoint that achieved the lowest Equal Error Rate (EER) on the validation set for evaluation.

To evaluate robustness against domain shift and unseen attacks, we tested on the ASVspoof2021 LA test set [13]. Following standard practices for fair comparison, we utilized a fixed 4-second cropped version of the dataset.

4.2 Evaluation Metrics

We report performance using two standard metrics established by the ASVspoof challenges:

Equal Error Rate (EER): The EER represents the operating point where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal. FAR is the proportion of spoofing attacks incorrectly accepted as bona fide, while FRR is the proportion of bona fide speech incorrectly rejected.

$$\text{EER} = \text{FAR}(\theta_{eer}) = \text{FRR}(\theta_{eer}) \tag{10}$$

where θ_{eer} is the decision threshold.

Minimum Tandem Detection Cost Function (min t-DCF): While EER assesses the standalone performance of the countermeasure (CM), the t-DCF [**?**] evaluates the impact of the CM on a fixed Automatic Speaker Verification (ASV) system. It considers the costs of different error types (C_{miss}, C_{fa}) and the prior probability of attacks (π_{spoof}).

$$\begin{aligned} \text{t-DCF}(\theta) = C_{miss} P_{miss}(\theta) \pi_{target} \\ + C_{fa} P_{fa}(\theta)(1 - \pi_{target}) \end{aligned} \tag{11}$$

A lower min t-DCF indicates that the CM is more effective in a practical tandem integration scenario.

4.3 Data Augmentation

To improve robustness against channel variations, we applied RawBoost [14] during training. Due to implementation constraints, we utilized Algorithm 2 (Impulsive Signal Dependent Noise) and Algorithm 3 (Stationary Signal Independent Noise), excluding Algorithm 1. This subset of augmentations effectively simulates additive background noise and transmission artifacts.

5. RESULTS

5.1 Overall Result

Table 2 and Table 3 compare our method with existing SOTA systems.On ASVspoof 2019 LA(Table 2), our baseline (1.37%) significantly outperforms the RawNet2 baseline (2.48%) and is competitive with OC-Softmax (1.25%). While SAMO (1.08%) performs slightly better, it requires speaker enrollment. Our method achieves comparable results in a strictly speaker-agnostic setting.

On ASVspoof 2021 LA(Table 3), which features severe channel variability, our TitaNet-Large model achieves an EER of 9.90%. This performance is comparable to RawNet2 (9.50%) but lags behind AASIST (5.59%) and large-scale pre-trained models like XLSR-Conformer (1.38%). This indicates that while our method is effective for detecting logical artifacts, the domain gap in telephony conditions remains a challenge for the finetuned TitaNet encoder compared to massive SSL models.

5.2 Score Distribution Analysis

Table 4 shows the statistics of the output scores. The Baseline system provides the clearest separation between Bonafide (Mean 1.16) and Spoof (Mean -1.00) distributions with relatively low standard deviation. In contrast, the experiment with encoder only (without memory bank) results in a much smaller margin (0.98 vs -0.14), confirming that the Memory Bank significantly enhances the discriminative power.

Table 1. Experimental Results on ASVspoof 2019 and 2021 LA.

Method	2019 EER (%)	2019 min t-DCF	2021 EER (%)
Baseline (Recon + OT)	**1.37**	**0.0412**	11.03
+ OC-Softmax	1.52	0.0438	18.39
+ Multi-Center OC	3.47	0.0714	11.92
+ Contrastive Loss	2.87	0.0510	12.76
+ Large Model	5.26	0.1150	**9.90**
+ Adaptive Margin	3.22	0.0635	10.19
+ Score Fusion	2.47	0.0626	11.02
Larger Memory (128 slots)	1.90	0.0584	10.25
TitaNet + OC (No Memory)	1.70	0.0548	10.50
- OT	5.80	0.0673	11.34

Table 2. Comparison with SOTA on ASVspoof 2019 LA (Eval)

