Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 3, No. 1, 2026

Blind Device-Response Interpolation for Intermediate-Domain Unsupervised Adaptation in Cross-Device Acoustic Scene Classification

Claire Whitmore, Lucas Bennett *

Department of Computer Science and Engineering, School of Engineering and Applied Sciences, University at Buffalo, NY, USA

Abstract: Recording-device mismatch is one of the dominant failure modes of acoustic scene classification: a model trained on one microphone degrades sharply on an unseen one. Aligning source and target in a single adversarial step is unstable when the device gap is large, and recent intermediate-domain formulations mitigate this by inserting bridge domains, but they build those bridges from hand-specified parametric device operators that are unavailable at deployment time. We propose BRIDA, an intermediate-domain unsupervised adaptation framework whose bridges are derived from the device response itself. A blind estimator recovers the relative log-mel magnitude response between the labelled source devices and the unlabelled target device from long-term spectral statistics alone, with an optional class-balanced refinement for the case where the two unlabelled pools differ in scene composition. Because a magnitude filter is additive in the log-mel domain, fractional powers of the estimate yield a continuum of label-preserving bridge domains at essentially zero cost; we traverse that continuum with a curriculum and regularise it with an ordinal path-position adversary, a residual maximum-mean-discrepancy anchor and confidence-gated target prototypes. Experiments use 1648 real recordings and 66 measured microphone impulse responses under a leave-one-device-out protocol with disjoint source, target and evaluation content. Averaged over 3 unseen devices and 3 seeds, BRIDA attains 53.00 ± 4.29% target accuracy and 52.99 ± 4.58% macro-F1, improving on source-only training by 7.83 points and on a discrete two-bridge intermediate-domain baseline by 0.61 points, with no inference-time overhead.

Keywords: Acoustic scene classification, unsupervised domain adaptation, recording-device mismatch, intermediate domains, blind channel estimation, adversarial learning.

1. Introduction

Acoustic scene classification (ASC) assigns an audio recording to a label describing the environment in which it was captured, and it underpins context-aware mobile computing, acoustic monitoring and machine listening [1], [2]. A recurring obstacle is that a classifier which is accurate on recordings from one device fails on recordings of the same scene captured by another. The cause is physical rather than semantic: microphones and their enclosures impose different band limits, resonances, spectral notches and effective noise floors, and these transformations reshape exactly the time-frequency evidence a classifier relies upon. Multi-device corpora and the DCASE challenge series have documented consistent gaps between dominant and low-resource devices, under increasingly strict complexity budgets [3-6].

Two families of remedies dominate. Domain generalisation broadens the training distribution so that no adaptation is needed: device impulse response augmentation convolves source audio with a library of measured microphone responses [13], and frequency-wise statistics perturbation such as Freq-MixStyle or relaxed instance frequency-wise normalisation randomises the spectral envelope [14-16]. Unsupervised domain adaptation (UDA) instead exploits unlabelled recordings from the specific device that will be deployed, by minimising a kernel or moment discrepancy [24-26] or by learning device-invariant features through a gradient-reversal discriminator [27-29]. Both families have been applied to ASC with clear benefit [17-21].

A single source-to-target alignment step is nevertheless fragile when the device gap is large. Early target predictions are unreliable, and an over-powered discriminator can equalise marginal distributions while mixing semantically distinct modes, a regime in which adaptation is worse than no adaptation at all. Gradual domain adaptation addresses this by decomposing one large shift into a sequence of smaller ones, and theory shows that error propagation along a sufficiently smooth path is controllable [34, 35]. Within ASC, the class-conditional intermediate-domain framework of Guo et al. [39] showed convincingly that inserting label-preserving bridge devices between source and target, and conditioning the domain discriminator on class predictions, recovers a large part of the accuracy lost to device shift while adding no inference-time parameters. That work also identified the main obstacle to deployment: its bridges are produced by explicitly specified device operators, so transferring the idea to a real target device requires an estimate of the intermediate responses which is not available a priori.

This paper takes that obstacle as its starting point. Our observation is that the quantity needed to build bridges is not the full device transfer function but only the relative magnitude response between the source devices and the target device, and that this relative response is identifiable from unlabelled audio whenever the two device populations record comparable content. We therefore estimate it blindly from long-term spectral statistics and generate a continuum of intermediate domains by applying fractional powers of the estimate. Because a magnitude-only filter becomes an additive offset in the log-mel domain, the entire bridge family is obtained without touching a waveform. The contributions are as follows.

1) A blind relative device-response estimator that uses unlabelled target audio only, is robust to transient scene events through nested band-wise medians, and admits an

optional class-balanced refinement for the case where the two unlabelled pools differ in scene composition.

2) A zero-cost construction of label-preserving intermediate domains as additive offsets in the log-mel domain, validated numerically against explicit waveform-domain filtering.

3) A continuous device-path curriculum together with an ordinal path-position adversary that regresses a soft position along the source-to-target path from the multilinear feature-prediction map, replacing discrete bridge indices and categorical domain discriminators.

4) A reproducible cross-device benchmark built from real recordings and measured microphone impulse responses with mutually disjoint source, target and evaluation content, and a leave-one-device-out comparison against eight domain-generalisation and adaptation baselines.

2. Related Work

2.1. Device Mismatch in Acoustic Scene Classification

Device robustness became a first-class objective once multi-device corpora made the gap measurable [3], and subsequent challenge analyses coupled it with strict complexity budgets [4-6]. Model-side responses include receptive-field control and knowledge distillation from large audio transformers [15], often on top of representations pretrained on large audio collections [10, 11]. Data-side responses simulate device diversity directly: convolving source audio with measured microphone impulse responses is now a standard and remarkably effective augmentation [13], and spectral masking or frequency-statistics mixing act similarly [12, 14, 16]. Feature-side responses adapt the representation rather than the model. Channel information conversion maps recordings between devices [21], band-wise statistics matching aligns the first- and second-order statistics of each frequency band of the target to those of the source without any retraining [19], and feature projection removes the device subspace [20]. Our blind estimator is closest in spirit to [19] and to the classical channel-normalisation view of RASTA filtering [22], but we use the estimated response constructively, to synthesise intermediate training domains, rather than subtractively at test time.

