In Search for a Generalizable Method for Source Free Domain Adaptation

Malik Boudiaf [1 2] **Tom Denton** [3] **Bart van Merriënboer** [3] **Vincent Dumoulin**[* 3] **Eleni Triantafillou**[* 3]

Abstract

Source-free domain adaptation (SFDA) is compelling because it allows adapting an off-the-shelf model to a new domain using only unlabelled data. In this work, we apply existing SFDA techniques to a challenging set of naturally-occurring distribution shifts in bioacoustics, which are very different from the ones commonly studied in computer vision. We find existing methods perform differently relative to each other than observed in vision benchmarks, and sometimes perform worse than no adaptation at all. We propose a new simple method which outperforms the existing methods on our new shifts while exhibiting strong performance on a range of vision datasets. Our findings suggest that existing SFDA methods are not as generalizable as previously thought and that considering diverse modalities can be a useful avenue for designing more robust models.

1. Introduction

Deep learning has made significant progress on a wide range of application areas. An important contributing factor has been the availability of increasingly larger datasets and models (Kaplan et al., 2020; Song et al., 2022). However, a downside of this trend is that training state-of-the-art models has also become increasingly expensive. This is not only wasteful from an environmental perspective, but also makes the training of such models inaccessible to some practitioners due to the prohibitive resources required, or potential difficulties with data access. On the other hand, directly reusing already-trained models is often not desirable, as their performance can degrade significantly in the presence of distribution shifts during deployment (Geirhos et al., 2020). Therefore, a fruitful avenue is designing *adaptation methods* for pre-trained models to succeed on a new target domain, without requiring access to the original (*source*) training data ("source-free"). Preferably this adaptation can be performed *unsupervised*. This is the problem of *source-free domain adaptation* (SFDA) that we target in this work.

Several models have been proposed recently to tackle SFDA. However, we argue that *evaluation* in this area is a significant challenge in and of itself: We desire SFDA methods that are *general*, in that they can be used for different applications to adapt an application-appropriate pre-trained model to cope with a wide range of distribution shifts. Unfortunately, the standard SFDA evaluation protocols focus on a narrow set of shifts in vision tasks, leaving us with a limited view of the relative merits among different methods, as well as their generalizability. In this work, we address this limitation by studying a new set of distribution shifts. We expand on the existing evaluation methods, in order to gain as much new information as possible about SFDA methods. We also argue that we should target distribution shifts that are naturally-occurring as this maximizes the chances of the resulting research advances being directly translated into progress in real-world problems.

To that end, we propose to study a new set of distribution shifts in the *audio* domain. We use a bird species classifier trained on a large dataset of bird song recordings as our pre-trained model. This dataset consists of *focalized recordings*, where the song of the identified bird is at the foreground of the recording. Our goal is to adapt this model to a set of passive recordings (*soundscapes*). The shift from focalized to soundscape recordings is substantial, as the recordings in the latter often feature much lower signal-to-noise ratio, several birds vocalizing at once, as well as significant distractors and environmental noise like rain or wind. In addition, the soundscapes we consider originate from different geographical locations, inducing extreme label shifts.

Our rationale for choosing to study these shifts is threefold. Firstly, they are challenging, as evidenced by the poor performance on soundscape datasets compared to focalized recordings observed by Goëau et al. (2018); Kahl et al. (2021). Secondly, they are naturally occurring and any progress in addressing them can support ecological monitoring and biodiversity conservation efforts and research. Finally, our resulting evaluation framework is "just different enough" from the established one: It differs in terms of i) the modality (vision vs. audio), ii) the problem setup (single-

[*]Equal contribution [1]ÉTS Montréal [2]Work done while interning at Google [3]Google Research, Brain Team. Correspondence to: Malik Boudiaf <malik.boudiaf.1@etsmtl.net>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

label vs multi-label classification), and iii) the degree and complexity of shifts (we study extreme covariate shifts that co-occur with extreme label-space shifts). But it's not out of reach: SFDA methods designed for vision can be readily applied in our framework, since audio is often represented as spectrograms and thus can be treated as images.

We perform a thorough empirical investigation of established SFDA methods on our new shifts. Interestingly, the relative performance of established approaches varies significantly from observations made in common vision benchmarks, and in some cases we observe a *degradation* with respect to the pre-trained model's baseline performance. This striking finding leads us to explore simple modelling principles which we demonstrate result in consistently strong performance in the bioacoustics task considered and (importantly) in vision benchmarks as well. In the presence of extreme shifts, we observe that the confidence of the pre-trained model drops significantly (possibly also leading to miscalibration), which poses a challenge for entropy-based approaches. On the other hand, Noisy Student (Xie et al., 2020) and similar approaches are less sensitive to low model confidence but exhibit poor stability and require careful early-stopping—which is infeasible for SFDA because domain-specific validation data is unavailable. We hypothesize that leveraging the model's *feature space* as an additional "source of truth" helps stabilize adaptation, as this space carries rich information about the relationship between examples. In particular, we propose adding a Laplacian regularizer to Noisy Student, which we name NOisy student TEacher with Laplacian Adjustment (NOTELA).

Our contributions are: (i) we evaluate existing SFDA approaches on a challenging benchmark derived from a bioacoustics task and make observations on their generalizability which were not previously surfaced by common vision benchmarks; (ii) stemming from these observations, we advocate for the necessity of expanding the scope of SFDA evaluation in terms of modalities and distribution shifts; and (iii) we exemplify the benefits of this expanded scope by exploring simple modelling principles which, when combined, yield a more generalizable SFDA method.

2. Related Work

See Table 6 in the Appendix for a summary.

Domain adaptation (DA). DA assumes a setting in which labelled data is available for a source domain, and unlabelled data for a target domain. The goal is to maximize performance on the target domain. DA methods can be roughly divided into three types (Sagawa et al., 2022): *domain-invariant training* (also called *feature alignment*) aims to ensure that the features generated by the model for the source and target domain are indistinguishable by some metric (Sun et al., 2016; Sun & Saenko, 2016; Tzeng et al., 2014; Long et al., 2015; Ganin et al., 2016; Long et al., 2018; Tzeng et al., 2017; Sankaranarayanan et al., 2018); *self-training* involves generating pseudo-labels for the unlabelled data (Xie et al., 2020); and *self-supervision* involves training an unsupervised/self-supervised model, later fine-tuned or jointly trained with supervision (Shen et al., 2022).

Source-Free Domain Adaptation (SFDA) and Test-time Adaptation (TTA). These methods additionally assume that the source data itself is not available, e.g., because of resource, privacy, or intellectual property concerns. The distinction between SFDA and TTA is subtle: the latter is transductive, motivated by an online setup where adaptation happens on (unlabelled) target examples as they appear and evaluation is subsequently performed on the same examples. SFDA considers an offline adaptation phase and the adapted model is then evaluated on held-out examples. In practice, though, the methods developed for either are similar enough to be applicable to both. Related problems also include black-box (Zhang et al., 2021), online (Yang et al., 2020), continual (Wang et al., 2022b), and universal (Kundu et al., 2020) source-free domain adaptation.

Of the three types of DA methods discussed above, self-training most easily transfers to the SFDA and TTA settings (Liang et al., 2020; Kim et al., 2021), and we focus on this category since it's also the most generalizable to new modalities. Other methods use output prediction uncertainty for adaptation (Yang et al., 2020; Wang et al., 2021; Roy et al., 2022) or generative training to transform target domain examples or synthesize new ones (Li et al., 2020; Hou & Zheng, 2020; Kurmi et al., 2021; Morerio et al., 2020; Sahoo et al., 2020). Interestingly, Boudiaf et al. (2022) show that previous methods suffer from large hyperparameter sensitivity, and may *degrade* the performance of the source model if not tuned in a scenario-specific manner; this violates the assumption that labelled target data is unavailable.

Test-time Training (TTT). TTT (Sun et al., 2020) is a related problem where, like in TTA, a pre-trained model is adapted on the target test examples using a self-supervised loss, before making a prediction on those examples. Unlike SFDA and TTA, though, TTT modifies the source training phase to incorporate a similar self-supervised loss there too.

Domain generalization (DG). In DG (Wang et al., 2022a), like in SFDA, the target domain is unknown. However, unlike SFDA, no adaptation set is available. Instead the aim is to train a robust source model which works directly on new target distributions. Another important distinction is that DG assumes that information about the source domain is available during deployment on the target domain. A popular strategy for DG is to increase the source model's generalizablity by exposing it to diverse "conditions" at training time via domain randomization (Tobin et al., 2017) or adversarial

data augmentation (Volpi et al., 2018; Zhou et al., 2020), or to learn domain-*invariant* representations by training to match all available training "environments" (Arjovsky et al., 2019; Creager et al., 2021), minimizing the worst-case loss over a set of such environments (Sagawa et al., 2020), or decomposing the latent space or model weights into domain-specific and domain-general components (Ilse et al., 2020; Khosla et al., 2012).

