MEMORY BANK-BASED UNSUPERVISED ANOMALOUS SOUND DETECTION EXPLOITING STEREO SPATIAL INFORMATION

Technical Report

Ryo Morita, Ryoya Kozai, Shota Sekino, Yusuke Kishi

KONICA MINOLTA, INC., Tokyo, Japan
{ryo.morita2, ryoya.kozai, shota.sekino, yusuke.kishi1}@konicaminolta.com

ABSTRACT

This report describes our four systems submitted to DCASE 2026 Challenge Task 2: Noise-Aware Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. All systems share a unified hypothesis: stereo spatial information is the key to domain-robust anomaly detection. Within a memory bank-based k-nearest neighbor (kNN) framework, we present two contrasting strategies for exploiting the two-channel recording. Systems 1–2 employ a coherence-weighted power ratio mask for signal-level noise suppression combined with a fully fine-tuned EAT-large transformer trained with ArcFace loss. Systems 3–4 adopt a training-free approach using domain invariant features (DIF) extracted from stereo signal processing combined with frozen EAT-large embeddings. Our best system (System 1) achieves $\Omega = 64.41\%$ on the development set, while the training-free System 3 achieves $\Omega = 62.34\%$ with zero trainable parameters. The two designs—learned adaptation and physics-based invariance—provide complementary strategies that may generalize differently to unseen evaluation machines.

***Index Terms*—** Anomalous sound detection, stereo signal processing, memory bank, kNN, coherence mask, domain invariant features, EAT

1. INTRODUCTION

DCASE 2026 Challenge Task 2 [1, 2, 3, 4] addresses noise-aware unsupervised anomalous sound detection (UASD) for machine condition monitoring. The task imposes five requirements: (1) unsupervised learning using only normal sounds, as anomalies are rare and diverse in real factories; (2) domain generalization against environmental noise and operational condition changes; (3) generalization to machine types completely unseen during development; (4) robustness both with and without attribute information; and (5) training and inference with **two-channel audio** recorded at different distances from the target machine. This stereo recording setup—a near-field microphone (ch0) capturing high-SNR machine sounds and a far-field microphone (ch1) capturing attenuated machine sounds mixed with environmental noise—is the key new element distinguishing this task from prior DCASE challenges. It enables systems to leverage spatial cues for noise suppression and domain-invariant feature extraction.

Recent top-performing ASD systems [5, 6] have demonstrated the effectiveness of large-scale pre-trained audio backbones for anomalous sound detection. Building upon these advances, we employ EAT-large embeddings within a memory bank-based framework and further exploit stereo spatial information at multiple levels. Within this unified framework (Figure 1), we present two contrasting strategies:

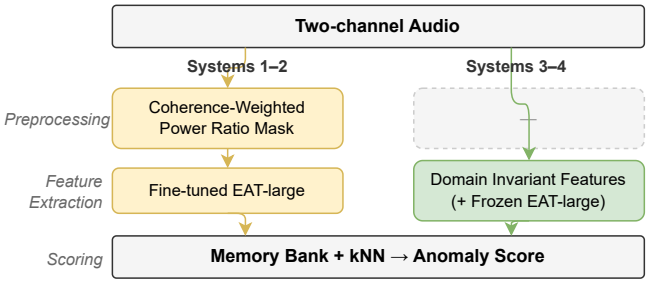

Figure 1: System overview. Systems 1–2 (left) and Systems 3–4 (right) follow independent architectures; both employ memory bank-based kNN scoring.

- **Fine-tuning approach (Systems 1–2):** Signal-level noise suppression via a coherence-weighted power ratio mask, followed by fully fine-tuned EAT-large [7] with ArcFace loss [8] and cosine kNN scoring.
- **Training-free approach (Systems 3–4):** Physics-based domain invariant features (DIF) combined with frozen EAT-large embeddings, ZCA whitening, and weighted kNN with percentile ensemble—requiring zero trainable parameters.