2.2. Unsupervised Domain Adaptation

Classical UDA theory bounds target error by source error, a divergence between the domains, and the disagreement of the optimal labelling functions [23]. Discrepancy-based methods minimise a kernel two-sample statistic [24, 25] or match second-order statistics [26]; adversarial methods train the encoder to defeat a domain classifier [27, 28]. Conditional adversarial adaptation reduces class mixing by conditioning the discriminator on the outer product of features and predictions [29], and decision-boundary views such as maximum classifier discrepancy provide complementary signals [30]. Semantic alignment through class centroids [31] and self-training with entropy or information-maximisation objectives [32, 33] are frequently combined with the above. In audio, adversarial UDA for ASC was introduced in [17] and extended with a Wasserstein critic in [18].

2.3. Intermediate and Gradual Domains

Gradual domain adaptation studies sequences of unlabelled distributions whose adjacent gaps are much smaller than the end-to-end gap; self-training along such a path enjoys favourable error-propagation guarantees [34], and later work characterises optimal paths [35] and removes the assumption that intermediate data are already indexed [36]. Related constructions insert a learned bridge inside the adversarial game [37] or synthesise intermediate distributions for dense prediction [38]. For cross-device ASC, Guo et al. [39] instantiate the idea physically, generating bridge devices by progressively stronger transfer functions applied to labelled source waveforms so that scene labels are inherited exactly, and combining bridge supervision with a conditional four-domain adversary and ordered class-centroid constraints. We retain that overall philosophy and depart from it in three respects: our bridges follow a blind estimate of the actual target response instead of chosen operators, the path is continuous rather than a fixed pair of stages, and the discriminator predicts a scalar position along the path instead of a categorical domain index.

3. Proposed Method

Fig. 1 summarises the training graph.

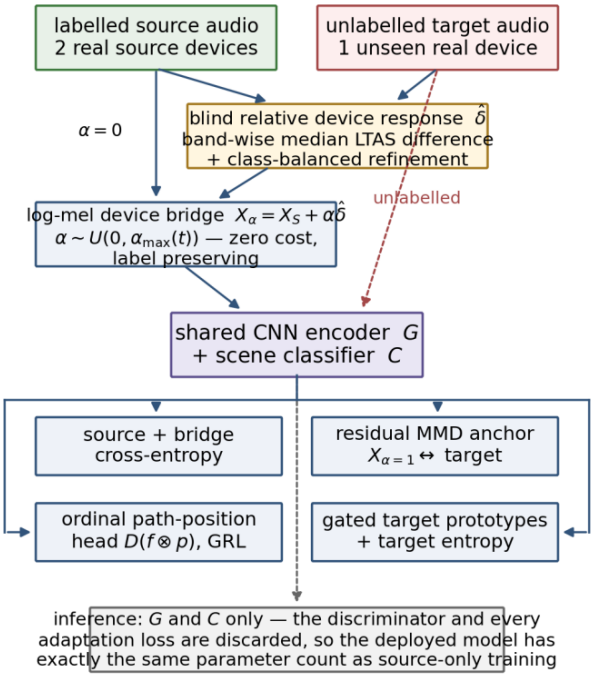

Figure 1. BRIDA training graph. The relative device response is estimated from unlabelled audio; fractional powers of it generate a continuum of label-preserving bridges directly in the log-mel domain; the discriminator regresses the position along that path. Inference uses the encoder and classifier only.

3.1. Problem setting

Let the labelled source set contain recordings of K scene classes captured by one or more devices, and let the unlabelled target set contain recordings of the same classes captured by a single device that is unseen during training. Write X for the log-mel spectrogram in decibels with M mel bands. A shared encoder G maps X to a feature f, and a linear classifier C produces posteriors p = softmax(C(f)). No target label is used at any point in training; target labels exist only to compute the reported metrics.

3.2. Blind Relative Device Response

If a source and a target recording of identical acoustic content differ only by the linear responses of the two devices,

their short-time magnitude spectra differ by a multiplicative factor, which is an additive offset in decibels. Real corpora are unpaired, but the offset is still recoverable in expectation from long-term averages when the two device populations contain comparable scene content. Writing L(X) for the temporal median of X across frames, we take a band-wise median across recordings and define the relative response as

$$\hat{\delta} = \text{med}_{D_T}\ L(x) - \text{med}_{D_S}\ L(x) \quad (1)$$

Followed by removal of the band average, which is redundant because per-example normalisation already discards broadband gain. Medians are used twice on purpose: the temporal median suppresses transient scene events that would otherwise dominate a mean spectrum, and the median across recordings suppresses atypical clips.

The estimator is unbiased only if the scene composition of the two unlabelled pools matches. To remove residual composition bias we add a single refinement step. After a warm-up, target clips whose maximum posterior exceeds a threshold are grouped by predicted class, and the class-conditional differences are recombined using the source class priors,

$$\hat{\delta}' = \Sigma_c\ w_c\ (\ \text{med}_{\hat{D}_{T,c}}\ L - \text{med}_{D_{S,c}}\ L\ ) \quad (2)$$

With w_c proportional to the number of labelled source clips of class c; the refined estimate is averaged with the initial one. Because only class-conditional spectral statistics are used, a wrong pseudo-label perturbs a median rather than injecting a wrong training label, which makes this step markedly safer than hard self-training. It is a safeguard rather than a core ingredient, and Section V-E reports that on our benchmark it never activated.

