SPECIALIZED EMBEDDING APPROXIMATION FOR EDGE INTELLIGENCE: A CASE STUDY IN URBAN SOUND CLASSIFICATION

Sangeeta Srivastava[⋆†] *Dhrubojyoti Roy*[⋆†] *Mark Cartwright*[§] *Juan P. Bello*[§] *Anish Arora*[⋆]

[⋆]Computer Science and Engineering, The Ohio State University, USA
[§]Music and Audio Research Laboratory, New York University, USA

{srivastava.206, roy.174}@osu.edu, {mark.cartwright, jpbello}@nyu.edu, arora.9@osu.edu

ABSTRACT

Embedding models that encode semantic information into low-dimensional vector representations are useful in various machine learning tasks with limited training data. However, these models are typically too large to support inference in small edge devices, which motivates training of smaller yet comparably predictive student embedding models through knowledge distillation (KD). While knowledge distillation traditionally uses the teacher's original training dataset to train the student, we hypothesize that using a dataset similar to the student's target domain allows for better compression and training efficiency for the said domain, at the cost of reduced generality across other (non-pertinent) domains. Hence, we introduce *Specialized Embedding Approximation* (SEA) to train a student featurizer to approximate the teacher's embedding manifold for a given target domain. We demonstrate the feasibility of SEA in the context of acoustic event classification for urban noise monitoring and show that leveraging a dataset related to this target domain not only improves the baseline performance of the original embedding model but also yields competitive students with >1 order of magnitude lesser storage and activation memory. We further investigate the impact of using random and informed sampling techniques for dimensionality reduction in SEA.

Index Terms— On-device machine learning, acoustic event detection, deep audio embeddings, knowledge distillation, urban noise classification.

1. INTRODUCTION

Embeddings serve as general purpose abstractions of data learned by representing semantically similar inputs close together in the embedding space. Besides their re-usability across different tasks, these representations help scale up the storage of data and speed up embedding-based retrieval [1]. Unfortunately, to learn such generic representations, a neural net requires a large amount of training data, which tends to be expensive to obtain with task-specific labels. Self-supervised learning [2, 3], one of the methods of representation learning, addresses this shortcoming by leveraging implicit labels in an unlabeled dataset to train neural network architectures on a pretext (or *upstream*) task. One such self-supervised embedding model, Look, Listen and Learn (L^3-Net) [4, 5], trains an Audio-Visual Correspondence (AVC) pretext task to learn rich visual and audio representations that can be successfully transferred to a variety of *downstream* tasks. However, the model has a significant storage and runtime memory requirement (18 MB flash and 12 MB of RAM) which prevents its adoption in resource constrained edge devices.[1]

Knowledge distillation (KD) can be used to learn a smaller *student* architecture by training it to mimic a larger *teacher* embedding network's output distribution [9], intermediate features [10], or inter-data relations [11]. However, the traditional KD setup assumes the availability of the upstream dataset and pretext task that the teacher was trained with. Although it is easy to have access to the pre-trained teacher model, growing concerns related to user privacy and proprietary information make it uncommon for developers to publicly release the training data [12, 13]. To address this problem, several works in Computer Vision make use of metadata [14] or generative adversarial networks (GANs) [15] to synthesize data (images) mimicking the data distribution of the teacher. But synthesizing high-fidelity, diverse and natural acoustic datasets using GANs is challenging due to the well-known problems of catastrophic forgetting and mode collapse.

Fortunately, it is often easy to obtain unlabeled data that is relevant for the downstream task at hand. For example, a deployed sensor network of Internet of Things (IoT) devices can inexpensively collect massive amounts of raw audio [16, 17] pertinent to tasks such as the classification of urban sounds. In this paper, we investigate the feasibility of using this new data to train *domain specialized* student embedding models without the need to define a pretext task for distillation.