Method	Backbone	EER (%)	min t-DCF
RawNet2 [15]	RawNet2	2.48	-
OC-Softmax [5]	AASIST	1.25	0.0415
SAMO [7]	AASIST	**1.08**	**0.0363**
Ours	TitaNet Small	1.37	0.0412

Table 3. Comparison on ASVspoof 2021 LA (Fixed 4s crop)

Method	2021 EER (%)
RawNet2 [15]	9.50
AASIST [1]	5.59
XLSR-Conformer [16]	1.38
XLSR-Conformer + TCM [3]	**1.03**
Ours (TitaNet Large)	9.90
Ours (TitaNet Small)	11.03

5.3 ASVspoof 2019 LA Performance

Table 1 summarizes the performance of our proposed methods. The Baseline configuration (TitaNet-Small + Dual Memory + OT) achieved the best performance with an EER of **1.37%** and min t-DCF of **0.0412**.

Surprisingly, increasing complexity (Enhanced Mode) did not yield improvements. Adding OC-Softmax (1.52%) and Multi-Center losses (3.47%) degraded performance. This suggests that the unsupervised dual memory mechanism is sufficient for capturing discriminative features, and additional supervised clustering losses may introduce optimization conflicts.

6. DISCUSSION

6.1 The Necessity of Optimal Transport

A critical question in memory-based networks is whether complex regularization like Optimal Transport is truly necessary, or if a simple reconstruction loss would suffice. To address this, we conduct an experiment that we removed the OT regularization and the Diversity Loss, relying solely on reconstruction error.

Table 4. Score Distribution Analysis (ASVspoof 2019 LA)

Method	Bonafide		Spoof	
	Mean	Std	Mean	Std
Baseline	1.16	0.20	-1.00	0.31
+ OC-Softmax	1.34	0.33	-1.60	0.39
+ Contrastive	1.31	0.22	-1.66	0.61
+ Large Model	1.10	0.27	-1.23	0.76
TitaNet Only	0.98	0.11	-0.14	0.20

The results were decisive: removing OT caused the EER to degrade drastically from **1.37%** to **5.80%**. We observed that without the normalization imposed by OT, the model suffered from severe mode collapse, utilizing only a small fraction of the available memory slots. This confirms that OT is not merely an auxiliary component but a fundamental requirement for learning a diverse and effective set of acoustic prototypes in an unsupervised manner.

6.2 Incompatibility of Supervised Clustering Losses

We initially hypothesized that explicitly enforcing global compactness via One-Class (OC) loss functions would complement the local manifold learning of the Memory Network. We tested two variants:

1. **Single-Center OC-Softmax:** Compacting all bona fide embeddings to a single point.
2. **Multi-Center OC-Softmax:** Learning $K = 20$ centers to capture speaker diversity.

The objective function was modified as:

$$L_{oc} = \frac{1}{N}\sum_{i} \log(1 + e^{\alpha(m_{real} - \cos(\mathbf{z}_i, \mathbf{c}))}) \tag{12}$$

$$L_{total} = L_{recon} + \lambda_{oc} \cdot L_{oc} + \lambda_{ot} \cdot L_{ot} + \lambda_{div} \cdot L_{div} \tag{13}$$

As shown in Table 1, both variants degraded performance compared to the Baseline (EER **1.37%**). Specifically, Multi-Center OC-Softmax performed significantly

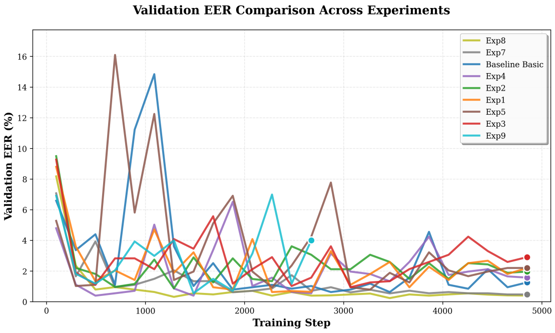

Figure 1. Validation EER curves during training. The oscillation indicates the sensitivity of the decision boundary to the embedding geometry.

worse (EER **3.47%**). We suspect this performance degradation stems from a conflict in optimization objectives. While the reconstruction loss requires the embeddings to retain enough variance to distinguish prototypes, the OC-Softmax loss forces them to collapse into a single point. This contradiction likely prevents the encoder from learning a stable feature space.