3.3. Fractional Device Bridges in the Log-Mel Domain

A bridge domain is a version of the labelled source data rendered as if it had passed through a device whose magnitude response is a fractional power of the estimated relative response. Applying a filter with magnitude $|H|^{\alpha}$ multiplies the power spectrum by $|H|^{2\alpha}$; if $|H|$ varies slowly within a mel band, the filterbank output is scaled by the same factor, so in decibels

$$X_{\alpha} \approx X_S + \alpha\ \hat{\delta}, \quad \alpha \in [0, 1] \quad (3)$$

Three properties follow. The construction is label preserving, because only the recording chain changes. It is anchored at both endpoints: $\alpha = 0$ recovers the source, and $\alpha = 1$ matches the long-term target spectrum by construction. And it costs one broadcast addition, so an unlimited number of intermediate devices is available without re-rendering audio or recomputing spectrograms. Section V-A quantifies the accuracy of the approximation against explicit waveform-domain filtering.

3.4. Continuous Device-Path Curriculum

Rather than fixing two bridge stages, we sample a fresh position for every bridge example. Let $\pi \in [0,1]$ denote training progress and $\alpha_{max}(\pi) = \min(1, \pi/r)$ with ramp r. Half of the bridge mini-batch is drawn from $U(0, \alpha_{max})$ and the other half is pinned at α_{max}: the uniform half densely covers the traversed part of the path, while the pinned half keeps the frontier populated. In preliminary runs, omitting the pinned half left the mean position near 0.3, so the bridge never reached the target response and the benefit of the path largely disappeared. Bridge examples are supervised with the inherited source labels, so with ℓ denoting the cross-entropy of the classifier output,

$$L_{cls} = E_S[\ \ell(X, y)\ ] + \lambda_b\ E_{\alpha}[\ \ell(X_{\alpha}, y)\ ] \quad (4)$$

3.5. Ordinal Path-Position Adversary

A categorical discriminator over a fixed number of domains cannot express a continuous path. We therefore keep the multilinear conditioning of [29], forming $q = f \otimes p$, and train a small head D to output a scalar whose sigmoid matches the true path position a: zero for source examples, α for a bridge example, one for target examples. With a gradient reversal layer scaled by $-\gamma$ under the standard logistic schedule, the adversarial term is a soft-label binary cross-entropy

$$L_{adv} = E[\ \text{softplus}(z) - a \cdot z\ ], \quad z = D(GRL_{\gamma}(q)) \quad (5)$$

Minimising over D estimates how far along the device path a representation lies; the reversed gradient asks the encoder to make that estimate impossible. Compared with a categorical alternative this head is smoother, requires no bridge indexing, and degrades gracefully when the estimated response is imperfect, because a mis-scaled estimate only relabels interior positions.

3.6. Residual Global Alignment and Gated Prototypes

Adversarial training alone can oscillate, so we retain a multi-kernel maximum mean discrepancy term as a low-variance global force [24, 25]. An important detail is that the discrepancy is computed between fully bridged source features at $\alpha = 1$ and target features rather than between raw source and target: the blind estimate has already removed the bulk of the first-order spectral shift, so the discrepancy term only has to close the residual gap. Finally, after a longer warm-up, target examples whose maximum posterior exceeds τ form confidence-gated class centroids that are pulled towards the labelled source and bridge centroids [31], and a small entropy term encourages decisive target predictions [33]. Both carry small weights. In preliminary runs a permissive threshold ($\tau = 0.70$) made the prototype term actively harmful, costing several accuracy points on the hardest device, which is consistent with the negative-transfer behaviour reported for early conditioning in [39]; we therefore gate it tightly and start it late.

3.7. Objective, Complexity and Inference

The complete objective is

$$L = L_{cls} + \alpha_m\ L_{MMD} + \beta\ L_{adv} + \xi\ L_{pro} + \varepsilon\ L_{ent} + \rho\ L_{mom} \quad (6)$$

Where L_mom matches the first and second feature moments of the source and bridge mini-batches. All adaptation terms and the discriminator are discarded after training, so the deployed model contains exactly the encoder and classifier used by source-only training, which matters in the low-complexity regime that dominates practical ASC [4]-[6].

4. Experimental Setup

4.1. Real Recordings and Real Device Responses

The audio corpus is ESC-50 [7], a collection of real

environmental recordings; we use the 1648 clips available in our environment, each 5 s long, and adopt the five coarse categories the corpus defines over its fifty classes (Animals; Natural/Water; Human non-speech; Interior/Domestic; Exterior/Urban) as the scene label space. Using a redistributable real corpus was a deliberate choice: the whole benchmark, including device rendering, can be regenerated from a public archive.

Device responses are 66 measured microphone impulse responses from the MicIRP library, redistributed under a Creative Commons licence with the DCASE'23 code release accompanying [15] and used as the source of device variability in [13]. Each response is normalised to unit energy so that only its spectral shape acts on the signal. We rank the responses by the standard deviation of their mel-band magnitude, take the 2 flattest as labelled source devices (IR_OktavaMK18Silver, Melodium_42B_1), and select 3 strongly deviating and mutually dissimilar responses as unseen target devices: T1 = Reslo_VMC2 (10.52 dB), T2 = IR_Lomo52A5M (9.35 dB), T3 = IR_Shure510C (8.92 dB), where the value in brackets is the root-mean-square log-spectral distance from the source mean. The remaining 61 responses form the pool used by the augmentation baseline and never contain a target device. Fig. 2(a) plots the selected responses and Fig. 2(b) illustrates the fractional bridges.