Contributions of the paper. We introduce *Specialized Embedding Approximation* (SEA), a teacher-student learning paradigm where the student aims at learning the teacher's embedding space for a *target domain* in the absence of the original training data. In a domain specialized setting, the student cares only about preserving a portion of the teacher's embedding space instead of the entire space which would have been the case if the original dataset was used for distilling knowledge in a generic setting. This helps achieve the joint benefits of superior compression and significantly improved training efficiency.

We illustrate SEA in the context of an urban sound classification case study with relevant upstream data consisting of hundreds of millions of recordings collected by an array of deployed sensors in New York City. However, for such a large upstream data, dimensionality reduction of teacher's embeddings, a typical pre-processing

† Authors contributed equally to this work

This work is supported by the NSF Frontiers 1544753 award.

[1] As shown in [6], the storage cost can be reduced by more than an order of magnitude through sparsification without affecting model performance; however, the resultant models still have similar run-time memory requirements as L^3-Net without specialized hardware or software support [7, 8].

step in SEA, can be computationally expensive. We assess the utility of random and informed sampling techniques to select a suitable training subset from the aforementioned upstream data. Finally, we investigate the extent of local versus global structure preservation [18, 19, 20] necessary in approximating the teacher's manifold for our case study.

Our significant findings can be summarized as follows:

1. We show that SEA with domain specialization alone can improve the teacher's baseline performance, which in our case study is achieved with $\sim$15$\times$ smaller activation memory.
2. We show that SEA can produce competitive students with up to 11.75$\times$ smaller storage, 31$\times$ lesser activation memory, and $5-10\times$ improved training efficiency over the baseline.
3. We show that relevance and diversity-aware sampling techniques for selecting training points for the dimensionality reduction step of SEA can improve downstream performance.

We have open-sourced the SEA training pipeline[2] (Fig. 1) and the evaluated student models[3].

2. SPECIALIZED EMBEDDING APPROXIMATION

Given a teacher neural network $f_{\theta_T}(.) \in \mathbb{R}^n$ with model parameters θ_T, trained with a data set D_T, *Specialized Embedding Approximation (SEA)* aims to learn model parameters θ_S for a student $f_{\theta_S}(.) \in \mathbb{R}^d$ trained with a data set D_S, where $D_S \neq D_T \vee d \leq n$. Formally, the $SEA(f_{\theta_T}, f_{\theta_S}, D_S, \phi)$ problem is to optimize:

$$\min_{\theta_S} \sum_{x_j \in D_S} ||f_{\theta_S}(x_j) - \phi(f_{\theta_T}(x_j))||_2^2 \quad (1)$$

where $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^d$ is a learned dimensionality reduction map from the teacher's embedding space to the student's embedding space, and is the identity function if $d = n$.

As shown in Fig. 1 and explained in Section 1, the pipeline for solving SEA involves two components: (i) Dimensionality Reduction and (ii) Knowledge Distillation. The first step corresponds to learning a dimensionality reduction function ϕ. In order to reduce the memory and compute complexity associated with learning ϕ without compromising the geometric interpretation in the $\mathbb{R}^d$ subspace, we use a sampling technique to choose a representative subset of data points from D_S. The second step is to transform the teacher's embeddings for D_S using ϕ, and then train the student to learn the resulting $\mathbb{R}^d$ in D_S using Mean Squared Error (MSE) loss.

3. URBAN SOUND CLASSIFICATION

Sounds of New York City (SONYC) [16] is a large-scale wireless sensor network of acoustic sensors deployed in Manhattan, Brooklyn, and Queens boroughs of New York, to facilitate monitoring and mitigation of urban noise complaints. Since its inception in 2016, the SONYC sensor network has collected >150 Million 10-second audio clips and corresponding sound pressure level (SPL) data.