Furthermore, in the Multi-Center OC-Softmax setting, without the explicit read-write mechanism of a memory bank or speaker labels (as in SAMO [7]), the learnable centers $\mathbf{c}_k$ struggled to converge to meaningful clusters, likely trapping the encoder in a suboptimal local minimum. This confirms that for unsupervised modeling of bona fide diversity, the baseline is a superior mechanism to simple learnable centers.

6.3 Training Dynamics and Stability

Deep metric learning on small datasets often suffers from training instability. We analyzed the evolution of the Validation EER throughout the training process (Figure 1).

We observed significant fluctuations in the Validation EER, even when the training loss decreased smoothly. For instance, in the Baseline experiment, the EER fluctuated between 1.37% (best) and 1.59% (last epoch). This volatility underscores the disconnect between the training loss and the threshold-dependent EER metric.

6.4 Generalization to the ASVspoof2021

While the Baseline (Small model) was best for 2019, the experiment with TitaNet-Large model as encoder achieved the best performance on the 2021 dataset (9.90% vs 11.03%). This suggests that the larger model capacity helps in generalizing to unknown channel conditions, even if it slightly overfits the source domain (ASVspoof2019). However, compared to SOTA models like XLSR-Conformer + TCM [3] (1.03%), our method still struggles with the extreme domain shift in ASVspoof 2021, highlighting the advantage of large-scale SSL pretraining speech encoders for in-the-wild scenarios.

6.5 Ineffectiveness of Adaptive Margins

While the reconstruction loss L_{recon} requires fixed margin as hyperparameter, it might be too difficult for the model to satisfy early in training. Inspired by curriculum learning, we implemented an *Adaptive Margin Scheduler* to progressively tighten the decision boundary. The margins for bona fide (m_{real}) and spoof (m_{fake}) were formulated as time-dependent functions:

$$m_{real}(t) = 0.7 + \rho(t) \cdot (0.95 - 0.7) \tag{14}$$

$$m_{fake}(t) = 0.3 - \rho(t) \cdot (0.3 - 0.1) \tag{15}$$

where $\rho(t)$ linearly increases from 0 to 1 after a warmup period.

Contrary to expectations, this strategy degraded the EER from **1.37%** (Baseline) to **3.22%**. We analyze that the dynamic shifting of decision boundaries prevents the memory prototypes from stabilizing. Since the Sinkhorn algorithm relies on stable feature distributions to compute optimal assignments, the constantly moving margins disrupted the convergence of the unsupervised clustering process.

6.6 Redundancy in Score Fusion

We explored whether combining the local reconstruction score (S_{mem}) with a global classification score (S_{oc}) could leverage complementary information. To investigate, we implemented a score fusion mechanism using Z-score normalization:

$$S_{final} = w \cdot \frac{S_{oc} - \mu_{oc}}{\sigma_{oc}} + (1 - w) \cdot \frac{S_{mem} - \mu_{mem}}{\sigma_{mem}} \tag{16}$$

where $S_{mem} = E_{real} - E_{spoof}$ and S_{oc} is the output of the OC-Softmax layer.

Although Score Fusion (EER **2.47%**) improved upon the OC-Softmax experiments, it failed to surpass the pure reconstruction Baseline (EER **1.37%**). This indicates that the discriminative information captured by the global OC-Softmax loss is largely redundant to the fine-grained features captured by the Dual Memory banks. Furthermore, the gradient conflict between minimizing reconstruction error (preserving variance) and minimizing OC loss (collapsing variance) during training likely weakened the quality of the TitaNet embeddings, rendering the fusion less effective than the single-stream memory approach.