Our contributions are: (1) a memory bank-based anomalous sound detection framework that systematically exploits stereo spatial information at multiple levels; and (2) two complementary stereo exploitation strategies—signal-level noise suppression via a coherence-weighted power ratio mask and feature-level domain-invariant stereo descriptors—demonstrating that the two-channel recording setup is an effective foundation for this task regardless of the exploitation approach.

2. PROPOSED METHODS

Figure 1 shows the overall system architecture. All four systems share exploitation of stereo spatial information and memory bank-based kNN anomaly scoring, but follow fundamentally different architectures as summarized in Table 1. Figure 2 details the Systems 1–2 pipeline, and Figure 3 details the Systems 3–4 dual-path architecture.

Table 1: System configuration summary.

	S1	S2	S3	S4
Preprocess	Coherence mask		None	
Feature	EAT-large (FT)		DIF + EAT	DIF only
Training	ArcFace		None (0 params)	
Augmentation	Temporal crop		Residual aug + Pseudo target	
Scoring	Cosine k=1		Weighted k=3	Euclidean k=1
Ensemble	Single	4-seed avg.	Percentile	Single

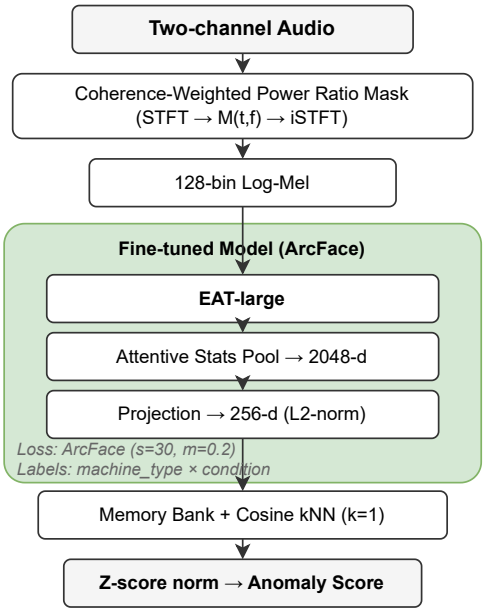

Figure 2: Systems 1–2 pipeline, expanding the left branch of Figure 1. Shaded region: training-phase components (ArcFace loss). Configuration details are listed in Table 1.

2.1. Stereo Spatial Preprocessing

2.1.1. *Coherence-Weighted Power Ratio Mask (Systems 1–2)*

We propose an STFT-domain noise suppression mask that exploits the near-field/far-field energy asymmetry and inter-channel coherence of the two-microphone setup. Let $X_0(t,f)$ and $X_1(t,f)$ denote the STFT of ch0 and ch1 (n_fft=400, hop=160). Given temporally-smoothed power spectral densities $P_i(t,f)$ (exponential moving average, α=0.9), the cross-power spectral density $\Phi_{01}(t,f) = \mathrm{E}[X_0(t,f)\,X_1^*(t,f)]$, and the magnitude squared coherence $\gamma(t,f) = |\Phi_{01}|^2/(P_0P_1+\varepsilon)$, the composite mask is:

$$M(t,f) = \max\Big(\sigma\big(\beta(\log P_0(t,f) - \log P_1(t,f))\big) \cdot \big(1 - w\gamma(t,f)\big),\ \delta\Big). \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function, β=5 controls the transition sharpness, w=0.7 is the coherence weight, and δ=0.2 is the floor value (corresponding to a maximum attenuation of approximately 14 dB). The enhanced signal is $Y(t,f) = M(t,f) \cdot X_0(t,f)$, followed by inverse STFT.

The mask is SNR-adaptive by construction: machines where ch0 strongly dominates (e.g., ToyCar, +21 dB mean power ratio) pass through with minimal modification (mean mask ≈ 0.96), while machines with comparable channel levels (e.g., fan, -0.6 dB) undergo substantial noise suppression (mean mask ≈ 0.50, 36% of T-F bins at the floor). This automatic calibration requires no per-machine tuning—it follows directly from the sigmoid's response to the log power ratio.