The goal of our SONYC case study is to use SEA to develop a small student embedding model that can act as a featurizer for training a downstream classifier to detect and classify acoustic events corresponding to *noise sources* while operating in situ in edge devices

[2]https://github.com/ksangeeta2429/embedding-approx

[3]https://pypi.org/project/edgel3/

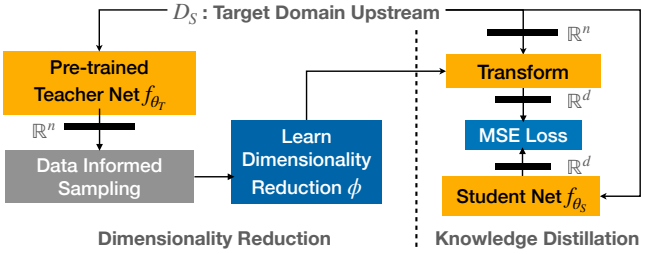

Fig. 1. Specialized Embedding Approximation (SEA) pipeline to train a student produce $\mathbb{R}^d$ embedding from a teacher with $\mathbb{R}^n$ output

we designed for SONYC. These IoT devices have an ARM Cortex-M7 MCU [21] with limited memory (1MB of RAM and 2MB of Flash) and computing to achieve long-lived self-powered operation. In addition to energy efficiency, it is desirable for the classifier to run on-device in SONYC for reasons of latency and privacy.

We use the data collected by the SONYC sensor network to derive an unlabeled dataset D_S for training student embedding model(s) with the afore-mentioned L^3 audio subnetwork as teacher[4]. Specifically, D_S consists of raw, unlabeled audio recordings collected by a subset of 15 sensors as our domain-specific upstream data, and is henceforth referred to as the *Upstream* SONYC dataset.

For training the downstream classifier from the student embeddings, we leverage SONYC Urban Sound Tagging (SONYC-UST) [22], a fraction of SONYC data annotated through crowdsourcing initiatives on the Zooniverse [23] platform, as the *Downstream* dataset. It is a multi-label dataset consisting of 3068 annotated 10-second audio recordings belonging to 8 classes: *engine*, *machinery-impact*, *non-machinery-impact*, *powered-saw*, *alert-signal*, *music*, *human-voice*, and *dog*. We use micro- and macro-averaged area under the precision-recall curve (AUPRC) as the evaluation metrics.

4. EXPERIMENTAL DESIGN

4.1. Student Models for Specialized Embedding Approximation

The L^3 audio subnetwork [5] comprises of 4 convolutional blocks (2 conv. layers in each) with 64, 128, 256, and 512 filters. It uses STFT-based mel spectrogram input representation calculated with a frame length of 2048 and a frame rate of $\sim$5 ms (equivalent to a hop size of 242 samples) over 48 kHz sampled audio. We evaluate 3 student models with at least 50% fewer filters in each layer and with 64, 128, and 256-dimensional outputs (Table 1), thereby reducing activation memory by more than 1.2 orders of magnitude (Students 1-3 in Table 1). Since computing the log mel spectrograms is a relatively heavy operation, we use a significantly coarser-grained input representation for all 3 students: sampling frequency of 8 KHz, 1024-point DFT, 64 mel filters, and 51 hops per STFT step. To isolate the impact of domain specialization, we also introduce a student model identical to L^3-Net albeit with the same low-resolution input representation (Student 0 in Table 1).

[4]We note that L^3 was selected owing to its superior performance in various acoustic event detection domains [4, 5]. SEA is agnostic to the teacher being used and L^3 can be replaced by any other embedding model.