7. CONCLUSION

In this paper, we proposed OTM-TitaNet, a lightweight and effective framework for audio deepfake detection. By combining a fine-tuned TitaNet encoder with a Dual Memory Network, we can detect attacks without requiring speaker enrollment.

Our experiments highlighted two key findings. First, **Optimal Transport is essential**. Without it, the memory network suffers from mode collapse and fails to learn useful features. Second, **simpler is better**. We found that adding complex supervised losses, such as OC-Softmax, actually degraded performance compared to our baseline (EER 1.37%). This indicates that simple memory reconstruction is robust enough for this task. While our method works well on the ASVspoof 2019 dataset, future work will focus on improving generalization to the unseen channel conditions in the ASVspoof 2021 dataset.

8. REFERENCES

[1] 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," *arXiv preprint arXiv:2110.01200*, 2021.

[2] E. Rosello, A. Gomez-Alanis, A. M. Gomez, and A. Peinado, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," in *Interspeech 2023*, 2023, pp. 5281–5285.

[3] D.-T. Truong, R. Tao, T. Nguyen, H.-T. Luong, K. A. Lee, and E. S. Chng, "Temporal-channel modeling in multi-head self-attention for synthetic speech detection," in *Interspeech 2024*, 2024, pp. 537–541.

[4] N. R. Koluguri, T. Park, and B. Ginsburg, "TitaNet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," *arXiv preprint arXiv:2110.04410*, 2021.

[5] Y. Zhang, F. Jiang, and Z. Duan, "One-Class Learning Towards Synthetic Voice Spoofing Detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[6] C. Y. Kwok, J. Q. Yip, Z. Qiu, C. H. Chi, and K. Y. Lam, "Bona fide Cross Testing Reveals Weak Spot in Audio Deepfake Detection Systems," in *Interspeech 2025*, 2025, pp. 2230–2234.

[7] S. Ding, Y. Zhang, and Z. Duan, "SAMO: Speaker attractor multi-center one-class learning for voice anti-spoofing," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[8] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms," *Proc. Interspeech*, pp. 3583–3587, 2020.

[9] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[10] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised Feature Learning via Non-Parametric Instance Discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3733–3742.

[11] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 9912–9924.

[12] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," 2020. [Online]. Available: https://arxiv.org/abs/1911.01601

[13] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang *et al.*, "ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *arXiv preprint arXiv:2109.00535*, 2021.

[14] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[15] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," *arXiv preprint arXiv:2011.01108*, 2021.

[16] E. Rosello, A. Gomez-Alanis, A. Gomez, and A. Peinado, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," in *Proc. Interspeech 2023*, 2023, pp. 5281–5285.

Algorithm 1 Memory Bank Initialization

```
Input: Number of slots $K$, Embedding dimension $D$
Output: Bonafide bank $\mathbf{M}_{real}$, Spoof bank $\mathbf{M}_{spoof}$
$\mathbf{M}_{real} \sim \mathcal{N}(0,1)^{K \times D}$
$\mathbf{M}_{spoof} \sim \mathcal{N}(0,1)^{K \times D}$
$\mathbf{M}_{real} \leftarrow$ RowL2Normalize$(\mathbf{M}_{real})$
$\mathbf{M}_{spoof} \leftarrow$ RowL2Normalize$(\mathbf{M}_{spoof})$
return $\mathbf{M}_{real}, \mathbf{M}_{spoof}$
```