The coherence term $(1 - w\gamma)$ provides a small but consistent performance gain (+1.3% Ω; see Section 3.3). The observed magnitude-squared coherence is generally low (mean $\gamma < 0.06$), indicating weak similarity between the two microphone signals under the near-/far-field recording setup. As a result, this term has limited global effect and mainly influences time-frequency bins where the power-ratio gate is uncertain (i.e., near the decision boundary). In such regions, it provides a lightweight spatial cue that slightly refines the mask rather than performing explicit source separation.

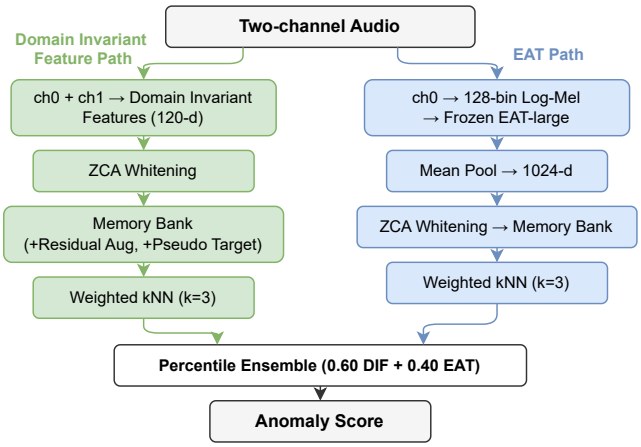

Figure 3: Systems 3–4 dual-path pipeline, expanding the right branch of Figure 1. The DIF path (left) extracts 120-d hand-crafted features from both channels; the EAT path (right) uses frozen EAT-large on ch0 only. Scores are fused via percentile ensemble. Configuration details are listed in Table 1.

2.1.2. *Domain Invariant Features (Systems 3–4)*

For Systems 3–4, we design hand-crafted features that are operationally robust to domain shift. Starting from 151 candidate features computed on development machines, we exclude 31 dimensions with large source/target distribution difference (Wasserstein distance $> \tau$), yielding 120-dim DIF features. We use the term "domain invariant" operationally: these features have small source/target shift by construction on development machines, though invariance to unseen evaluation domains is not guaranteed.

Ch0 Features (80-dim): Delta MFCC Std (26-d), MFCC Mean (13-d), Spectral Contrast (14-d), Band Energy ratios (11-d), Spectral Shape (8-d), and Impulsiveness factors (8-d). All are ratio-based or relative measures inherently invariant to recording gain and distance.

Stereo Features (40-dim): Basic stereo (7-d: LR correlation, RMS ratio, M/S ratio, centroid diff, HF ratio diff), Per-band LR energy difference (9-d), MFCC L-R difference (13-d), and Coherence Profile as band-wise MSC (11-d). These capture spatial relationships between the two microphones, providing direct sensitivity to anomalies that alter the acoustic radiation pattern.

2.2. Feature Extraction: EAT-large

Both approaches use EAT-large [7], but employ it differently.

Systems 1–2 (Fine-tuned): All transformer parameters are fully fine-tuned. Attentive statistics pooling [9] aggregates the patch sequence into a 2048-d representation (weighted mean + std), and a projection head ($2048 \to 512 \to$ BN $\to$ ReLU $\to 256 \to$ L2-norm) maps to a compact embedding space. Since anomaly labels are unavailable, we repurpose operating-condition metadata as proxy classes for ArcFace loss [8]. For machines with attribute annotations (e.g., fan speed, gearbox production lot), each unique attribute combination defines a class; for machines without attributes, source and target domains serve as the two classes—yielding 43 classes total across 7 development machines. This encourages the feature space to separate operationally distinct normal states, structuring the normal manifold for downstream kNN scoring. For machines where source/target attributes are fully disjoint (e.g., ToyCarEmu, gearboxEmu), the resulting clusters may partially reflect domain identity; however, the structured embedding still benefits kNN anomaly detection empirically (Section 3.3). Temporal crop augmentation randomly extracts 60–100% of the input duration during training, providing implicit speed perturbation without modifying spectral content.