3. Background

3.1. Problem formulation

Notation. In SFDA for classification, we assume access to a pre-trained model $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^C$, where $\mathcal{X}$ denotes the input space and C the number of classes. This model was trained on a source dataset $\mathbb{D}_s$ sampled from a source distribution $p_s(\mathbf{x})$, and needs to be adapted to a shifted target distribution $p_t(\mathbf{x}) \neq p_s(\mathbf{x})$. We assume we only have access to unlabelled data $\mathbb{D}_t^{\text{adapt}}$ sampled from p_t. The goal is to formulate an adaptation procedure $\mathcal{A} : (\boldsymbol{\theta}_s, \mathbb{D}_t^{\text{adapt}}) \to \boldsymbol{\theta}_t$ that produces an adapted version of the original model using the unlabelled dataset. The adapted model's performance is then evaluated on held-out data $\mathbb{D}_t^{\text{test}}$ sampled from p_t.

Single- vs. multi-label classification. SFDA methods have traditionally addressed single-label classification, in which exactly one category of interest is present in a given sample. Multi-label classification relaxes this assumption by considering that any number of categories (or none) may be present in a given sample, which is common in real-world data. In the single-label case the output probability for sample $\mathbf{x}_i$ is noted $\mathbf{p}_{i,\boldsymbol{\theta}} = \text{softmax}(f_{\boldsymbol{\theta}}(\mathbf{x}_i)) \in [0,1]^C$. In the multi-label case, the predictions for each class are treated as separate binary classification problems, resulting in $\mathbf{p}_{i,\boldsymbol{\theta}} = [\sigma(f_{\boldsymbol{\theta}}(\mathbf{x}_i)), 1 - \sigma(f_{\boldsymbol{\theta}}(\mathbf{x}_i))]^\top \in [0,1]^{2\times C}$, where σ is the logistic function.

3.2. Bioacoustics task

We use *Xeno-Canto* (XC; Vellinga & Planqué, 2015) as the source dataset $\mathbb{D}_s$ for bird species classification in the audio domain. XC is a large collection of user-contributed recordings of wild birds from across the world. Recordings are *focal* (targeted recordings of an individual captured in natural conditions, instead of the *passive* capture of all ambient sounds). Each recording is labeled with the species of the targeted individual; other birds may appear in the recording.

To evaluate adaptation to distribution shifts, we use multiple collections of passive (also called *soundscape*) recordings from various geographical locations as our target datasets. The soundscape datasets exhibit major covariate and label distribution shift from the source dataset. By virtue of being passively recorded, the annotated species are more occluded by environmental noise and distractors. Additionally, the geographical concentration of the datasets means that only a subset of XC's large number of species is present in each dataset, and the label distribution of those species does not necessarily follow that of XC. As a result, models trained on focal recordings have trouble generalizing to soundscapes recordings (Goëau et al., 2018; Kahl et al., 2021).

3.3. Vision tasks

We evaluate on several vision robustness benchmarks, most of which are used by prior SFDA approaches: (i) CIFAR-10-C and ImageNet-C (Hendrycks & Dietterich, 2019), a collection of corruptions applied to the CIFAR-10 and ImageNet test sets spanning 15 corruption types and 5 levels of severity; (ii) ImageNet-R (Hendrycks et al., 2021), which consists of 30,000 images of 200 of ImageNet's classes obtained by querying for renditions such as "art", "cartoon", "graffiti", etc.; (iii) ImageNet-Sketch (Wang et al., 2019), which consists of 50,000 images from querying Google Images for "sketch of {class}" for all ImageNet classes; and (iv) VisDA-C (Peng et al., 2017), which contains images of 12 object classes spanning synthetic and real domains.

3.4. Evaluation methodology

Adaptation in SFDA is fully unsupervised and hence there is no annotated validation set for each target domain. This is a significant challenge for model selection and evaluation, as evidenced by recent work on domain generalization (Gulrajani & Lopez-Paz, 2021). In line with recommendations made by Gulrajani & Lopez-Paz (2021), we disclose the model selection strategy used in our work, which for simplicity is shared across evaluated approaches. We hold out one domain for audio and one domain for vision which are used for validation and hyperparameter selection (details in section 5). An extensive hyperparameter search is conducted for every approach. In line with the methodology prescribed by SFDA, we also partition the evaluation data for each distribution shift into the adaptation set $\mathbb{D}_t^{\text{adapt}}$ (75% of the data) and the test set $\mathbb{D}_t^{\text{test}}$ (25% of the data).

3.5. Categorization of approaches

We will evaluate several methods which have been proposed for (source-free) domain adaptation on our new shifts. We choose to investigate methods that fit two criteria: generality across tasks and modalities, and strong performance. In this section we present our categorization of methods we consider, which we then build upon in Section 4.

Entropy Minimization (EM). These methods enforce the *cluster* assumption, i.e., that the boundaries described by the model's head should not cross any high-density region of samples in the feature space. Geometrically, this enforces

large margins between the classifier's boundaries and the provided samples by encouraging the model to output increasingly confident predictions for the unlabelled samples (Grandvalet & Bengio, 2004).

We evaluate **TENT** (Wang et al., 2021) as a representative example of the EM approach. TENT adapts the source model by minimizing the entropy of its predictions through tuning the normalization layers' channel-wise scaling and shifting parameters, and updating the layers' population statistics estimates accordingly.

Teacher-Student (TS). This is a self-training paradigm where a teacher provides *pseudo-labels* for the unlabelled examples, and a student is then trained to predict them. More formally, TS minimizes

$$\min_{\mathbf{y}_{i:N},\boldsymbol{\theta}} \ \mathrm{Tr}\left(-\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^\top \log\left(\mathbf{p}_{i,\tau(\boldsymbol{\theta})}\right)\right) \quad \text{s.t} \quad \mathbf{1}^\top\mathbf{y}_i = \mathbf{1},\ \mathbf{y}_i \geq 0, \tag{1}$$

where $\mathbf{y}_i$ and $\mathbf{p}_{i,\tau(\boldsymbol{\theta})}$ represent the pseudo-label and the model's soft predictions for the sample $\mathbf{x}_i$.[1] The pseudo-label is dependent on the model's weight's, $\boldsymbol{\theta}$, which are transformed using a weight transformation, τ, which is typically set to the identity (which we denote Id) in TS methods.

Optimization happens in an alternating fashion between student and teacher updates: The *teacher-step* minimizes equation 1 w.r.t. the pseudo-labels $\{\mathbf{y}_i, \dots, \mathbf{y}_N\}$ while the *student-step* minimizes equation 1 w.r.t. $\boldsymbol{\theta}$ using gradient descent. Different TS methods utilize soft (Xie et al., 2020) or hard (Lee et al., 2013) pseudo-labels.

Intuitively, TS can be seen as an indirect way of minimizing entropy: By training to predict each example's pseudo-label (i.e., its *most likely* label, based on its confidence), the model reinforces its own predictions, thereby increasing its confidence. At the same time, though, it has a *consistency maximization* flavour: Because of the alternation between teacher and student updates, predicting the correct pseudo-label requires consistency throughout time.

We consider three methods from this category in our investigation: First, **pseudo-labelling** (PL; Lee et al., 2013) assigns pseudo-labels to unlabelled examples by picking the maximum-probability class according to the trained model. In the context of SFDA, this translates into assigning pseudo-labels to unlabelled target domain examples using the model trained on the source dataset. Second, we consider an adaptation of **DUST** (Khurana et al., 2021), originally developed for DA (instead of SFDA), for automated speech recognition. DUST is also based on pseudo-labels, but only uses *reliable* samples, as measured by the consistency of the predictions obtained by performing multiple stochastic forward passes through the model. Finally, **SHOT** (Liang et al., 2020) adapts the feature extractor to the target domain while freezing the classifier head. It performs adaptation through a combination of nearest-centroid pseudo-labelling and an information maximization loss.

Denoising Teacher-Student (DTS). This method builds upon the TS framework by adding some form of "noise" to the student, while keeping the teacher clean. Intuitively, predicting clean pseudo-labels from noisy student predictions leads to maximizing another type of consistency: between different views of the same inputs. Mathematically, DTS differs from TS by setting $\tau \neq \mathrm{Id}$ during the student's forward pass, while keeping it set to the identity during the teacher's forward pass.