Algorithm 2 Top-K Sparse Reconstruction

```
Input: Embedding $\mathbf{z} \in \mathbb{R}^{B \times D}$, Memory Bank $\mathbf{M} \in \mathbb{R}^{K \times D}$, Top-$k$ parameter $k$
Output: Reconstructed $\hat{\mathbf{z}}$, Error $E$, Similarity $\mathbf{S}$
$\hat{\mathbf{M}} \leftarrow$ RowL2Normalize$(\mathbf{M})$
$\mathbf{S} \leftarrow \mathbf{z} \cdot \hat{\mathbf{M}}^\top$                              ▷ Cosine Similarity
$\mathbf{V}_{top}, \mathbf{I}_{top} \leftarrow$ TopK$(\mathbf{S}, k)$                       ▷ Select top-k slots
$\mathbf{W} \leftarrow$ Softmax$(\mathbf{V}_{top})$                                         ▷ Compute weights
$\mathbf{M}_{sel} \leftarrow$ Gather$(\hat{\mathbf{M}}, \mathbf{I}_{top})$
$\hat{\mathbf{z}} \leftarrow \sum_{j=1}^{k} \mathbf{W}_{:,j} \cdot \mathbf{M}_{sel,:,j}$    ▷ Weighted Sum
$E \leftarrow \|\mathbf{z} - \hat{\mathbf{z}}\|_2^2$                                        ▷ MSE Calculation
return $\hat{\mathbf{z}}, E, \mathbf{S}$
```

Algorithm 3 Sinkhorn-Knopp Algorithm (OT Regularization)

```
Input: Logits $\mathbf{L} \in \mathbb{R}^{B \times K}$, Smooth $\epsilon$, Iterations $T$
Output: Optimal Assignment Matrix $\mathbf{Q}$
$\mathbf{Q} \leftarrow \exp(\mathbf{L}/\epsilon)$
for $t = 1$ to $T$ do
    $\mathbf{Q} \leftarrow \mathbf{Q} \oslash (\mathbf{Q} \cdot \mathbf{1}_K \cdot \mathbf{1}_K^\top)$    ▷ Row Norm
    $\mathbf{Q} \leftarrow \mathbf{Q} \oslash (\mathbf{1}_B \cdot \mathbf{1}_B^\top \cdot \mathbf{Q})$    ▷ Col Norm
end for
$\mathbf{Q} \leftarrow \mathbf{Q} \oslash (\mathbf{Q} \cdot \mathbf{1}_K \cdot \mathbf{1}_K^\top)$        ▷ Final Row Norm
return $\mathbf{Q}$
```

Algorithm 4 Dual Reconstruction Loss

```
Input: Errors $E_{real}, E_{spoof}$, Labels $y$, Margin $m$
$\mathcal{B} \leftarrow \{i \mid y_i = 0\}$                  ▷ Bonafide indices
$\mathcal{S} \leftarrow \{i \mid y_i = 1\}$                  ▷ Spoof indices
$\mathcal{L} \leftarrow 0$
if $|\mathcal{B}| > 0$ then
    $\mathcal{L} \leftarrow \mathcal{L} +$ Mean$(E_{real}[\mathcal{B}])$
    $\mathcal{L} \leftarrow \mathcal{L} +$ Mean(ReLU$(m - E_{spoof}[\mathcal{B}])$)
end if
if $|\mathcal{S}| > 0$ then
    $\mathcal{L} \leftarrow \mathcal{L} +$ Mean$(E_{spoof}[\mathcal{S}])$
    $\mathcal{L} \leftarrow \mathcal{L} +$ Mean(ReLU$(m - E_{real}[\mathcal{S}])$)
end if
return $\mathcal{L}$
```

Algorithm 5 OT Loss Computation

```
Input: Logits $\mathbf{L}$, Target Assignment $\mathbf{Q}$ (from Sinkhorn)
Output: Loss scalar $\mathcal{L}_{ot}$
$\mathbf{P} \leftarrow$ LogSoftmax$(\mathbf{L})$
$\mathbf{Q}_{target} \leftarrow$ Detach$(\mathbf{Q})$                ▷ Stop gradient for target
$\mathcal{L}_{ot} \leftarrow -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{K} \mathbf{Q}_{target,i,j} \cdot \mathbf{P}_{i,j}$
return $\mathcal{L}_{ot}$
```