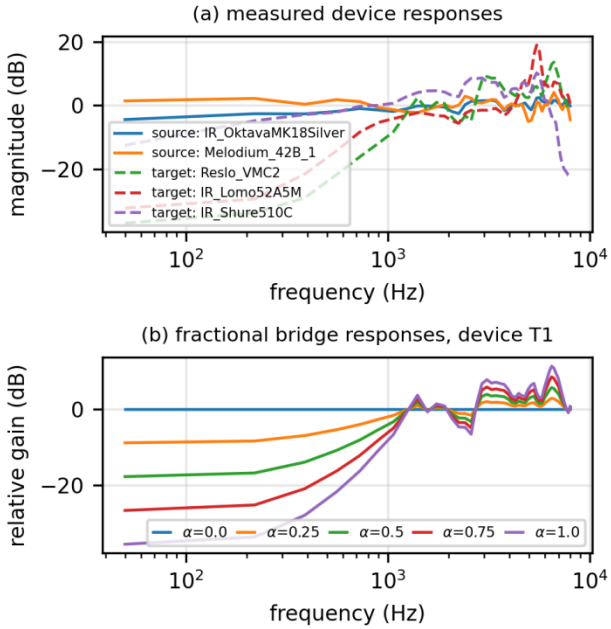

Figure 2. (a) Mel-band magnitude responses of the two source and three unseen target microphones, all measured. (b) Fractional bridge responses obtained by scaling the estimated relative response of device T1.

4.2. Leave-one-device-out Protocol

Clips are partitioned by the official ESC-50 folds so that no recording appears in more than one role. Folds 1 and 2 are rendered through the source devices and form the labelled training set; fold 3 is rendered through a single target device and forms the unlabelled target set; fold 4 is rendered through the same target device and is used only for evaluation, while its source-device rendering measures the retained source accuracy. Because the folds are disjoint, no source waveform reappears in the target domain, which removes the paired-content shortcut a purely synthetic benchmark can accidentally provide. Table 1 summarises the composition. The experiment is repeated for each of the 3 target devices and each of 3 seeds, giving 9 runs per method.

Table 1. Composition of the cross-device benchmark

Role	ESC-50 fold	Devices	Examples	Labels
Source training	1, 2	2 source	1600	yes
Target adaptation	3	1 unseen	400	no
Target evaluation	4	1 unseen	400	eval only
Source evaluation	4	2 source	800	eval only
Augmentation pool	1, 2	61 unused	3200	yes

4.3. Features, Network and Training

Audio is resampled to 16 kHz and converted to 48-band log-mel spectrograms with a 1024-point Hann window and a hop of 1024 samples. Training uses 2 s segments; at evaluation the posteriors of the four segments of a clip are averaged. Each segment is normalised by its own global mean and standard deviation, which removes broadband gain while preserving the spectral tilt that carries the device signature. The encoder has three 3×3 convolutional blocks with 16, 32 and 64 channels, each with batch normalisation [40], ReLU and 2×2 max pooling, followed by adaptive pooling, a 96-dimensional projection with dropout 0.2 and a linear classifier; the discriminator is a 64-unit head on the multilinear map. All methods share this backbone and are trained for 14 epochs with Adam [41], batch size 64, cosine-decayed learning rate 1e-3, weight decay 1e-4 and gradient-norm clipping at 5. Coefficients are $\lambda\_b = 1.0$, $\alpha\_m = 0.45$, $\beta = 0.035$, $\xi = 0.05$, $\varepsilon = 0.01$, $\rho = 0.02$, $r = 0.4$ and $\tau = 0.9$, and are shared across every device and seed.

4.4. Baselines

We compare against (i) source-only training; (ii) band-wise cepstral mean and variance normalisation, a training-free feature-domain counterpart of statistics matching [19]; (iii) DIR-Aug, convolving source audio with responses from the disjoint pool [13]; (iv) Freq-MixStyle, mixing frequency-wise statistics across examples [14, 16]; (v) DAN, a multi-kernel MMD objective [25]; (vi) DANN [27]; (vii) CDAN [29]; and (viii) a discrete intermediate-domain variant following the formulation of [39], with two fixed bridges at $\alpha = 1/3$ and 2/3, bridge supervision at the same weight as our method, an unconditional four-way domain discriminator and the same MMD anchor. Baselines (i)-(iv) never see target audio; (v)-(viii) do. All use identical data, backbone and schedule.

5. Results

5.1. Is the Log-Mel Bridge Faithful?

Table 2 compares the feature-domain bridge of (3) with an explicit waveform-domain rendering through a linear-phase filter whose mel magnitude response equals the scaled estimate. At full strength the applied spectral shift averages 6.95 dB, while the mean absolute discrepancy between the two constructions is 1.039 dB, about 14.9% of the shift. The discrepancy grows in proportion to α, so the relative error is essentially constant along the whole path, and part of it is attributable to the finite-length filter used for the reference rendering rather than to the mel approximation itself. The construction is therefore faithful enough that intermediate

devices can be synthesised inside the feature cache, which is what makes an arbitrarily fine path affordable.