Ideas related to DTS have been explored in semi-supervised (Xie et al., 2020; Miyato et al., 2018) and self-supervised learning (Grill et al., 2020; Chen et al., 2020). Notably, Noisy Student (Xie et al., 2020) is a popular representative that we build upon. However, to keep the approach light, both in terms of computation and hyper-parameter load, we consider a simplified model in our investigation, where the same network is used for both the teacher and the student, and dropout is the sole source of noise. We refer to this variant as **dropout student (DS)**.

Manifold regularization (MR). Manifold regularization exploits the assumption that target data forms well-defined clusters in the feature space of the source model, by explicitly enforcing the *cluster assumption*. From this family of methods, we consider **NRC** (Yang et al., 2021), which forces a target sample and its nearest-neighbors, as well as its 'extended nearest-neighbors', to have similar predictions.

Our proposed model, presented in section 4, combines elements of MR and DTS.

4. Laplacian Adjustment

As we will demonstrate in subsection 5.1, existing SFDA methods struggle to perform consistently well on bioacoustics distribution shifts. This motivates the exploration of a combination of simple modelling principles which we hypothesize will result in consistently strong performance on different datasets, problem settings and modalities. We name this combination NOisy student TEacher with Laplacian Adjustment (**NOTELA**).

In our experiments, we observe that EM struggles in the face of severe distribution shifts resulting in models with very low confidence. We also note that TS consistently improves upon EM, signaling that consistency maximization

[1] Recall from Section 3.1 that in the multi-label case we consider C separate binary classification problems, and hence $\mathbf{y}_i$ and $\mathbf{p}_{i,\tau(\boldsymbol{\theta})}$ are $2 \times C$ matrices. The trace in Equation 1 sums the objectives for each of these C problems.

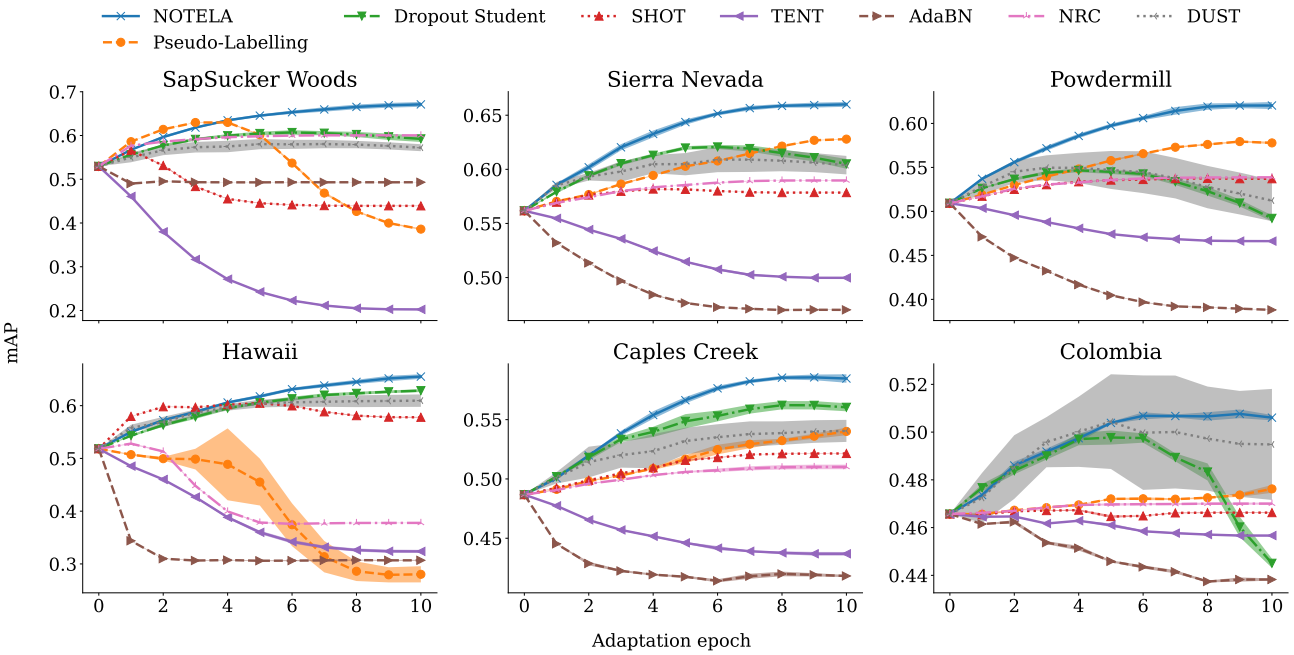

Figure 1. All but one SFDA method fail to consistently improve the source model in terms of mAP on $\mathbb{D}_t^{test}$ across distribution shifts. Dropout Student succeeds, but only through early stopping, which is infeasible in the SFDA setting. NOTELA achieves consistently stable convergence while improving upon Dropout Student's performance. Note that this analysis's purpose is to illustrate the failure modes of SFDA methods; those plots cannot be used for hyperparameter selection because we are looking at the test sets.

is a useful "auxiliary task", perhaps due to being more robust to low model confidence. The fact that DTS—with the additional consistency maximization "task" between clean and noisy views—further improves performance in our experiments strengthens this hypothesis. Starting from DTS, we seek to further improve it by utilizing another source of information that we hypothesize is also robust to low model confidence. Specifically, while the model outputs may have low confidence, its *feature space* may still carry useful information about the relationship between examples.

From this hypothesis, and drawing inspiration from classic ideas on manifold regularization (Belkin et al., 2006), we suggest probing the feature space directly as an *auxiliary source of truth*. NOTELA instantiates this idea by encouraging that nearby points in the feature space be assigned similar pseudo-labels, similar to NRC. We hypothesize this may also help with stability, since the targets that the student is asked to predict will vary less over time, due to the slower-changing similarity in feature space. NRC uses a sophisticated definition of neighbourhood which weighs reciprocal, non-reciprocal, and extended nearest neighbours differently. In contrast, NOTELA simplifies this by considering only reciprocal nearest-neighbours. NRC and NOTELA also differ by their class-marginal pseudo-label prior (uniform vs. no prior) and the loss function used (dot product vs. cross-entropy).

Formalization. We augment the denoising teacher-student formulation in equation 1 with a Laplacian regularization:

$$\min_{\mathbf{y}_{1:N},\boldsymbol{\theta}} \operatorname{Tr}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^\top\left[-\log\left(\mathbf{p}_{i,\tau(\boldsymbol{\theta})}\right)+\alpha\log(\mathbf{y}_i) - \lambda\sum_{j=1}^{N} w_{ij}\,\mathbf{y}_j\right]\right). \tag{2}$$

There are two changes in Equation 2 compared to Equation 1. First, we introduce a scalar weight $\alpha \in \mathbb{R}$ that controls the softness of pseudo-labels, which we treat as a hyperparameter. Second, we have added the third term that represents a Laplacian regularizer. The value w_{ij} denotes the *affinity* between samples i and j and is obtained by the penultimate layer of the network.

Optimization. Disregarding the pairwise Laplacian term allows to directly obtain a closed-form solution to Equation 2, namely $\mathbf{y}_i \propto \mathbf{p}_i^{1/\alpha}$. However, adding the pairwise affinities makes optimization more challenging. We simplify the problem by linearizing the Laplacian term, which allows us to recover a closed-form solution:

$$\mathbf{y}_i \propto \mathbf{p}_i^{1/\alpha} \odot \exp\left(\frac{\lambda}{\alpha}\sum_{j=1}^{N} w_{ij}\mathbf{p}_j\right). \tag{3}$$

The full proof of Equation 3 can be found in the Appendix. Furthermore, we show in Appendix F that under the assumption of positive semi-definite affinity matrix (w_{ij}), Equation 3 becomes an instance of a concave-convex procedure (CCP; Yuille & Rangarajan, 2003).

Complexity. Theoretically, the added Laplacian regularization scales quadratically in the number of samples N. In practice, we set $w_{ij} = w_{ji} = \frac{1}{d}$ where $0 < d \leq k$ is the number of mutual k-nearest neighbours (Brito et al., 1997) of samples i and j, and $w_{ij} = w_{ji} = 0$ if these samples are not in each other's k-nearest neighbours lists. Finding the k-nearest neighbours can be done with $\mathcal{O}(N \log N)$ average time complexity. Equation 3 scales as $\mathcal{O}(NCk)$, but can be fully vectorized across samples.

5. Experiments

5.1. Bioacoustics Task

Data processing and source model. XC is a large dataset containing a total of 10,932 species. We process XC recordings by resampling the audio to 32 kHz and extracting 6-second slices of relevant audio (see Appendix B for details). We process soundscape recordings by extracting 5-second slices using the provided bounding-box labels (see Appendix C for details).