Algorithm 6 Diversity Loss (Entropy Maximization)

```
Input: Attention Weights $\mathbf{W} \in \mathbb{R}^{B \times K}$
Output: Loss scalar $\mathcal{L}_{div}$
$\bar{\mathbf{w}} \leftarrow \frac{1}{B} \sum_{i=1}^{B} \mathbf{W}_{i,:}$
$H \leftarrow -\sum_{j=1}^{K} \bar{\mathbf{w}}_j \cdot \log(\bar{\mathbf{w}}_j + \epsilon)$
$\mathcal{L}_{div} \leftarrow -H$                             ▷ Maximize entropy
return $\mathcal{L}_{div}$
```

Algorithm 7 Multi-Center OC-Softmax Loss (Exp 1 & 2)

```
Input: Embeddings $\mathbf{z}$, Centers $\mathbf{C}$, Labels $y$, Margins $m_{real}, m_{fake}$, Scale $\alpha$
$\mathbf{S} \leftarrow$ L2Normalize($\mathbf{z}$) $\cdot$ L2Normalize($\mathbf{C}$)$^\top$
$s_{max} \leftarrow \max_j(\mathbf{S}_{:,j})$                                  ▷ Max similarity
$\mathcal{L} \leftarrow 0$
for each sample $i$ in batch do
    if $y_i = 0$ then                                                          ▷ Bonafide
        $\mathcal{L} \leftarrow \mathcal{L} +$ Softplus$(\alpha(m_{real} - s_{max,i}))$
    else                                                                       ▷ Spoof
        $\mathcal{L} \leftarrow \mathcal{L} +$ Softplus$(\alpha(s_{max,i} - m_{fake}))$
    end if
end for
return Mean($\mathcal{L}$)
```

Algorithm 8 Contrastive Memory Loss (Exp 3)

```
Input: Embedding $\mathbf{z}$, Memory $\mathbf{M}$, Labels $y$, Temp $\tau$, Margin $m$
$\mathbf{S} \leftarrow (\mathbf{z} \cdot \mathbf{M}^\top)/\tau$
$\mathcal{L}_{pull} \leftarrow 0, \mathcal{L}_{push} \leftarrow 0$
if Bonafide samples exist then
    $\mathcal{L}_{pull} \leftarrow -$Mean(LogSumExp($\mathbf{S}$[Bonafide]))
end if
if Spoof samples exist then
    $\mathcal{L}_{push} \leftarrow$ Mean(ReLU(max($\mathbf{S}$[Spoof]) $+ m$))
end if
return $\mathcal{L}_{pull} + \mathcal{L}_{push}$
```

Algorithm 9 Adaptive Margin Scheduler (Exp 5)

```
Input: Current Step $t$, Warmup $T_{warm}$, Total Steps $T_{total}$
Output: Current margins $m_{real}, m_{fake}$
Hyperparams: $m_{real}^{start} = 0.7, m_{real}^{end} = 0.95$
Hyperparams: $m_{fake}^{start} = 0.3, m_{fake}^{end} = 0.1$
if $t < T_{warm}$ then
    $p \leftarrow 0$
else
    $p \leftarrow \frac{t - T_{warm}}{T_{total} - T_{warm}}$
    $p \leftarrow \min(p, 1.0)$
end if
$m_{real} \leftarrow m_{real}^{start} + p \cdot (m_{real}^{end} - m_{real}^{start})$
$m_{fake} \leftarrow m_{fake}^{start} - p \cdot (m_{fake}^{start} - m_{fake}^{end})$
return $m_{real}, m_{fake}$
```

Table 5. Main Training Configuration (Baseline)