Table 1. Input shape, number of filters, trainable parameters and memory requirements of the L^3 Audio teacher and student models

Model	Input Shape	#Filters in conv. blocks				#Train Params (Million)	Model Size (MB)	Act. Mem. (MB)
		1	2	3	4			
L^3 Audio	(256, 199, 1)	64	128	256	512	4.69	18.80	12.74
Student 0	(64, 50, 1)	64	128	256	512	4.69	18.80	0.82
Student 1	(64, 50, 1)	32	64	128	256	1.17	4.70	0.41
Student 2	(64, 50, 1)	32	64	128	128	0.58	2.34	0.41
Student 3	(64, 50, 1)	32	64	128	64	0.40	1.60	0.41

4.2. Dimensionality Reduction

Since SEA uses MSE loss for its knowledge distillation step, the teacher and student embedding dimensions need to be kept consistent. For the case study, we use Principal Component Analysis (PCA) [24] to map $\mathbb{R}^n$ to $\mathbb{R}^d$, where $n = 512$ and $d = 64, 128,$ and 256. In the following subsections, we respectively investigate the impact of sampling the upstream data for training the dimensionality reduction models, and whether preserving local or global structure is necessary in distilling the teacher's manifold.

4.2.1. Sampling Domain-Specific Upstream Data

The ability of the SONYC sensor network to continuously monitor results in the collection of a large amount of data where there are no acoustic events of interest, i.e. "background" and only a small fraction of the data contain "foreground" events corresponding to noise complaints. To ensure that a representative subset is selected for our study with low redundancy as well as good coverage with informative data points, we use different sampling methods. From each of the 15 sensors, we arrange the timestamped data into *daily* timeframes spanning an entire year, and use the following four strategies to create the training set for dimensionality reduction:

1. **Random Sampling.** It is the simplest technique that uniformly selects embeddings from each day over a year.
2. **Sound Pressure Level (SPL) Informed Sampling.** This sampling technique selects embeddings using their corresponding recorded SPL values. We use the following strategy: using a 2-hour sliding window, we first assess the *relative loudness* of each embedding from its associated SPL values by ranking the SPL values and normalizing these ranks to get a relevance score for each frame. We then convert these 2-hour relevance scores to a sampling probability distribution over a day[5].
3. **Determinantal Point Process (DPP) Sampling.** Using DPP [26] as a probabilistic model of diversity, we increase the probability of sets that are more spread out and diverse in the embeddings space. Two flavors of DPPs are considered:
   (a) *Diversity Only (dpp_div)*. It uses a kernel matrix that defines similarity between pairs of embeddings, to ensure that more similar embeddings are less likely to co-occur.
   (b) *Diversity and Quality (dpp)*. In addition to the kernel matrix for ensuring diversity in the embedding subset, we use the aforementioned audio frame relative loudness measurements as a quality signal to ensure inclusion of the most relevant events.

We use these techniques to sample 500K training points for PCA across the selected sensors, yielding ∼33K points per sensor.

[5]A more detailed overview can be found in the Appendix [25].

Table 2. SEA improves baseline as well as student performances on the target task

Model	Original (AVC)		Specialized (SEA)	
	Micro-AUPRC	Macro-AUPRC	Micro-AUPRC	Macro-AUPRC
L^3 Audio	0.810	0.567	-	-
Student 0	0.812	0.591	**0.823**	**0.595**
Student 1	0.789	0.562	0.793	0.552
Student 2	0.790	0.552	0.797	0.559
Student 3	0.784	0.543	0.784	0.559

4.2.2. Model Selection

Upon training a dimensionality reduction model using data sampled with the above methods, we first transform the SONYC-UST dataset to assess its downstream performance. The best downstream classifier[5] is selected from combinations of 0, 1, and 2 hidden layers with 128 and 256 dimensions and using Multi-Instance Learning aggregation [27] across frames in an audio clip. The best dimensionality reduction model thus obtained is then used to transform the upstream data prior to the KD step.