During training, we take a random 5-second crop and the audio signal's gain is normalized to a random value between 0.15 and 0.25. A technique similar to LEAF (Zeghidour et al., 2021) is used to convert the waveform into a spectrogram. Unlike LEAF, we do not learn a Gaussian lowpass filter and instead apply the Gabor filters with a stride of 320. The resulting output is a power-spectrogram with a time resolution of 100 Hz.

The most comprehensive bird species classifier we are aware of is BirdNet (Kahl, 2019; Kahl et al., 2021), however the publicly available checkpoints are trained with a combination of focal and soundscape recordings (including many of the test sets considered in this paper),[2] which is incompatible with our SFDA methodology. We instead use an EfficientNet-B1 (Tan & Le, 2019) model we trained ourselves. The output of this model is flattened and projected into a 1280-dimensional embedding space. In addition to species prediction, the model is trained with three auxiliary losses for the bird's order, family, and genus (each having a 0.25 weight).

Metrics. Unless otherwise specified, we use sample-wise mean average precision (mAP) and class-wise mean average precision (cmAP) for evaluation. Both of these metrics are threshold-free and appropriate for multi-label scenarios. Each can be interpreted as a multi-label generalization

[2]Personal correspondence with the authors.

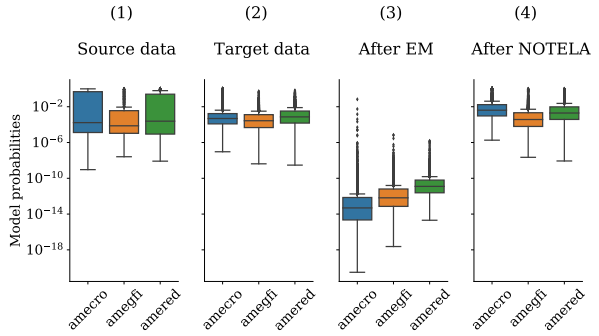

Figure 2. (1) and (2): When shifted to a new domain, the pre-trained model exhibits *lower confidence* than on the original source data, as seen from the lower per-class probability distributions assigned by the model. (3): Applying entropy minimization only worsens the problem, and leads to *general collapse*. (4): Our method does not suffer from the same phenomenon. Species names are represented using their corresponding eBird (Sullivan et al., 2009; 2014) species code.

of mean reciprocal rank, where ranking of model logits is performed either per-sample (for mAP) or per-class (for cmAP). See Appendix D in the appendix for formal definitions.

The mAP metric measures the ability of the model to assign higher logits to any species present in an example. By contrast, cmAP is the mean of the model's per-species classification quality (similar to the average of per-species AUC scores). Note that class-averaging in cmAP corrects for class imbalance, while mAP reflects the natural data distribution. To avoid noisy measurements we only consider species with at least five vocalizations in the dataset when computing cmAP.

Baselines. In addition to **TENT**, **pseudo-labelling (PL)**, **SHOT**, **dropout student (DS)**, **DUST** and **NRC** which we described in section 3.5, we also consider **AdaBN** (Li et al., 2018). This method recomputes the population statistics of batch normalization (Ioffe & Szegedy, 2015) layers in the pre-trained model using the unlabelled target dataset.

Hyperparameter selection. We use the High Sierras dataset as the held-out domain. For every method, we search over hyperparameters such as the learning rate and its schedule, the subset of parameters to adapt (all or only batch norm), and whether to use dropout during adaptation. Additionally, we search over specific hyperparameters such as the β weight in SHOT, the confidence threshold in PL, or the weights $\{\alpha, \lambda\}$ in NOTELA. The overview of tuned hyperparameters, resulting in $O(200)$ experiments per method, can be found in the appendix.

Results We found neither AdaBN nor TENT were able to improve the source model, regardless of the hyperparameter configuration or the domain (Table 1; also see Figure 1 for a

Table 1. Test results on the 6 test target domains (averaged over 5 random seeds) using hyperparameters selected on the validation domain.

Method	S. Nevada		Powdermill		Hawai'i		Caples		SSW		Colombia	
	mAP	cmAP	mAP	cmAP	mAP	cmAP	mAP	cmAP	mAP	cmAP	mAP	cmAP
Source	56.2	36.4	50.9	33.0	51.8	33.2	48.7	41.8	53.0	33.7	46.6	38.5
AdaBN	47.0	34.3	38.8	30.7	30.7	28.0	41.8	38.8	49.3	34.2	43.8	35.5
SHOT	57.8	37.8	53.7	33.1	57.7	29.1	52.1	43.7	43.9	24.9	46.6	39.3
TENT	50.0	35.1	46.6	32.7	32.4	25.7	43.7	39.6	20.2	14.0	45.7	37.6
PL	62.8	39.3	57.8	33.3	28.0	25.6	54.0	43.8	38.6	30.6	47.6	39.6
DS	60.5	37.2	49.2	33.3	62.8	**36.4**	56.0	41.7	59.2	33.2	44.5	41.6
DUST	60.4	35.1	51.2	31.5	61.0	35.7	54.1	40.3	57.2	32.3	49.5	35.6
NRC	59.0	38.0	53.9	33.3	37.8	28.7	51.0	43.5	60.0	39.3	47.0	39.5
NOTELA	**66.0**	**40.0**	**62.0**	**34.7**	**65.5**	36.2	**58.5**	**44.7**	**67.1**	**42.7**	**50.6**	**44.7**

Table 2. Test accuracy on the 6 test target domains in the *single-label* scenario (averaged over 5 random seeds) using hyperparameters selected on the validation domain.

Method	S. Nevada	Powdermill	Hawai'i	Caples	SSW	Colombia
Source	29.9	51.4	24.0	15.4	33.5	19.6
AdaBN	14.2	30.9	13.0	10.1	25.5	14.2
DUST	49.8	62.9	44.2	34.8	43.2	**32.7**
SHOT	23.9	15.2	6.8	17.6	21.9	18.9
TENT	7.3	74.3	29.9	8.0	5.4	5.0
PL	60.2	**93.1**	50.0	33.7	58.7	24.0
NRC	42.8	55.2	26.9	27.5	16.7	29.4
DS	62.5	44.6	**56.1**	39.3	49.4	27.7
NOTELA	**73.5**	90.5	29.1	**43.9**	**63.3**	**32.5**

Table 3. Test results on the validation domain (High Sierras) using hyperparameters selected on that domain.

Method	mAP	cmAP
Source	63.8	46.9
AdaBN	53.0	38.5
SHOT	67.5	50.3
TENT	53.5	38.6
PL	72.0	51.1
DS	76.8	52.7
DUST	76.0	51.6
NRC	66.1	49.3
NOTELA	**78.1**	**54.4**

hindsight analysis). This is surprising for TENT, since EM has also previously shown strong performance on various tasks (Wang et al., 2021; Vu et al., 2019). We hypothesize this is due to a combination of extreme distribution shift as well as the multi-label nature of our problem. Indeed, in the single-label scenario minimizing entropy leads to forcing the model to choose a single class for each example in an increasingly confident manner. On the other hand, in our multi-label scenario there is no constraint that a class should be chosen. This fact—combined with the low confidence caused by very large distribution shifts (Figure 2, (1,2))—can drive the model to a collapsed state where all class probabilities are zero (Figure 2, (3)).

As for SHOT and PL, we were able to find hyper-parameters that improved over the source model's performance on the validation set, as observed in Table 3. We observed however that the gains did not consistently translate to the test domains (Table 1). For example, PL boosted the source model by 8.2% mAP on the validation domain, while degrading it by more than 23% mAP on the Hawai'i test domain.

On the other hand, we find that DTS works significantly better on our challenging shifts, but suffers from stability issues: it often displays a plateau followed by a degradation of the model's performance (Figure 1; see e.g. Powdermill). NRC and DUST are also able to improve upon the source model in some cases, but are not consistent in achieving this for all of the shifts we study, and also suffer from the same plateau-and-degradation trend (e.g. see Hawaii for NRC and Powdermill for DUST). This trend would necessitate a precise early-stopping procedure to pinpoint the number of adaptation updates to perform before degradation starts.

Table 4. Validation and ablation results from the High Sierras bioacoustics dataset. Dropout Noise, Softness and Laplacian regularization ingredients **act symbiotically** to provide the best performances. Removing any ingredient leads to significant drops in performances.

Dropout	Softness	Laplace Reg.	mAP	cmAP
pre-trained model (no adaptation)			63.8	46.9
	✓	✓	61.3	33.3
✓		✓	72.1	49.4
✓	✓		77.0	52.0
✓	✓	✓	**78.1**	**54.4**

This is a serious drawback in the context of SFDA, given the absence of a domain-specific labelled validation set for tuning the training schedule or performing early-stopping.

In contrast, we find that NOTELA not only addresses these stability issues, but also outperforms all considered baselines, setting the state of the art on the bioacoustics shifts.