Table 2. Log-mel bridge versus explicit waveform filtering

α	Applied shift (dB)	Mean abs. error (dB)	Relative error (%)
0.25	1.74	0.253	14.5
0.5	3.48	0.508	14.6
0.75	5.21	0.769	14.7
1	6.95	1.039	14.9

5.2. How Good is the Blind Estimate?

Table 3. Blind relative-response estimate against the measured response

Unseen device	Shift RMS (dB)	Correlation	RMSE (dB)
T1 Reslo_VMC2	10.52	0.998	0.74
T2 IR_Lomo52A5M	9.35	0.998	0.66
T3 IR_Shure510C	8.92	0.997	0.70

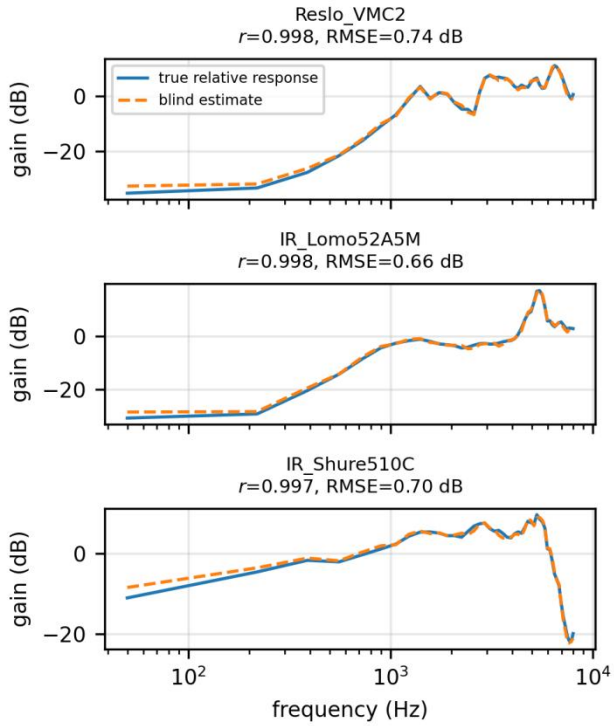

Figure 3. Blind estimate of the relative device response (dashed) against the response derived from the measured impulse responses (solid), for the three unseen target devices.

Fig. 3 overlays the blind estimate with the relative response computed from the measured impulse responses, which is available to us for validation only. Table 3 reports the agreement. Across the 3 devices the correlation lies between 0.997 and 0.998, and the root-mean-square error is 0.66-0.74 dB against shifts whose own magnitude is 8.92-10.52 dB root-mean-square. The estimator therefore recovers the direction and most of the magnitude of the device shift from unlabelled audio alone, which is exactly the input the bridge construction requires.

5.3. Main Comparison

Table 4 and Fig. 4 report target accuracy and macro-F1 averaged over 3 unseen devices and 3 seeds. Source-only training reaches 58.82 ± 2.12% on held-out source-device recordings but only 45.17 ± 7.40% on unseen devices, confirming a substantial device gap. Purely generative robustness helps without closing it: DIR-Aug and Freq-MixStyle reach 46.14 ± 7.45% and 46.03 ± 6.84%. Band-wise normalisation is the weakest option at 46.28 ± 1.41%, because discarding every band-wise statistic also discards a discriminative spectral profile; this is a useful reminder that a compensation which is beneficial as a test-time transform is not automatically beneficial as a training-time representation.

Among methods that consume unlabelled target audio, DAN reaches 46.31 ± 6.70%, DANN 49.33 ± 5.72% and CDAN 49.06 ± 5.73%. The discrete intermediate-domain baseline built on the estimated response is already strong at 52.39 ± 4.90%, reproducing the central claim of [39] in a setting with real device responses and unpaired content. BRIDA is the best method at 53.00 ± 4.29% accuracy and 52.99 ± 4.58% macro-F1, i.e. 7.83 accuracy points and 10.12 macro-F1 points above source-only training, 6.72 points above the best target-free baseline and 0.61 points above the discrete intermediate-domain variant. Across the 9 device-and-seed-matched pairs, BRIDA improves on the discrete variant in 6 cases with a mean paired difference of 0.61 points; with 3 seeds we report this as consistent directional evidence rather than a significance claim.

Table 4. Main comparison: mean ± sample standard deviation over all runs (%)

Method	Target acc.	Target macro-F1	Source acc.
Source-only	45.17 ± 7.40	42.87 ± 10.19	58.82 ± 2.12
Band-wise CMVN	46.28 ± 1.41	45.81 ± 1.39	45.97 ± 1.33
DIR-Aug [13]	46.14 ± 7.45	44.23 ± 9.91	57.14 ± 1.35
Freq-MixStyle [14, 16]	46.03 ± 6.84	43.99 ± 8.57	55.07 ± 1.56
DAN (MMD) [25]	46.31 ± 6.70	44.47 ± 8.91	59.15 ± 1.76
DANN [27]	49.33 ± 5.72	48.44 ± 7.01	59.07 ± 1.90
CDAN [29]	49.06 ± 5.73	48.23 ± 7.04	59.03 ± 1.56
Discrete IDA [39]	52.39 ± 4.90	52.31 ± 5.11	55.14 ± 2.51
BRIDA (proposed)	53.00 ± 4.29	52.99 ± 4.58	55.33 ± 1.92

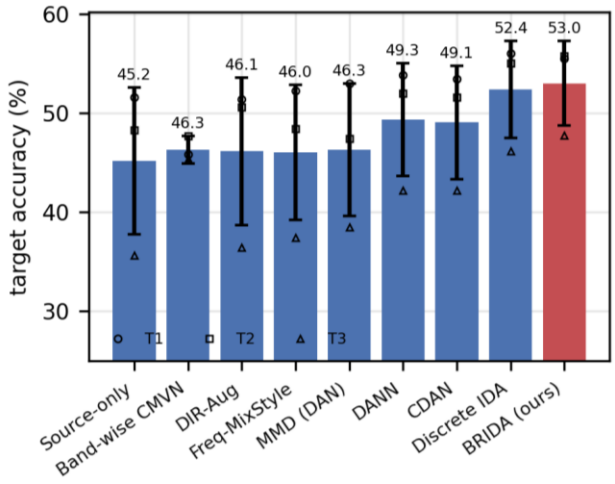

Figure 4. Target accuracy averaged over three unseen devices and three seeds. Error bars are the sample standard deviation over all runs of a method.