5. RESULTS

5.1. Specialized Embedding Approximation

5.1.1. Impact of Domain Specialization

Table 2 presents an overall evaluation of SEA, specifically highlighting the impact of domain specialization for the urban noise monitoring application by replacing the original training points and AVC task with an SEA task involving the more relevant upstream dataset. It can be seen that training a specialized student model through SEA not only removes the dependency on the original training dataset but, in fact, improves the baseline performance on SONYC-UST even with a coarser input representation when the CNN is identical to the teacher (cf. Student 0). Similar improvements are observed in Students 1-3; they are extremely competitive with the teacher's performance with $>$1.2 orders of magnitude lesser activation memory while generally outperforming their AVC-trained counterparts. Thus, the proposed technique produces state-of-the-art models for noise monitoring on edge devices and illustrates the value in collecting domain-specific upstream data for the target tasks at hand without relying on the availability of pre-existing datasets.

Performance on Out-of-Domain Tasks. Domain specialization techniques such as using SEA aim at training an embedding model optimized for a target domain and the tasks related to it. While this can certainly boost in-domain performance (cf. Table 2), the performance on out-of-domain datasets can be reduced. This is exemplified in Table 3 where the performance of SONYC-UST (in-domain dataset) is compared with two domain-shifted datasets: (i) **US8K** [28], where the domain shift results from different acoustic recording conditions and (ii) **ESC-50** [29], which differs significantly from SONYC-UST not only in terms of its recording conditions but also in terms of event characteristics and label space. Student 0, with the same architecture as the L^3 audio except for coarser input representation, while improving SONYC-UST's performance by 1.6%, incurs a performance loss of 6.8% and 16% in US8K and ESC-50 respectively. As the model shrinks from Student 0 to Student 3, the degradation is more substantial in the out-of-domain datasets as

Table 3. Specialized models sacrifice out-of-domain performances

Model	Accuracy (US8K)	Accuracy (ESC-50)	Micro-AUPRC (SONYC-UST)
L^3 Audio	0.759	0.737	0.810
Student 0	0.707	0.613	0.823
Student 1	0.655	0.512	0.793
Student 2	0.673	0.498	0.797
Student 3	0.648	0.440	0.784

Table 4. Effect of sampling techniques on dimensionality reduction

Model	Emb. Dim.	Samp. Type	Micro-AUPRC	Macro-AUPRC	
				Overall	5 Common Classes
PCA	64	dpp	0.781	0.555	0.674
		dpp_div	**0.783**	0.561	0.685
		random	0.782	**0.574**	**0.686**
		spl	0.779	0.560	0.675
	128	dpp	**0.796**	0.578	**0.706**
		dpp_div	0.796	0.575	0.700
		random	0.793	0.574	0.702
		spl	0.795	**0.582**	0.700
	256	dpp	**0.796**	0.582	**0.702**
		dpp_div	0.795	0.575	0.693
		random	0.795	**0.585**	0.699
		spl	0.795	0.578	0.697

compared to the in-domain SONYC-UST with ESC-50 being the worst because of its dissimilar task.

5.1.2. Training Efficiency

Finally, as compared to the L^3-Net AVC training, the student distillation in SEA is significantly more efficient in two aspects: (i) requires 10x lesser data as compared to the former, (ii) converges 5x (10x) faster with a learning rate of 10^{-5} (10^{-4}).

5.2. Sampling

Table 4 lists the impact of sampling techniques in selecting the data for training dimensionality reduction model, PCA. While the DPP techniques seem to improve micro-AUPRC on the SONYC-UST dataset, random and SPL-based techniques yield the highest macro-AUPRC. This apparent anomaly can be explained by studying the per-class AUPRCs obtained by the four sampling methods. Note that macro-AUPRC aggregates the per-class results *without* considering their overall distributions in the dataset. In particular, the underperformance of DPP on *music* and *dog* classes, which constitute only 9% and 5% of the test points respectively, biases the metric in favor of random/SPL sampling. However, diversity sampling is seen to generally outperform the other techniques on the 5 most commonly occuring classes in the dataset: *engine* (52%), *human-voice* (45%), *alert-signal* (22%), *machinery-impact* (21%), and *non-machinery-impact* (11%) (last column in Table 4). Thus, informed sampling techniques that balance relevance with diversity benefit downstream performances for the target task.