Disentangling the multi-label effect We previously conjectured that the multi-label nature of the bio-acoustics tasks contributed to SFDA methods' failure. We challenge that supposition by generating a new version of each bioacoustics dataset, in which all recordings containing more than a single bird annotation are filtered. That allows us to fall back onto the standard single-label setting, in which the model's logits can be constrained into a single probability distribution through a softmax operator. Results are presented in Table 2. While an exact apple-to-apples comparison with the multi-label case is not possible, because a lot of recordings have been filtered, we can already draw interesting insights. First, the single-label constraint (through the use of softmax) appears to be a strong regularizer that can substantially contribute to the success of SFDA methods. For instance, the pseudo-labelling baseline, which dramatically failed on 2 test cases in the multi-label scenario, now systematically outperforms the baseline, and even reaches a state-of-the art performance on Powdermill. However, Table 2 also confirms our premise that the multi-label factor cannot alone fully explain methods failing in the audio domain, as TENT and AdaBN still suffer from dramatic model collapses on several tests domains. Similarly, SHOT which obtains strong results on single-label visual tasks, as shown in subsection 5.2, performs poorly in this setting. Overall, NOTELA preserves its ability to systematically improve over the baseline, and obtains state-of-the-art results on 4 out of 6 target domains.

Ablations on NOTELA We find in Table 4 that the presence of the noise is crucial, and that removing it leads to worse performance than the non-adapted model. Furthermore, removing softness in labels ($\alpha = 0$) or Laplacian Regularization ($\lambda = 0$) significantly under-performs the full method, thereby highlighting the symbiosis of all three components.

5.2. Vision Tasks

While we have shown that recent SFDA approaches are not as generalizable as previously thought, we have yet to show that the simple NOTELA approach developed in the context of bioacoustics distribution shifts is generalizable enough to perform well on vision tasks.

Data processing and source models. We process vision datasets in accordance with their respective established practices. We adopt model architectures from previous works, namely a ResNet-50 (He et al., 2016) for ImageNet benchmarks, a Wide ResNet 28-10 (Zagoruyko & Komodakis, 2016) for CIFAR-10 benchmarks, and a ResNet-101 for VisDA-C. We use the same CIFAR-10 Wide ResNet model checkpoint (provided by Croce et al., 2021) as Wang et al. (2021) and the ResNet-101 checkpoint provided by Yang et al. (2021) for fair comparison.

Metrics and hyperparameter selection. We report top-1 accuracy for all vision SFDA benchmarks. We use the three most challenging corruptions from ImageNet-C (*contrast, glass blur and snow*) as the validation domain for all vision tasks (we use the average accuracy across those three).

Results. From Table 5, we find that NOTELA consistently exhibits strong performance on vision benchmarks, despite the change in modality and the single-label setup. It performs strongly on all datasets and in fact outperforms all baselines on both ImageNet-R and ImageNet-Sketch.

Interestingly, SHOT performs uncharacteristically well on VisDA-C, whereas NRC underperforms SHOT, TENT, and NOTELA. We note that several factors could explain this observation. We purposefully split VisDA-C's target domain into an adaptation and an evaluation set, unlike the TTA procedure used in the NRC paper. We also performed a full hyperparameter search for NRC (as well as all other approaches considered) using ImageNet-C as a validation domain, as described earlier, which naturally may yield less optimistic results compared to performing model selection on the target domain. In line with our overall takeaway, this is possibly indicative of a lack of performance consistency across different evaluation methodologies.

6. Discussion and Conclusion

We investigated the generalizability of recent source-free domain adaptation methods developed in the context of vision tasks. We observe that when applied to a new modality and problem setting (multi-label classification of audio recordings of wild bird vocalizations), these methods' performance characteristics differ greatly from observations

Table 5. Top-1 accuracy (averaged over 5 random seeds) on vision test benchmarks. NOTELA approaches best methods on CIFAR-10-C, and surpasses them on ImageNet variants. We report per-corruption confidence-intervals for CIFAR-10-C in Appendix G

Method	CIFAR-10-C (Average across severity 5 corruptions)	ImageNet-R	ImageNet-Sketch	Vis-DA-C
Source	56.66	23.16 ± 0.0	21.67 ± 0.0	51.8 ± 0.0
AdaBN (Li et al., 2018)	80.10	24.54 ± 0.08	22.55 ± 0.14	65.63 ± 0.3
SHOT (Liang et al., 2020)	82.63	24.49 ± 0.14	26.16 ± 0.1	$\mathbf{76.06} \pm 0.14$
TENT (Wang et al., 2021)	82.69	28.78 ± 0.04	27.56 ± 0.16	71.03 ± 0.25
PL (Lee et al., 2013)	**83.85**	26.54 ± 3.43	2.03 ± 1.5	50.16 ± 0.67
NRC (Yang et al., 2021)	80.58	25.73 ± 0.1	23.66 ± 0.1	69.13 ± 0.22
DUST (Khurana et al., 2021)	80.07	24.61 ± 0.09	22.54 ± 0.07	65.29 ± 0.21
Dropout Student	82.18	29.14 ± 0.14	27.49 ± 0.11	69.9 ± 0.18
NOTELA (ours)	82.21	$\mathbf{32.97} \pm 0.11$	$\mathbf{30.92} \pm 0.1$	72.08 ± 0.33

made when evaluating on vision tasks (and in some cases fail to live up to expectations set in the vision domain).

In light of this, our first message is that as more and more SFDA approaches are developed for vision tasks, they may become increasingly co-adapted to the characteristics of that domain, up to a point where progress there ceases to translate into progress on other modalities and problems. Which characteristics the methods co-adapt to is not yet clear, as the audio classification task considered in this work differs in many ways from the image classification tasks used as standardized SFDA benchmarks. Beyond the shift from vision to audio modality, and from a single-label to multi-label problem, other confounders could be responsible for our observations, such as the severe class imbalance, the extremity of the distribution shifts and the complex nature of bird vocalizations that have distinct modes (songs vs. calls) even within the same class.

Our work intends to draw attention to the surprising lack of generalizability of SFDA methods and encourage practitioners to expand the scope of their evaluation (for instance using our proposed bird song classification task). We also hope to encourage future work into properly characterizing which of the above factors contribute to the observed differences in performance characteristics, in order to gain a deeper understanding into existing SFDA methods and on how to improve them.

Our second message is that consistent and generalizable performance is a valuable attribute for SFDA approaches given the nature of the problem (no target labels, no well-defined validation set) and the resulting challenge in performing model selection. One may spend a lot of time and effort developing a strong approach for a particular pair of source and target domains only to see it underperform in a different context. Mitigating this danger requires careful research methodology, isolating whole development datasets for hyper-parameter tuning. From this perspective, an approach which performs consistently well in multiple contexts (even if sometimes worse than top-performing approaches in each individual domains) is valuable in practice. While we don't believe that NOTELA is the ultimate SFDA method, we believe it is a strong baseline in terms of this consistency desideratum while performing very competitively: it sets the state-of-the-art on our challenging bioacoustics shifts and is a strong competitor on vision benchmarks.

Moreover, we argue that beyond evaluation, our proposed bioacoustics task is important for model development in and of itself, as it surfaces differences in performance characteristics which, when addressed in our proposed NOTELA approach, resulted in the desired generalizable performance.

Acknowledgements

Each author of this paper contributed in the following way:

- Malik proposed the method presented in this work. He implemented it and all methods we compare against, conducted all experiments, and produced all figures and tables.
- Tom implemented several soundscapes datasets used by Malik for his experiments and was a significant contributor to the codebase used by Malik for his experiments.
- Bart implemented the bioacoustics classifier and was a significant contributor to the codebase used by Malik for his experiments.
- Vincent co-advised Malik on the project. He contributed the Xeno-Canto dataset implementation and was a significant contributor to the codebase used by Malik for his experiments.
- Eleni co-advised Malik on the project. She contributed to the framing of the proposed approach.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Belkin, M., Niyogi, P., and Sindhwani, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 2006.

Boudiaf, M., Mueller, R., Ben Ayed, I., and Bertinetto, L. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Brito, M., Chávez, E., Quiroz, A., and Yukich, J. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. *Statistics & Probability Letters*, 35(1):33–42, 1997. ISSN 0167-7152. doi: https://doi.org/10.1016/S0167-7152(96)00213-1. URL https://www.sciencedirect.com/science/article/pii/S0167715296002131.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, 2020.

Chronister, L. M., Rhinehart, T. A., Place, A., and Kitzes, J. An annotated set of audio recordings of eastern north american birds containing frequency, time, and species information, 2021.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *Proceedings of the International Conference on Machine Learning*, 2021.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. In *NeurIPS Track on Datasets and Benchmarks*, 2021.