Systems 3–4 (Frozen): EAT-large is used without fine-tuning. The ch0 signal (high-SNR) is input as 128-bin log-mel, with mean pooling over patch tokens yielding 1024-d embeddings. Freezing reduces the risk of parameter overfitting to development machines.

2.3. Memory Bank Construction

Systems 1–2: L2-normalized 256-d embeddings from normal training clips are stored directly. Test-time augmentation (10 temporal crops with feature-level averaging) stabilizes the embedding estimates for both bank construction and inference.

Systems 3–4: ZCA whitening is applied to both DIF and EAT feature spaces, making Euclidean distance equivalent to Mahalanobis distance while accommodating multimodal normal distributions through kNN (unlike single-Gaussian Mahalanobis).

Adaptive Residual Augmentation (DIF only): The 10 target clips provide insufficient coverage of the 40-dim stereo subspace, causing false positives. We augment by varying ch1 while keeping ch0 fixed:

$$R_{\text{aug}} = (1-s) \cdot y_0 + s \cdot y_1, \quad s \sim \mathcal{U}(0.5, 1.5) \tag{2}$$

This affects only the 40 stereo dimensions, safely densifying the stereo subspace without generating physically implausible machine states. The augmentation ratio adapts in three tiers based on LR correlation. Here, the augmentation ratio is defined relative to the number of source samples, i.e., the total number of generated augmented target samples normalized by the source dataset size (e.g., an augmentation ratio of 1.0 corresponds to generating augmented target samples equal in number to the entire source dataset). Specifically, aug_ratio $= 0.0$ when LR_corr > 0.6 (ch0$\approx$ch1, residual negligible), 0.1 when $0.3 <$ LR_corr ≤ 0.6, and 1.0 when LR_corr ≤ 0.3 (large spatial difference).

Pseudo Target (DIF only): Source bank diversity is transferred to the target domain via parallel translation: pseudo_target $=$ $\text{bank}_{\text{source}} + (\bar{\mu}_{\text{target}} - \bar{\mu}_{\text{source}})$. This translation assumes approximately linear domain shift in the DIF space, which holds by construction (Wasserstein-based exclusion ensures small shifts). However, if domain shift on unseen machines exceeds development-set levels, the pseudo-target may map anomalous patterns into the augmented normal manifold, increasing false negatives.

2.4. Anomaly Scoring

System 1 (Primary): Cosine-distance kNN (k=1) in the 256-d L2-normalized space:

$$\text{score}(x) = 1 - \max_{z \in Z_{\text{train}}} \cos\big(f(x), f(z)\big) \tag{3}$$

System 2: 4-seed ensemble of k=1 scorers. Each model's raw scores are Z-score normalized using leave-one-out training set self-scores before averaging, calibrating across seeds with different score scales.

System 3 (Percentile Ensemble): Weighted kNN (k=3) with negative weights for local density correction:

$$\text{score} = 1.0 \cdot d_1 - 0.35 \cdot d_2 - 0.05 \cdot d_3 \tag{4}$$

The negative weights on d_2, d_3 provide local density correction, suppressing false positives in sparse bank regions; these coefficients are robust to perturbation. DIF and EAT scores are each converted to their percentile rank within the per-machine training LOO score distribution, ensuring commensurable scales before weighted fusion: $0.60 \times \text{pct}_{\text{DIF}} + 0.40 \times \text{pct}_{\text{EAT}}$. DIF receives a higher weight because it showed stronger domain-shift robustness on the development set.

System 4 (DIF kNN k=1): Nearest-neighbor L2 distance in the DIF space only. We prioritize the parameter-free formulation for maximum stability on unseen machines.