Counterintuitively, sampling for diversity does not seem to favor the underrepresented classes in the downstream dataset. This is partly attributable to the fact that the current coarse label space of the SONYC-UST is quite limited compared to the wide range of sounds in urban soundscapes. It would be interesting to assess the impact of diversity sampling as more sound sources are annotated over time; we relegate this exploration to future work.

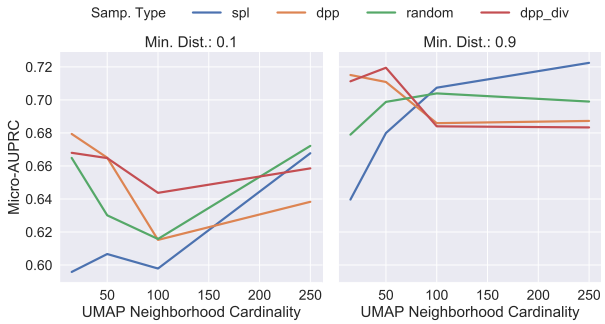

Fig. 2. Impact of structure preservation (emb. dim.=128)

5.3. Impact of Structure Preservation

Previous research [30, 31] has claimed that a single characterization, either global or local, is insufficient to represent the underlying structures of real-world data. In this section, we further explore this trade-off using Uniform Manifold Approximation and Projection (UMAP) [32], where the balance between local and global structure in the teacher's manifold can be tuned via two hyperparameters: (i) `n_neighbors`: the number of points in the local clusters created in $\mathbb{R}^d$, and (ii) `min_dist`: the minimum distance between points which determines how tightly the points are "clumped" together in their respective clusters. Note that an increase in either hyperparameter favors global over local structure.

Fig. 2 compares the micro-AUPRC of Student 2 for progressively increasing `n_neighbors` at the local and global extremeties of `min_dist`. It can be observed that, unlike with PCA, SPL-based sampling seems to outperform diversity sampling on micro-AUPRC. We rationalize this as follows: since diversity aims at reducing redundancy among high-information points, it yields a more complex manifold that requires smaller cluster sizes (i.e., more local clusters) to approximate and possibly overfits the downstream dataset. SPL, on the other hand, allows some redundancy of informative points which creates a simpler manifold and helps reduce this overfitting.

Further, it can be seen that a higher `min_dist` is universally better for all techniques. Through these observations, we postulate that preserving the global structure of the teacher's embedding manifold matters more than preserving local structure in our case study. This is corroborated by the observation that PCA-based dimensionality reduction, which only preserves gross global structure, dominates the downstream performances over UMAP.

6. CONCLUSION

We propose *Specialized Embedding Approximation* to distill knowledge from a teacher embedding model to smaller, domain-specialized student audio models. For our case study in urban sound classification, SEA produces competitive students that are substantially more training efficient and smaller in size, albeit at the cost of cross-domain performances. As future work, we wish to address this limitation by studying the feasibility of SEA in conjunction with meta-learning across downstream datasets, potentially yielding even smaller students that can easily specialize to new task contexts or unseen audio domains.

7. REFERENCES

[1] J. Huang, A. Sharma, et al., "Embedding-based retrieval in facebook search," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2553–2561.

[2] I. Misra, C. L. Zitnick, et al., "Shuffle and learn: unsupervised learning using temporal order verification," in *ECCV*. Springer, 2016, pp. 527–544.

[3] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *ICCV*, 2017, pp. 2051–2060.

[4] R. Arandjelović and A. Zisserman, "Look, listen and learn," in *IEEE ICCV*, 2017, pp. 609–617.

[5] J. Cramer, H. H. Wu, et al., "Look, Listen, and Learn more: Design choices for deep audio embeddings," in *IEEE ICASSP*, 2019.

[6] S. Kumari, D. Roy, et al., "EdgeL3: Compressing L3-Net for mote scale urban noise monitoring," in *IPDPSW*. IEEE, 2019, pp. 877–884.