5.4. Per-device and Per-Class Behaviour

Table 5 breaks the comparison down by target device. Difficulty does not simply follow the overall log-spectral distance: Reslo_VMC2 has the largest distance from the source (10.52 dB) yet the smallest source-only degradation, which indicates that where a response deviates matters more than how much it deviates in aggregate. The margin of the proposed method nevertheless tracks the source-only degradation exactly: 3.92 points on T1, 7.50 points on T2, 12.08 points on T3, ordered identically to how far source-only training falls on each device. That is the expected signature of a mechanism acting on the device shift itself rather than on generic regularisation. Fig. 5 shows aggregate per-class recall. Source-only training is weakest on Interior/Domestic (38%) and Human non-speech (41%); the proposed method lifts them to 50% and 60% while leaving the categories that were already easier essentially unchanged, which is why the macro-F1 gain (10.12 points) exceeds the accuracy gain (7.83 points).

Table 5. Target accuracy per unseen device (%)

Method	T1	T2	T3	Mean
Source-only	51.58 ± 2.36	48.25 ± 1.00	35.67 ± 1.01	45.17
Band-wise CMVN	45.83 ± 1.15	47.67 ± 0.38	45.33 ± 1.38	46.28
DIR-Aug [13]	51.42 ± 1.51	50.58 ± 1.88	36.42 ± 1.81	46.14
Freq-MixStyle [14, 16]	52.25 ± 2.38	48.42 ± 0.52	37.42 ± 1.81	46.03
DAN (MMD) [25]	53.00 ± 3.54	47.42 ± 0.76	38.50 ± 2.38	46.31
DANN [27]	53.83 ± 3.26	52.00 ± 0.87	42.17 ± 1.13	49.33
CDAN [29]	53.42 ± 4.45	51.58 ± 0.76	42.17 ± 1.38	49.06
Discrete IDA [39]	56.00 ± 2.14	55.00 ± 0.25	46.17 ± 1.91	52.39
BRIDA (proposed)	55.50 ± 2.29	55.75 ± 1.56	47.75 ± 1.98	53.00

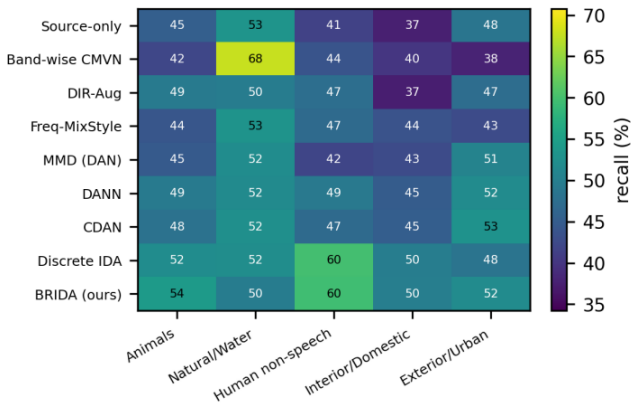

Figure 5. Aggregate per-class recall on the unseen target devices, averaged over devices and seeds.

5.5. Ablation

Table 6 and Fig. 6 remove one component at a time, comparing against the full model on exactly the same device-and-seed runs. Bridging only at the endpoint instead of traversing the path costs 0.81 accuracy points, the largest single effect we observe: covering the interior of the device path, not merely compensating its endpoint, is what stabilises the representation, which mirrors the hierarchy advocated in [39]. Removing the ordinal path-position adversary changes accuracy by +0.22 points and removing the gated prototype and entropy terms changes accuracy by +0.39 points; both magnitudes are smaller than the seed-to-seed spread of the full model, so on this benchmark we report them as neutral rather than as demonstrated contributors, and we do not claim they are required. Dropping the class-balanced refinement of the blind estimate changes nothing whatsoever, and a diagnostic run explains why: with a five-way task and a compact backbone, no target clip reached the confidence gate $\tau = 0.9$ at the refinement epoch, so the step never activated. We retain it in the description because it is the natural safeguard when the two unlabelled pools differ in scene composition, but we make no empirical claim for it here. Finally, replacing the blind estimate with the relative response computed from the measured impulse responses, an oracle a deployed system cannot have, changes accuracy by +0.83 points: this bounds the cost of blind estimation and indicates that the estimator is not the bottleneck of the pipeline. Fig. 6 plots the same quantities as paired differences. Every one of them, including the curriculum, has a sample standard deviation across the nine matched runs that spans zero, so the figure should be read as an ordering of point estimates rather than as a set of individually established effects.

Table 6. Ablation of BRIDA components (%)

Variant	Target acc.	Target macro-F1	Δ acc.
w/o path curriculum (α fixed to 1)	52.19 ± 4.15	52.17 ± 4.32	-0.81
w/o path-position adversary	53.22 ± 3.78	53.22 ± 4.05	+0.22
w/o gated prototypes / entropy	53.39 ± 4.14	53.41 ± 4.34	+0.39
w/o estimator refinement	53.00 ± 4.29	52.99 ± 4.58	0.00
BRIDA (full)	53.00 ± 4.29	52.99 ± 4.58	0.00
measured-response oracle	53.83 ± 3.29	53.99 ± 3.31	+0.83

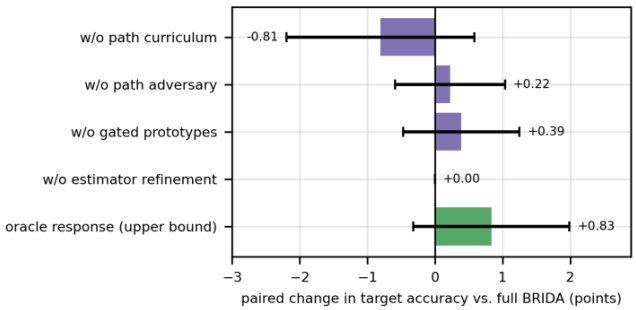

Figure 6. Ablation of the components of BRIDA. The oracle bar replaces the blind estimate by the relative response derived from the measured impulse responses.