3. EXPERIMENTAL RESULTS

3.1. Setup

We evaluate on the DCASE 2026 Task 2 development dataset [1, 2, 3, 4]: 7 machine types, each with 990 source and 10 target normal training clips (16 kHz stereo). Test sets contain 50 clips per machine/domain. The official metric Ω is the harmonic mean of AUC(source), AUC(target), and pAUC across all machines.

3.2. Overall Results

Table 2 shows the per-machine results of all four systems. System 1 achieves the best overall Ω=64.41%. The fine-tuning approach dominates on machines with large domain shift and time-frequency localized sounds—gearboxEmu (S2 target AUC 74.64% vs. S4 54.08%) and valveEmu (S2 pAUC 66.00% vs. S4 53.37%)—where ArcFace's condition-aware clustering absorbs the source–target gap. Conversely, the training-free approach excels when anomalies manifest as impulsiveness or spatial pattern changes: bearingEmu (+6.9% pAUC over S1, captured by DIF impulsiveness factors) and fan (+13.4% AUC(source) for S4, where the mask suppresses 36% of T-F bins containing broadband machine signal while DIF stereo features detect radiation pattern changes directly).

3.3. Ablation Study

Table 3 shows the cumulative component contribution for System 1.

Coherence term contribution: Comparing w=0.7 vs. w=0.0 (power ratio only) under full fine-tuning yields a mean improvement of +1.3% Ω (all runs positive). Despite low observed MSC values

Table 2: Per-machine results across systems (dev set, %).

Machine	Metric	S1	S2	S3	S4
ToyCar	AUC(s)	87.12	88.08	81.30	78.28
	AUC(t)	53.56	52.80	49.69	46.46
	pAUC	62.84	61.84	56.95	56.63
ToyCarEmu	AUC(s)	61.42	55.42	70.68	70.10
	AUC(t)	88.56	91.04	70.52	67.58
	pAUC	55.05	52.11	56.50	58.53
bearingEmu	AUC(s)	57.74	55.80	60.48	61.40
	AUC(t)	55.82	51.10	61.34	63.76
	pAUC	53.79	52.74	59.89	60.68
fan	AUC(s)	75.60	80.76	85.29	89.04
	AUC(t)	60.70	59.42	62.29	67.02
	pAUC	58.68	57.32	51.89	55.00
gearboxEmu	AUC(s)	82.68	81.62	74.26	76.14
	AUC(t)	71.46	74.64	57.57	54.08
	pAUC	56.58	65.26	55.79	56.21
sliderEmu	AUC(s)	76.80	75.70	70.47	69.56
	AUC(t)	57.12	53.66	54.89	54.26
	pAUC	54.16	52.89	53.45	52.58
valveEmu	AUC(s)	75.48	73.86	75.11	71.88
	AUC(t)	85.73	88.24	78.44	76.14
	pAUC	62.53	66.00	54.50	53.37
Ω(%)		**64.41**	63.72	62.34	62.15

Table 3: Cumulative ablation of System 1 components.

Configuration	Ω(%)	Δ
Frozen EAT-large (baseline)	56.27	—
+ Full fine-tuning	62.07	+5.80
+ Coherence mask (w=0.7)	**64.41**	+2.34
Total improvement		+8.14

(mean γ=0.004–0.056), the coherence term consistently improves performance.

Systems 3–4 ablation: DIF alone achieves Ω=62.18%, EAT alone 60.96%, and the percentile ensemble yields 62.34% (+0.16% over DIF alone). The complementarity arises because DIF excels for machines with rich stereo information (e.g., fan, LR_corr= 0.86) while EAT captures broad spectral patterns.

Design choice validation (System 1–2): Among angular margin losses (vs. ArcFace m=0.2 baseline), CosFace [10] (-0.8% Ω), ElasticFace [11] (-2.5%), and MagFace [12] (-2.6%) all underperform. Full fine-tuning outperforms LoRA [13] (rank=16) by $+3.3\%$ Ω (seed-matched), reflecting the large domain gap between AudioSet and industrial machine sounds.