[7] S. Han, X. Liu, et al., "EIE: Efficient inference engine on compressed deep neural network," *ACM SIGARCH Computer Architecture News*, vol. 44, Feb 2016.

[8] D. Kim, J. Ahn, et al., "ZeNA: Zero-aware neural network accelerator," *IEEE Design & Test*, vol. 35, no. 1, pp. 39–46, 2017.

[9] G. Hinton, O. Vinyals, et al., "Distilling the knowledge in a neural network," *preprint arXiv:1503.02531*, 2015.

[10] A. Romero, N. Ballas, et al., "FitNets: Hints for thin deep nets," *preprint arXiv:1412.6550*, 2014.

[11] W. Park, D. Kim, et al., "Relational knowledge distillation," in *Proceedings of the IEEE CVPR*, 2019, pp. 3967–3976.

[12] Y. Taigman, M. Yang, et al., "DeepFace: Closing the gap to human-level performance in face verification," in *CVPR*, 2014, pp. 1701–1708.

[13] Y. Wu, M. Schuster, et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," *preprint arXiv:1609.08144*, 2016.

[14] K. Bhardwaj, N. Suda, et al., "Dream distillation: A data-independent model compression framework," *arXiv preprint arXiv:1905.07072*, 2019.

[15] H. Chen, Y. Wang, et al., "Data-free learning of student networks," in *Proceedings of the IEEE ICCV*, 2019, pp. 3514–3522.

[16] J. P. Bello, C. Silva, et al., "SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution," *CACM*, vol. 62, no. 2, pp. 68–77, Jan. 2019.

[17] C. Catlett, P. Beckman, et al., "Array of Things: A scientific research instrument in the public way: platform design and early lessons learned," in *SCOPE*. ACM, 2017, pp. 26–33.

[18] M. Noroozi, A. Vinjimoor, et al., "Boosting self-supervised learning via knowledge transfer," in *CVPR*, 2018, pp. 9359–9367.

[19] J. Huang, Q. Dong, et al., "Unsupervised deep learning by neighbourhood discovery," *preprint arXiv:1904.11567*, 2019.

[20] Y. Liu, J. Cao, et al., "Knowledge distillation via instance relationship graph," in *CVPR*, 2019, pp. 7096–7104.

[21] "STM32H7," `https://www.st.com/en/microcontrollers/stm32h7-series.html`.

[22] M. Cartwright, A. Mendez, et al., "SONYC urban sound tagging (SONYC-UST): a multilabel dataset from an urban acoustic sensor network," *DCASE*, 2019.

[23] "Zooniverse," `https://www.zooniverse.org/`.

[24] I. T. Jolliffe, *Principal Component Analysis and Factor Analysis*, pp. 115–128, Springer New York, 1986.

[25] S. Srivastava, D. Roy, et al., "Appendix," `https://doi.org/10.5281/zenodo.4536076`.

[26] G. Gautier, G. Polito, et al., "DPPy: DPP sampling with python.," *Journal of Machine Learning Research*, vol. 20, no. 180, pp. 1–7, 2019.

[27] B. McFee, J. Salamon, et al., "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

[28] J. Salamon, C. Jacoby, et al., "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[29] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *ACM Int. Conf. on Multimedia*, 2015, pp. 1015–1018.

[30] X. Zhu, S. Zhang, et al., "Local and global structure preservation for robust unsupervised spectral feature selection," *TKDE*, vol. 30, no. 3, pp. 517–529, 2017.

[31] Y. Song, Y. Li, et al., "Preserving global and local structures for supervised dimensionality reduction," in *MATEC Web of Conferences*. EDP Sciences, 2015, vol. 28, p. 06003.

[32] L. McInnes, J. Healy, et al., "UMAP: Uniform manifold approximation and projection for dimension reduction," *preprint arXiv:1802.03426*, 2018.