Desai, M. and Rao, V. A characterization of the smallest eigenvalue of a graph. *Journal of Graph Theory*, 18(2): 181–194, 1994.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Goëau, H., Kahl, S., Glotin, H., Planqué, R., Vellinga, W.-P., and Joly, A. Overview of BirdCLEF 2018: monospecies vs. soundscape bird identification. In *Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum*, 2018.

Grandvalet, Y. and Bengio, Y. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, 2004.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 2020.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *Proceedings of the International Conference on Learning Representations*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the International Conference on Learning Representations*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the International Conference on Computer Vision*, 2021.

Hou, Y. and Zheng, L. Source free domain adaptation with image translation. *arXiv preprint arXiv:2008.07514*, 2020.

Ilse, M., Tomczak, J. M., Louizos, C., and Welling, M. Diva: Domain invariant variational autoencoders. In *Proceedings of the Conference on Medical Imaging with Deep Learning,*, 2020.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*, 2015.

Kahl, S. *Identifying Birds by Sound: Large-scale Acoustic Event Recognition for Avian Activity Monitoring*. PhD thesis, Chemnitz University of Technology, 2019.

Kahl, S., Wood, C. M., Eibl, M., and Klinck, H. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

Kahl, S., Charif, R., and Klinck, H. A collection of fully-annotated soundscape recordings from the Northeastern United States, August 2022a. URL https://doi.org/10.5281/zenodo.7079380.

Kahl, S., Wood, C. M., Chaon, P., Peery, M. Z., and Klinck, H. A collection of fully-annotated soundscape recordings from the Western United States, September 2022b. URL https://doi.org/10.5281/zenodo.7050014.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., and Torralba, A. Undoing the damage of dataset bias. In *Proceedings of the European Conference on Computer Vision*, 2012.

Khurana, S., Moritz, N., Hori, T., and Le Roux, J. Unsupervised domain adaptation for speech recognition via uncertainty driven self-training. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6553–6557. IEEE, 2021.

Kim, Y., Cho, D., Han, K., Panda, P., and Hong, S. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, 2021.

Kundu, J. N., Venkat, N., Babu, R. V., et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Kurmi, V. K., Subramanian, V. K., and Namboodiri, V. P. Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013.

Li, R., Jiao, Q., Cao, W., Wong, H.-S., and Wu, S. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Li, Y., Wang, N., Shi, J., Hou, X., and Liu, J. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.

Liang, J., Hu, D., and Feng, J. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning*, 2020.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *Proceedings of the International Conference on Machine Learning*, 2015.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. *Advances in Neural Information Processing Systems*, 2018.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Morerio, P., Volpi, R., Ragonesi, R., and Murino, V. Generative pseudo-label refinement for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020.

Navine, A., Kahl, S., Tanimoto-Johnson, A., Klinck, H., and Hart, P. A collection of fully-annotated soundscape recordings from the Island of Hawai'i, September 2022. URL https://doi.org/10.5281/zenodo.7078499.

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Roy, S., Trapp, M., Pilzer, A., Kannala, J., Sebe, N., Ricci, E., and Solin, A. Uncertainty-guided source-free domain adaptation. In *Proceedings of the European Conference on Computer Vision*, 2022.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proceedings of the International Conference on Learning Representations*, 2020.

Sagawa, S., Koh, P. W., Lee, T., Gao, I., Xie, S. M., Shen, K., Kumar, A., Hu, W., Yasunaga, M., Marklund, H., et al. Extending the WILDS benchmark for unsupervised adaptation. In *Proceedings of the International Conference on Learning Representations*, 2022.

Sahoo, R., Shanmugam, D., and Guttag, J. Unsupervised domain adaptation in the absence of source data. In *ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

Sankaranarayanan, S., Balaji, Y., Castillo, C. D., and Chellappa, R. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Shen, K., Jones, R. M., Kumar, A., Xie, S. M., and Liang, P. How does contrastive pre-training connect disparate domains?, 2022. URL https://openreview.net/forum?id=vBn2OXZuQCF.

Song, H., Dong, L., Zhang, W.-N., Liu, T., and Wei, F. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. *arXiv preprint arXiv:2203.07190*, 2022.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., and Kelling, S. ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, 142(10):2282–2292, 2009.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., et al. The ebird enterprise: An integrated approach to development and application of citizen science. *Biological conservation*, 169:31–40, 2014.

Sun, B. and Saenko, K. Deep CORAL: Correlation alignment for deep domain adaptation. In *Proceedings of the European conference on computer vision*, 2016.

Sun, B., Feng, J., and Saenko, K. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., and Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the International Conference on Machine Learning*, 2020.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

Vellinga, W.-P. and Planqué, R. The xeno-canto collection and its relation to sound recognition and classification. *CLEF (Working Notes)*, 2015.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., and Savarese, S. Generalizing to unseen domains via adversarial data augmentation. *Advances in Neural Information Processing Systems*, 2018.

Vu, T.-H., Jain, H., Bucher, M., Cord, M., and Pérez, P. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526, 2019.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the International Conference on Learning Representations*, 2021.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 2019.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Yu, P. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022a.

Wang, Q., Fink, O., Van Gool, L., and Dai, D. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves ImageNet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Yang, S., Wang, Y., van de Weijer, J., Herranz, L., and Jui, S. Casting a BAIT for offline and online source-free domain adaptation. *arXiv preprint arXiv:2010.12427*, 2020.

Yang, S., van de Weijer, J., Herranz, L., Jui, S., et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 34:29393–29405, 2021.

Yuille, A. L. and Rangarajan, A. The concave-convex procedure. *Neural Computation*, 2003.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 2016.

Zeghidour, N., Teboul, O., Quitry, F. d. C., and Tagliasacchi, M. LEAF: A learnable frontend for audio classification. In *Proceedings of the International Conference on Learning Representations*, 2021.

Zhang, H., Zhang, Y., Jia, K., and Zhang, L. Unsupervised domain adaptation of black-box source models. In *Proceedings of the British Machine Vision Conference*, 2021.

Zhou, K., Yang, Y., Hospedales, T., and Xiang, T. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Ziko, I., Granger, E., and Ben Ayed, I. Scalable laplacian k-modes. *Advances in Neural Information Processing Systems*, 2018.

Table 6. **Relationship of problem settings**. x and y denote inputs and labels, and s and t "source" and "target", respectively (note that in some cases, as in DG, s might be a union of source domains / environments). For TTA and SFDA, the $*$ in their training data and loss reflects that they are entirely agnostic to how source training is performed, allowing the use of any off-the-shelf model.

	DA	DG	TTT	TTA	SFDA
Data used for training	x^s, y^s, x^t	x^s, y^s	x^s, y^s	$*$	$*$
Data used for adaptation	x^s, y^s, x^t	x^s, y^s, x^t	x^t	x^t	x^t
Training loss	$\mathcal{L}(x^s, y^s) + \mathcal{L}(x^s, x^t)$	$\mathcal{L}(x^s, y^s)$	$\mathcal{L}(x^s, y^s) + \mathcal{L}(x^t)$	$*$	$*$
Adaptation loss	—	—	$\mathcal{L}(x^t)$	$\mathcal{L}(x^t)$	$\mathcal{L}(x^t)$

A. Bioacoustics Datasets

We use **Xeno-Canto** (XC; Vellinga & Planqué, 2015) as the source dataset for bird species classification in the audio domain. XC is a growing, user-contributed collection of Creative Commons recordings of wild birds across the world. Our snapshot, downloaded on July 18, 2022, contains around 675,000 files spanning 10,932 bird species. Recordings are *focal* (purposefully capturing an individual's vocalizations in natural conditions, as opposed to *passively* capturing all ambient sounds), and each is annotated with a single foreground label (for the recording's main subject) and optionally a varying number of background labels (for other species vocalizing in the background).

For our distributionally-shifted datasets, we use multiple collections of passive (also called *soundscape*) recordings from various geographical locations. We use a 75/25 % split to obtain $\mathbb{D}_t^{adapt}$—used to adapt the model—and $\mathbb{D}_t^{test}$—used to evaluate the adapted model.