Category	**Parameter**	**Value**
Dataset	Dataset Train split Dev split Eval split	ASVspoof 2019 LA Official train (A01–A06 attacks) Official dev (A01–A06 attacks) Official eval (A07–A19 attacks, unseen)
Audio Processing	Sample rate Max length Normalization	16 kHz 64,600 samples (≈ 4 seconds) Per-utterance mean-variance
Model Architecture	Backbone Embedding dimension Freeze encoder Memory slots (per bank) Top-K attention	TitaNet-Small (10M parameters) 192 (L2-normalized) **False** 64 (Bonafide and Spoof banks) 10
Training	Optimizer Weight decay Initial learning rate Learning rate schedule Warm-up steps Max training steps Batch size Gradient clipping	AdamW 2×10^{-3} 1×10^{-4} Warm-up + Cosine annealing 500 5,000 64 5.0
Loss Weights	λ_{recon} λ_{ot} λ_{oc} λ_{div} $\lambda_{\text{contrastive}}$	1.0 0.2 0.0 0.1 0.0
Sinkhorn OT	Iterations Epsilon (ϵ)	3 0.05
Data Augmentation	RawBoost	Enabled (Algorithms: ISD, SSI)
Hardware	Accelerator Devices Precision	GPU 2 RTX4090 FP32

Table 6. Enhanced Mode Configuration

Category	**Parameter**	**Value**
Mode	Configuration	Enhanced
OC-Softmax	Centers m_{real} m_{fake} α	20 0.9 0.3 20.0
Loss Weights (Enhanced)	λ_{recon} λ_{ot} λ_{oc} λ_{div} $\lambda_{\text{contrastive}}$	1.0 0.2 0.5 0.1 0.3
Score Fusion	Fusion strategy Fusion weight (w) Normalization	Reconstruction + OC-Softmax (weighted combination) 0.5 (OC score weight) Z-score (per-score normalization)
Adaptive Margin	Use adaptive margin Warmup steps Total steps	False/True 500 5,000

Table 7. Experiment Configurations Comparison (Baseline to Exp 9)

Configuration	**Baseline**	**Exp 1**	**Exp 2**	**Exp 3**	**Exp 4**	**Exp 5**	**Exp 6**	**Exp 7**	**Exp 8**	**Exp 9**
Model										
TitaNet	Small	Small	Small	Small	**Large**	**Large**	**Large**	Small	Small	Small
Memory slots	64	64	64	64	128	128	128	128	**0**	64
Top-K	10	10	10	10	10	10	10	**20**	**0**	10
Loss Functions										
Reconstruction	✓	✓	✓	✓	✓	✓	✓	✓	×	✓
OT regularization	✓	✓	✓	✓	✓	✓	✓	✓	×	×
OC-Softmax	×	✓	✓	✓	✓	✓	✓	×	✓	×
Contrastive	×	×	×	✓	✓	✓	✓	×	×	×
Loss Weights										
λ_{recon}	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	**0.0**	1.0
λ_{ot}	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	**0.0**	**0.0**
λ_{oc}	0.0	0.5	0.5	0.5	0.5	0.5	0.5	0.0	0.5	0.0
$\lambda_{\text{contrastive}}$	0.0	0.0	0.0	0.3	0.3	0.3	0.3	0.0	0.0	0.0
OC-Softmax										
Centers	–	1	20	20	20	20	20	–	1	–
m_{real}	–	0.9	0.9	0.9	0.9	0.9	0.9	–	0.9	–
m_{fake}	–	0.3	0.3	0.3	0.3	0.3	0.3	–	0.3	–
Adaptive margin	–	×	×	×	×	✓	✓	–	×	–
Mode & Scoring										
Mode	Basic	Basic	Enhanced	Enhanced	Enhanced	Enhanced	Enhanced	Basic	Basic	Basic
Score fusion	Recon	Recon	Recon	Recon	Recon	Recon	**Combined**	Recon	**OC**	Recon
Score weight	–	–	–	–	–	–	**0.7**	–	–	–