3.4. Discussion

Why signal-level preprocessing helps: The mask improves the effective SNR of the input to EAT-large by suppressing T-F bins dominated by environmental noise. This is SNR-adaptive: machines with low near-field advantage (fan, gearboxEmu) receive substantial denoising, while high-SNR machines (ToyCar) pass through nearly unmodified. By cleaning the input representation, the mask allows ArcFace fine-tuning to learn machine-specific features rather than noise patterns. We also explored domain adaptation at the embedding level (mean shift, L2 normalization, Z-score normalization) and the model level (DANN [14], MixStyle [15]), but none provided meaningful improvement—DANN yielded a maximum gain of only +0.2% Ω, which is negligible compared to the improvements obtained from signal-level preprocessing and full fine-tuning. Full fine-tuning of all transformer blocks outperforms parameter-efficient alternatives such as LoRA [13] ($+3.3\%$ Ω, seed-matched), suggesting that the domain gap between AudioSet pre-training and industrial machine sounds requires full-rank adaptation beyond low-rank subspaces.

Stereo spatial information is effective at both levels: For Systems 1–2, the coherence mask provides $+2.3\%$ Ω at the signal level. For Systems 3–4, 40 of 120 DIF dimensions explicitly encode stereo spatial relationships, contributing most strongly for machines with pronounced inter-channel differences. Both approaches suggest that the two-channel recording setup enables meaningful performance improvements regardless of the exploitation strategy. However, each approach exhibits characteristic limitations. The masked fine-tuned pipeline underperforms the DIF-only system on fan (by 13.4% in AUC(source)), whereas the DIF-based approach underperforms the ArcFace-based system on gearboxEmu (by 20.6% in AUC(target)). These complementary behaviors motivate the use of a dual-paradigm design to improve robustness to unseen machine characteristics.

4. CONCLUSION

We presented four systems for DCASE 2026 Task 2, unified by the hypothesis that stereo spatial information is the key enabler for domain-robust anomalous sound detection. The fine-tuning approach (Systems 1–2) combines a coherence-weighted power ratio mask with fully fine-tuned EAT-large, achieving Ω=64.41%. The training-free approach (Systems 3–4) combines domain invariant features with frozen EAT-large embeddings, achieving Ω=62.34% with zero trainable parameters. The two contrasting designs—learned adaptation and physics-based invariance—provide complementary strategies that may generalize differently to unseen evaluation machines.

5. REFERENCES

[1] T. Nishida, N. Harada, D. Takeuchi, D. Niizumi, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2026 challenge task 2: Noise-aware unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2606.01578*, 2026.

[2] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1–5.

[3] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes*

and Events 2022 Workshop (DCASE2022), Nancy, France, November 2022.

[4] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," *Proceedings of 31st European Signal Processing Conference (EUSIPCO)*, pp. 191–195, 2023.

[5] T. Fujimura, I. Kuroyanagi, and T. Toda, "The NU Systems for DCASE 2025 Challenge Task 2," DCASE 2025 Challenge, Tech. Rep., 2025.

[6] X. Zheng, A. Jiang, B. Han, S. Zhang, W.-Q. Zhang, X. Chen, C. Lu, P. Fan, J. Liu, and Y. Qian, "SJTU-AITHU System for DCASE 2025 Anomalous Sound Detection Challenge," DCASE 2025 Challenge, Tech. Rep., 2025.

[7] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, K. Larson, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2024, pp. 3807–3815, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2024/421

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF CVPR*, 2019, pp. 4690–4699.

[9] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.

[10] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF CVPR*, 2018, pp. 5265–5274.

[11] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "ElasticFace: Elastic margin loss for deep face recognition," in *Proc. IEEE/CVF CVPRW*, 2022, pp. 1578–1587.

[12] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A universal representation for face recognition and quality assessment," in *Proc. IEEE/CVF CVPR*, 2021, pp. 14 225–14 234.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.

[14] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.

[15] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with MixStyle," in *Proc. ICLR*, 2021.