- **Sapsucker Woods** (SSW; Kahl et al., 2022a) contains soundscape recordings from the Sapsucker Woods bird sanctuary in Ithaca, NY, USA.
- **Sierra Nevada** (Kahl et al., 2022b) contains soundscape recordings from the Sierra Nevada in California, USA.
- **Hawai'i** (Navine et al., 2022) contains soundscape recordings from Hawai'i, USA. Some species, particularly endangered honeycreepers, are endemic to Hawai'i and many are under-represented in the Xeno-Canto training set.
- **Powdermill** ((Chronister et al., 2021)) contains high-activity dawn chorus recordings captured over four days in Pennsylvania, USA.
- **Caples** is an unreleased dataset collected by the California Academy of Science at the Caples Creek area in the central Californian Sierra Nevadas. Work is underway to open-source this dataset.
- **Colombia** is an unreleased dataset, previously used as part of the test set for the BirdCLEF 2019 competition.
- **High Sierras** is an unreleased dataset, previously used as part of the test set for the Kaggle Cornell Birdcall Identification challenge. Recordings are typically sparse, but with very low SNR due to wind noise. Work is underway to open-source this dataset.

B. Xeno-Canto data processing

Xeno-Canto recordings range from less than 1 second to several hours long. To extract 6-second segments we use a heuristic to identify segments with strong signal.

1. If the audio is shorter than 6 seconds, pad the recording evenly left and right using wrap-around padding.
2. Convert the audio into a log mel-spectrogram.
3. Denoise the spectrogram:

Table 7. Summary of Bioacoustic Soundscape Dataset Characteristics.
'XC/Species' indicates the average number of Xeno-Canto training example files per species. 'Low Data Species' are species with fewer than 50 training examples available. Labels per example is computed on the peak-sliced data, while hours and number of labels refer to the original raw dataset.

	Hours	#Species	#Labels	Labels/ Example	Climate	XC/ Species	Low data
Sapsucker	285	96	50,760	1.5	Temperate	367.9	3%
Sierra Nevada	33	56	20,147	1.9	Temperate	416.0	0%
Hawai'i	51	27	59,583	1.5	Tropical	166.3	44%
Powdermill	6	48	16,052	3.2	Temperate	360.0	0%
Caples	6	78	4,993	1.4	Temperate	334.8	10%
Colombia	8	63	1,489	1.4	Tropical	215.8	6%
High Sierras	34	19	14,494	1.2	Alpine	323.5	5%

(a) For each channel calculate the mean and standard deviation. All values that lie more than 1.5 standard deviations from the mean are considered outliers.
(b) Calculate a robust mean and standard deviation using the inliers[3].
(c) Any values that lie more than 0.75 robust standard deviations away from the robust mean are considered signal. Shift the signal in each channel by its robust mean, and set all noise to zero.

4. Sum all channels in the denoised spectrogram to create a signal vector.
5. Use SciPy's `find_peaks_cwt` function to retrieve peaks, using 10 Ricker wavelets with widths linearly spaced between 0.5 and 2 seconds
6. Select windows of 0.6 seconds centred at each peak and discard the peak if the maximum value in this window is smaller than 1.5 times the mean of the signal vector.
7. Keep only up to 5 peaks, with the highest corresponding values in the signal vector.
8. Select a 6 second window centred at each peak. If the window overlaps the start or beginning of the boundary, shift the window accordingly.

C. Soundscapes data processing

We extract 5-second segments from soundscapes recording by cross-referencing the bounding box labels with the same heuristic used to extract 6-second segments from XC recordings:

1. Use the procedure outlined in Appendix B to extract 5-second (rather than 6-second) windows from up to 200 (rather than 5) high-signal peaks per source file.
2. For all 5-second windows:
   (a) If it does not overlap in time with any bounding box label, drop it.
   (b) Otherwise, find all overlapping bounding box labels and label the window with the union of all their labels.

D. Metrics

D.1. mAP

Define $\mathrm{Prec}_X(s, c)$ as the generalized inverse rank of a ground-truth positive label c in an observation s, computed as:

$$\mathrm{Prec}_X(s, c) = \frac{1}{\mathrm{Rank}_X(s, c)} \sum_{r=1}^{\mathrm{Rank}_X(s,c)} \mathbf{1}[\mathrm{Label}(s, c) \in \mathcal{C}(s)]$$

[3]In the calculation of the mean and variance we add 1 to the denominator to avoid division by zero in the case that all values are considered outliers.

Table 8. Grid used for tuning hyper-parameters.

Method	Hyperparameter	Grid
All	Learning rate	{ 1e-5, 1e-4, 1e-3 }
	Trainable parameters	{ BatchNorm scale and bias, all }
	Use of dropout	{True, False }
	Use source BN statistics	{True, False }
	Learning rate cosine decay	{True, False }
SHOT (Liang et al., 2020)	Pseudo-labels weight β	{0., 0.3, 0.6, 0.9}
Pseudo-labelling (Lee et al., 2013)	Confidence threshold	{0., 0.5, 0.9, 0.95}
Dropout Student	Softness weight α	{0.1, 1.0}
	Pseudo-label update frequency	{Every iteration, Every epoch}
DUST	Number of random passes	{2, 3, 4}
	KL threshold	{0.8, 0.9, 0.99}
NRC	k nearest neighbors	{5, 10, 15}
	k extended nearest neighbors	{5, 10, 15}
	Base affinity	{0.1, 0.2}
NOTELA	k nearest neighbors	{5, 10, 15}
	Softness weight α	{0.1, 1.0}
	Pseudo-label update frequency	{Every iteration, Every epoch}

where X is the corpus over which we perform rankings, $\mathrm{Rank}_X\, s, c$ is the rank of the score for class c in observation s in the corpus X, $\mathrm{Label}(s, r)$ is the ground-truth label of class r in observation s, and $\mathcal{C}(s)$ is the set of ground-truth positive classes in observation s. Finally, $\mathbf{1}[\mathrm{Label}(s, r) \in \mathcal{C}(s)]$ is the indicator function for whether observation s is a ground-truth member of class c.

The mAP metric measures the per-example precision, averaged over the set $\mathcal{E}$ of all examples in the dataset:

$$\frac{1}{|\mathcal{E}|} \sum_{s \in \mathcal{E}} \frac{1}{|\mathcal{C}(s)|} \sum_{c \in \mathcal{C}(s)} \mathrm{Prec}_{\mathcal{E}}(s, c)$$

D.2. cmAP

Class-wise mean average precision (cmAP) is defined as:

$$\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|c \in \mathcal{E}|} \sum_{s \in \mathcal{E}} \mathrm{Prec}_{\mathcal{C}}(s, c)$$

Here $|c \in \mathcal{E}|$ denotes the total number of ground-truth positive examples of class c in the dataset. Notice that in this case, the precision ranking is over the logits for each target class, instead of ranking the logits in a single observation.

E. Hyperparameter validation

As mentioned in section 5, we reproduce all methods and ensure fairness of comparisons by (i) using the same pre-trained models and (ii) re-tuning each method's hyperparameters. All experiments are carried out with a batch size set to 64 (both audio and vision). Table 8 displays the grids used for hyper-parameter tuning (also both for audio and vision tasks). Note that to reduce the load of hyperparameters to tune in NOTELA, we make the design choice of using $\lambda = \alpha$, which effectively removes one degree of freedom.

F. Proof of Equation 3

We hereby provide the proof, as well as a more formal justification for the updates given in Equation 3. We start by restating the objective we want to minize:

$$\begin{aligned} \min_{\mathbf{y}_{1:N}} \quad & \mathrm{Tr}\left(-\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^\top \log(\mathbf{p}_i) + \frac{\alpha}{N}\sum_{i=1}^{N}\mathbf{y}_i^\top \log(\mathbf{y}_i) - \frac{\lambda}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\, \mathbf{y}_i^\top \mathbf{y}_j\right) \\ \text{s.t} \quad & \mathbf{1}^\top \mathbf{y}_i = 1,\ \mathbf{y}_i \geq 0\,. \end{aligned} \tag{4}$$

General case. Let us consider an easier problem, in which the Laplacian term has been linearized around the current solution $\mathbf{y}_i = \mathbf{p}_i$,

$$\begin{aligned} \min_{\mathbf{y}_{1:N}} \quad & \mathrm{Tr}\left(-\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^\top \log(\mathbf{p}_i) + \frac{\alpha}{N}\sum_{i=1}^{N}\mathbf{y}_i^\top \log(\mathbf{y}_i) - \frac{\lambda}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\, \mathbf{y}_i^\top \mathbf{p}_j\right) \\ \text{s.t} \quad & \mathbf{1}^\top \mathbf{y}_i = 1\,. \end{aligned} \tag{5}$$