5.6. Complexity

Table 7 lists parameter counts and measured single-core CPU training times. Adversarial variants add 31,044 discriminator parameters during training, so the training graph holds 128,761 parameters, but every method deploys the same 97,717 inference parameters. Bridge construction adds no feature extraction because it is a broadcast addition in the cached log-mel tensor, so the additional cost of BRIDA over source-only training stems only from the extra forward and backward passes on the bridge and target mini-batches.

Table 7. Complexity and measured training cost

Method	Train params	Infer. params	Mean s / run
Source-only / DAN	97,717	97,717	47.8
DANN / CDAN	128,761	97,717	70.3
Discrete IDA [39]	128,761	97,717	96.7
BRIDA (proposed)	128,761	97,717	96.5

6. Discussion and Limitations

The results support a simple picture. Device shift has a large, low-dimensional, first-order component: a smooth relative magnitude response. That component is identifiable from unlabelled audio, and once identified it is far more useful as a generator of supervised intermediate data than as a test-time correction, because the labels of the source recordings transfer for free to every point of the generated path. What remains after the first-order component has been traversed is a smaller residual; in our ablation a light adversarial and prototype treatment neither helps nor hurts beyond run-to-run variation, which suggests that on this benchmark almost all of the recoverable device shift is first order. This division of labour also explains why an aggressive conditional adversary applied directly to the raw gap can be counter-productive, and why the ordered, gradual treatment advocated by Guo et al. [39] transfers so well to a setting in which the intermediate devices are estimated rather than designed.

Several limitations should be stated plainly. First, device shift is simulated by convolving real recordings with measured microphone impulse responses. Those responses are real and the resulting mismatch is genuinely a device mismatch, but a linear time-invariant model omits automatic gain control, nonlinear compression, codec artefacts and device-dependent noise floors, all of which occur in real mobile recordings; covering the last of these would require adding a noise-floor term to the estimator. Second, the label space is the five coarse ESC-50 categories rather than the ten urban scenes of TAU Urban Acoustic Scenes, so absolute accuracies are not comparable with DCASE submissions; validating the mechanism on TAU with low-resource devices held out is the natural next step, and the estimator applies there unchanged. Third, we report three seeds and three target devices with a compact backbone; the margin over the discrete intermediate-domain variant is consistent in sign but modest relative to seed variance, and a larger study with at least five seeds, several backbones and preregistered coefficients would be needed for a significance claim. Fourth, the estimator assumes comparable scene composition in the two unlabelled pools. The class-balanced refinement is intended to mitigate that assumption, but it never activated in our runs because the model is not confident enough for the gate we set, so the assumption is in effect untested here; a benchmark with deliberately mismatched target priors, and an open-set target requiring an explicit rejection mechanism, are both left to future work.

7. Conclusion

We introduced BRIDA, an intermediate-domain unsupervised adaptation method for cross-device acoustic scene classification whose bridge domains are derived from a blind estimate of the relative device response instead of hand-specified device operators. Because a magnitude response is additive in the log-mel domain, fractional powers of the estimate yield a continuum of label-preserving intermediate devices at negligible cost, which we traverse with a curriculum and regularise with an ordinal path-position adversary, a residual MMD anchor and confidence-gated prototypes. On a leave-one-device-out benchmark built from real recordings and measured microphone impulse responses with disjoint source, target and evaluation content, the method attains 53.00 ± 4.29% target accuracy and 52.99 ± 4.58% macro-F1, improving on source-only training by 7.83 points and on a discrete two-bridge variant by 0.61 points without any inference-time overhead. Extending the estimator with a noise-floor component and validating it on recorded multi-device corpora are the immediate directions for future work.

References

[1] Barchiesi, D., Giannoulis, D., Stowell, D., & Plumbley, M. D. (2015). Acoustic scene classification: Classifying environments from the sounds they produce. IEEE Signal Processing Magazine, 32(3), 16–34.

[2] Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., & Plumbley, M. D. (2015). Detection and classification of acoustic scenes and events. IEEE Transactions on Multimedia, 17(10), 1733–1746.

[3] Mesaros, A., Heittola, T., & Virtanen, T. (2018). A multi-device dataset for urban acoustic scene classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE) (pp. 9–13).

[4] Zhang, H. (2024). Graph-aware multi-task learning for early prediction of timing violations and routing congestion in digital IC physical design. World Journal of Information Technology, 2(1), 13–20.

[5] Wang, J., Tan, Y., Jiang, B., Wu, B., & Liu, W. (2025). Dynamic marketing uplift modeling: A symmetry-preserving framework integrating causal forests with deep reinforcement learning for personalized intervention strategies. Symmetry, 17(4), 610.

[6] Zhao, X., Liu, J., Wang, Y., & Wang, J. (2026). CryptoMamba-SSM: Linear complexity state space models for cryptocurrency volatility prediction. IEEE Open Journal of the Computer Society, 7, 226–243.

[7] Wang, J., Liu, J., Zheng, W., & Ge, Y. (2025). Temporal heterogeneous graph contrastive learning for fraud detection in credit card transactions. IEEE Access.

[8] Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access, 12, 116546–116557.

[9] Zhang, H. (2025). Reinforcement learning approaches for layout optimization in electronic design automation with electromagnetic compatibility constraints. Frontiers in Robotics and Automation, 2(2), 77–93.

[10] Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-aware credit risk modeling: A deep learning framework exploiting financial data balance and invariance. Symmetry, 17(3), 341.

[11] Chen, J., Cui, Y., Zhang, X., Yang, J., & Zhou, M. (2024). Temporal convolutional network for carbon tax projection: A data-driven approach. Applied Sciences, 14(20), 9213.