We purposefully omitted the $\mathbf{y}_i \geq 0$ constraint, which will be satisfied later on by our solution to Equation 4. The Lagrangian of this problem is

$$\mathcal{L}(\mathbf{y}_{1:N}) = \mathrm{Tr}\left(-\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i^\top \log(\mathbf{p}_i) + \frac{\alpha}{N}\sum_{i=1}^{N}\mathbf{y}_i^\top \log(\mathbf{y}_i) - \frac{\lambda}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\, \mathbf{y}_i^\top \mathbf{p}_j\right) \tag{6}$$

$$+ \frac{1}{N}\sum_{i=1}^{N}\gamma_i(\mathbf{1}^\top \mathbf{y_i} - \mathbf{1})\,, \tag{7}$$

and the gradient of this Lagrangian with respect to $\mathbf{y}_i$ is

$$N.\nabla_{\mathbf{y}_i}\mathcal{L} = -\log(\mathbf{p}_i) + \alpha(\log(\mathbf{y}_i) + \mathbf{1}) - \lambda\sum_{j=1}^{N} w_{ij}\mathbf{p}_j + \gamma_i\mathbf{1}\,. \tag{8}$$

Solving for $\mathbf{y}_i$ yields

$$\mathbf{y}_i = \exp\left(-\frac{\alpha+\gamma_i}{\alpha}\right)\exp\left(\frac{\lambda}{\alpha}\sum_{j=1}^{N} w_{ij}\mathbf{p}_j\right)\odot \mathbf{p}_i^{1/\alpha}\,. \tag{9}$$

Now γ_i is chosen such that the constraint $\mathbf{y}_i^\top\mathbf{1} = 1$ is satisfied, resulting in

$$\mathbf{y}_i \propto \mathbf{p}_i^{1/\alpha}\odot \exp\left(\frac{\lambda}{\alpha}\sum_{j=1}^{N} w_{ij}\mathbf{p}_j\right)\,. \tag{10}$$

Concavity. Let $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{N\times N}$ be the matrix of affinity weights. An additional assumption on $\mathbf{W}$ allows a more formal justification of the linearization of the Laplacian term. Specifically, we can justify that if $\mathbf{W} + \mathbf{W}^\top$ is positive semi-definite, then the last term in Equation 4 is concave.

To show this rewrite $\mathrm{Tr}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\,\mathbf{y}_i^\top\mathbf{y}_j\right)$ as $\mathbf{1}^\top(\boldsymbol{W}\odot\boldsymbol{Y}^\top\boldsymbol{Y})\mathbf{1}$ where $\boldsymbol{Y} = (\mathrm{vect}(\mathbf{y}_i))$, which is in $\mathbb{R}^{N\times C}$ or $\mathbb{R}^{N\times 2C}$ for the single- and multi-label case respectively. The Hessian of this function is $\boldsymbol{W}\otimes\mathbb{I}+\boldsymbol{W}^\top\otimes\mathbb{I}$, whose eigenvalues are multiplicities of those of $\boldsymbol{W}+\boldsymbol{W}^\top$.

Hence equation 4 can be solved by a concave-convex procedure (CCP; Yuille & Rangarajan, 2003; Ziko et al., 2018; Boudiaf et al., 2022), which is suited to cases in which one part of the objective is convex (the two first terms in our case) and the

Table 9. Top-1 accuracy, averaged over 5 random seeds, along with 95% confidence interval for each corruption in CIFAR-10-C.

method corruption	Baseline	AdaBN	DUST	Dropout Student	NOTELA	NRC	Pseudo-Labelling	SHOT	TENT
brightness	91.28 ± 0.0	92.2 ± 0.03	92.2 ± 0.0	92.43 ± 0.16	92.53 ± 0.1	92.35 ± 0.05	92.49 ± 0.24	**92.68** ± 0.16	92.5 ± 0.09
contrast	52.84 ± 0.0	87.3 ± 0.02	87.02 ± 0.13	88.89 ± 0.36	89.75 ± 0.23	87.55 ± 0.11	**89.94** ± 1.63	89.21 ± 0.12	89.08 ± 0.32
defocus	53.16 ± 0.0	87.4 ± 0.09	87.35 ± 0.07	88.42 ± 0.24	88.47 ± 0.22	87.61 ± 0.09	**89.01** ± 0.43	88.45 ± 0.13	88.68 ± 0.07
elastic	73.28 ± 0.0	77.02 ± 0.1	77.09 ± 0.05	78.6 ± 0.39	78.4 ± 0.22	77.55 ± 0.07	**80.2** ± 0.39	79.07 ± 0.1	79.02 ± 0.07
fog	74.08 ± 0.0	85.92 ± 0.04	85.91 ± 0.04	87.3 ± 0.26	86.94 ± 0.38	85.99 ± 0.05	**88.49** ± 0.33	87.22 ± 0.16	87.46 ± 0.13
frost	59.6 ± 0.0	82.45 ± 0.05	82.46 ± 0.07	84.47 ± 0.37	84.55 ± 0.3	83.26 ± 0.04	**85.34** ± 0.7	84.71 ± 0.15	84.86 ± 0.06
frosted	46.36 ± 0.0	65.38 ± 0.03	65.26 ± 0.1	69.38 ± 0.43	69.05 ± 0.11	66.07 ± 0.1	**71.7** ± 0.5	70.58 ± 0.21	70.08 ± 0.15
gaussian	28.16 ± 0.0	73.31 ± 0.03	73.26 ± 0.07	75.75 ± 0.36	75.44 ± 0.18	73.93 ± 0.08	**78.62** ± 0.53	76.62 ± 0.08	76.58 ± 0.13
impulse	28.04 ± 0.0	64.36 ± 0.05	64.34 ± 0.06	67.74 ± 0.5	67.82 ± 0.32	65.45 ± 0.08	**70.86** ± 1.05	68.86 ± 0.11	68.96 ± 0.17
jpeg	69.12 ± 0.0	72.57 ± 0.06	72.49 ± 0.06	75.57 ± 0.2	75.32 ± 0.14	73.14 ± 0.09	**79.71** ± 0.27	76.01 ± 0.17	76.24 ± 0.21
motion	65.16 ± 0.0	86.29 ± 0.05	86.33 ± 0.07	87.72 ± 0.09	87.96 ± 0.12	86.51 ± 0.05	**88.12** ± 0.3	87.98 ± 0.12	87.9 ± 0.1
pixelate	41.84 ± 0.0	80.7 ± 0.04	80.66 ± 0.03	83.5 ± 0.2	83.66 ± 0.29	81.3 ± 0.08	**85.71** ± 0.66	83.66 ± 0.15	84.1 ± 0.11
shot	34.52 ± 0.0	75.68 ± 0.07	75.7 ± 0.05	78.64 ± 0.21	78.58 ± 0.35	76.36 ± 0.04	**81.42** ± 0.4	79.54 ± 0.15	79.88 ± 0.24
snow	75.12 ± 0.0	82.88 ± 0.03	82.91 ± 0.05	84.98 ± 0.2	85.17 ± 0.17	83.18 ± 0.06	**85.92** ± 0.36	85.22 ± 0.15	85.35 ± 0.15
zoom	57.36 ± 0.0	88.02 ± 0.05	87.99 ± 0.04	89.25 ± 0.43	89.5 ± 0.2	88.41 ± 0.09	**90.1** ± 0.37	89.62 ± 0.16	89.64 ± 0.14

other is concave (the Laplacian term). CCP proceeds by minimizing a sequence of *pseudo-bounds*, i.e., an upper bound that is tight at the current solution, obtained by linearizing the concave part of the objective at the current solution. Therefore, our proposed updates from Equation 3 can be interpreted as the first iteration of this procedure. Starting from the initial solution, $\mathbf{y}_i^{(0)} = \mathbf{p}_i$, unrolling the CCP procedure would consist of performing the following updates until convergence:

$$\mathbf{y}_i^{(t)} \propto \mathbf{p}_i^{1/\alpha} \odot \exp\left(\frac{\lambda}{\alpha} \sum_{j=1}^{N} w_{ij} \mathbf{y}_j^{(t-1)}\right) . \tag{11}$$

Affinity weights. Let $\boldsymbol{A}$ be the adjacency matrix of the mutual k-nearest neighbours graph, and $\boldsymbol{D}$ the diagonal matrix with the node degrees (i.e., $(\boldsymbol{D})_{ii}$ is the number of mututal neighbours for sample i). We know that $\boldsymbol{A} + \boldsymbol{D}$ is positive semi-definite (Desai & Rao, 1994) and hence matrix $\boldsymbol{W} = \boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{A} + \boldsymbol{D})\boldsymbol{D}^{-\frac{1}{2}}$ is also positive semi-definite. Note that this is equivalent to the case where w_{ij} is set to the reciprocal of the number of mutual neighbours, and $w_{ii} = 1$.

The terms w_{ii} act as an L2 regularizer on the values $\mathbf{y}_i$. Note that in our experiments we deviate from the theory and set $w_{ii} = 0$ to have better control over the regularization of $\mathbf{y}_i$.

G. Additional results

CIFAR-10-C per-corruption results. We provide the per-corruption results on CIFAR-10-C in Table 9, as well as the 95% confidence intervals.