[12] Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. Asian Business Research Journal, 10(11), 1–13.

[13] Chen, J., Wang, M., & Sun, T. (2025). Intelligent tax systems and the role of natural language processing in regulatory interpretation. American Journal of Machine Learning, 6(4), 74–94.

[14] Chen, Z., Liu, J., & Chen, J. (2025). Machine learning methods for financial forecasting in enterprise planning: Transitioning from rule-based models to predictive analytics. Frontiers in Artificial Intelligence Research, 2(3), 541–564.

[15] Chen, J., Liu, J., Liang, Y., & Zhou, M. (2026). KE-MLLM: A knowledge-enhanced multi-sensor learning framework for explainable fake review detection. Applied Sciences, 16(6), 2909.

[16] Yue, X., Yang, J., & Liu, W. (2026). FraudDebate-Agent: A multi-agent LLM framework with an evidence-based debate mechanism for financial statement fraud detection. Mathematics, 14(15), 2695.

[17] Zhu, R., & Yue, X. (2024). An explainable machine learning framework for predicting dividend sustainability and financial risk of U.S. REITs under interest-rate shocks. Journal of Trends in Finance and Economics, 1(2), 48–57.

[18] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. World Journal of Management Science, 2(1), 63–72.

[19] Zi, B. (2024). Cloud-native distributed systems for real-time payment intelligence. AI and Data Science Journal, 1(1), 51–56.

[20] Zhang, S., & Qiu, L. (2023). An interpretable, uncertainty-aware framework for bridge deterioration prediction and risk-based maintenance prioritization. Academic Journal of Architecture and Civil Engineering, 1(2), 21–28.

[21] Jin, J., Xing, S., Ji, E., & Liu, W. (2025). Xgate: Explainable reinforcement learning for transparent and trustworthy API traffic management in IoT sensor networks. Sensors, 25(7), 2183.

[22] Xing, S., Wang, Y., & Liu, W. (2025). Multi-dimensional anomaly detection and fault localization in microservice architectures: A dual-channel deep learning approach with causal inference for intelligent sensing. Sensors, 25(11), 3396.

[23] Xing, S., & Wang, Y. (2025). Proactive data placement in heterogeneous storage systems via predictive multi-objective reinforcement learning. IEEE Access, 13, 117986–117998.

[24] Ji, E., Wang, Y., Xing, S., & Jin, J. (2025). Hierarchical reinforcement learning for energy-efficient API traffic optimization in large-scale advertising systems. IEEE Access.

[25] Xing, S., & Wang, Y. (2025). Cross-modal attention networks for multi-modal anomaly detection in system software. IEEE Open Journal of the Computer Society.

[26] Liu, Y., Ren, S., Wang, X., & Zhou, M. (2024). Temporal logical attention network for log-based anomaly detection in distributed systems. Sensors, 24(24), 7949.

[27] Ren, S., Jin, J., Niu, G., & Liu, Y. (2025). ARCS: Adaptive reinforcement learning framework for automated cybersecurity incident response strategy optimization. Applied Sciences, 15(2), 951.

[28] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. IEEE Access, 12, 195974–195985.

[29] Hu, X., Zhao, X., Wang, J., & Yang, Y. (2025). Information-theoretic multi-scale geometric pre-training for enhanced molecular property prediction. PLOS ONE, 20(10), e0332640.

[30] Wang, M., Zhang, X., Yang, Y., & Wang, J. (2025). Explainable machine learning in risk management: Balancing accuracy and interpretability. Journal of Financial Risk Management, 14(3), 185–198.

[31] Yang, J., Li, P., Cui, Y., Han, X., & Zhou, M. (2025). Multi-sensor temporal fusion transformer for stock performance prediction: An adaptive Sharpe ratio approach. Sensors, 25(3), 976.

[32] Zhao, X., Sun, T., Ren, S., Yang, J., & Liu, Y. (2025). RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment. Frontiers in Artificial Intelligence Research, 2(3), 501–520.

[33] Lin, Y., Wang, Y., & Bian, M. (2024). XOR-EW: An explainable and cost-aware artificial intelligence framework for operational risk early warning in financial institutions. AI and Data Science Journal, 1(1), 94–102.

[34] Zhang, C., & Wang, Y. (2024). A data-driven multi-objective production scheduling framework for battery pack manufacturing under material arrival uncertainty. AI and Data Science Journal, 1(1), 86–93.

[35] Rhee, M., Zou, J., & Qin, Y. (2024). CLARITY: A closed-loop framework for duplicate-aware failure prediction and budgeted automated remediation in cloud infrastructure. AI and Data Science Journal, 1(1), 76–85.

[36] Wang, Y., & Zhang, C. (2023). AERIS: Adaptive executor-scoped resource management for privacy-preserving

data processing in Apache Spark. Eurasia Journal of Science and Technology, 5(2), 25–33.

[37] Qin, Y., Zou, J., & Rhee, M. (2022). An AI-driven intelligent transaction execution framework for internet finance platforms: Cost-sensitive real-time integrity interdiction. Eurasia Journal of Science and Technology, 4(2), 36–44.

[38] Zou, J., Qin, Y., & Rhee, M. (2024). Beyond model scale: A task-aware reliability benchmark for large language models in clinical decision support. Innovation and Technology Studies, 1(1), 39–47.

[39] Bian, M., Wang, Y., & Lin, Y. (2024). Explainable supplier risk prediction for critical telecommunications infrastructure procurement: An empirical XGBoost–SHAP framework. Social Science and Management, 1(1), 64–77.

[40] Jiao, Y., & Liang, Y. (2024). An integrated process-mining and explainable machine-learning framework for bottleneck and internal-control risk detection in the multi-entity financial close. Social Science and Management, 1(3), 96–107.