Dataset Distillation for Audio Classification: A Data-Efficient Alternative to Active Learning

Gautham Krishna Gudur **Edison Thomaz**
University of Texas at Austin
{gauthamkrishna, ethomaz}@utexas.edu

Abstract

Audio classification tasks like keyword spotting and acoustic event detection often require large labeled datasets, which are computationally expensive and impractical for resource-constrained devices. While active learning techniques attempt to reduce labeling efforts by selecting the most informative samples, they struggle with scalability in real-world scenarios involving thousands of audio segments. In this paper, we introduce an approach that leverages dataset distillation as an alternative strategy to active learning to address the challenge of data efficiency in real-world audio classification tasks. Our approach synthesizes compact, high-fidelity coresets that encapsulate the most critical information from the original dataset, significantly reducing the labeling requirements while offering competitive performance. Through experiments on three benchmark datasets – Google Speech Commands, UrbanSound8K, and ESC-50, our approach achieves up to a $\sim$3,000x reduction in data points, and requires only a negligible fraction of the original training data while matching the performance of popular active learning baselines.

1 Introduction

Deep learning has advanced audio classification tasks like keyword spotting (KWS) [19] and acoustic event detection and is widely used in voice assistants for detecting keywords and identifying environmental sounds [12]. These tasks often require large labeled datasets and substantial computational resources. With the rise of resource-constrained devices, such as smartphones, there is a growing need for data-efficient methods that reduce computational and labeling costs while maintaining high performance.

Traditional and modern active learning (AL) techniques often attempt to mitigate such labeling burdens by selecting the most informative subset of training samples for user annotation [14, 2]. While these methods exhibit efficacy, their practicality diminishes in such real-world applications, where datasets typically comprise thousands of windowed audio segments.

To this end, we propose leveraging *dataset distillation* [17] to generate compact, high-fidelity data summaries as a complementary strategy to active learning. We illustrate this key difference in Figure 1. Dataset distillation synthesizes representative data points effectively encapsulating the most critical information from the original dataset. These distilled data points, representing the most informative samples, can then be sent to the oracle (user/expert) for annotation, substantially reducing the amount of labeled data required. Notably, the proposed approach necessitates only 0.01-0.15% of the total training data while achieving comparable performance to existing active learning methods, enhancing its practicality and deployability for audio classification tasks.

Our scientific contributions are as follows: **(1)** Introducing *dataset distillation as an effective alternative strategy to active learning* and defining the key differences between data subset selection and data subset generation via dataset distillation. **(2)** Proposing a dataset distillation approach that

4th NeurIPS Efficient Natural Language and Speech Processing Workshop (ENLSP-IV 2024).

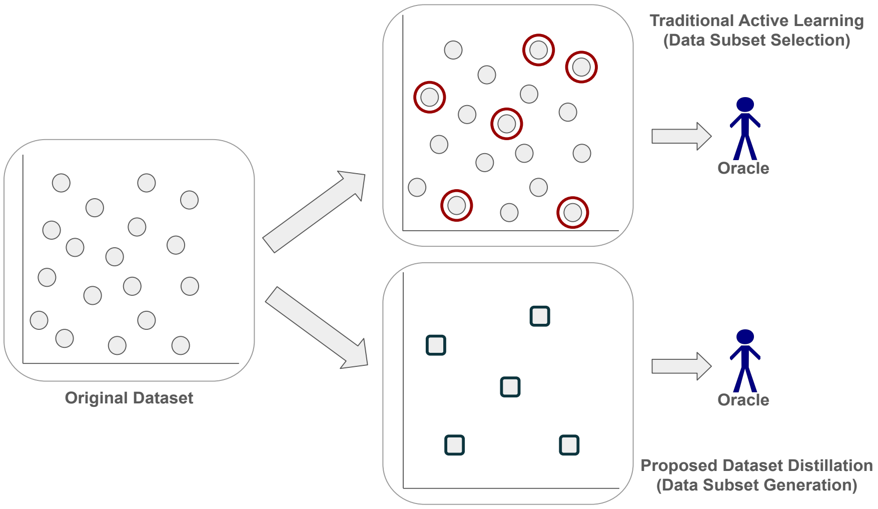

Figure 1: This figure illustrates the distinction between traditional active learning and our proposed dataset distillation approach, highlighting how dataset distillation can be leveraged as an alternative data-efficient strategy to active learning.

synthesizes compact, high-fidelity data summaries to significantly reduce labeled data requirements for audio classification tasks in resource-constrained environments, balancing data efficiency and model performance. **(3)** Conducting extensive evaluations across three diverse datasets – Google Speech Commands, UrbanSound8K, and ESC-50, demonstrating up to a ∼3,000x reduction in data points with minimal accuracy trade-offs, achieving competitive performance against traditional and modern active learning methods.

2 Related Work

Conventional active learning seeks to optimize labeling efficiency by selecting the most informative samples for annotation. Traditional methods such as uncertainty sampling and query-by-committee [14] have established the foundation for more sophisticated techniques in the deep learning era. Recently, strategies incorporating deep neural networks, such as Bayesian Active Learning by Disagreement [6] and core-set active learning [13] have demonstrated substantial improvements in performance. These methods prioritize samples that maximize learning efficiency, significantly reducing the annotation burden while offering competitive accuracy.

Dataset distillation is a technique designed to reduce dataset size while preserving their essential informational content [17]. With the growing need to manage large datasets efficiently, methods such as dataset condensation and core-set selection have gained prominence in recent years. Techniques like gradient matching [22], distribution matching [21], and differentiable Siamese augmentation (DSA) [20] demonstrate the effectiveness of synthesizing compact data representations that preserve the essential characteristics of the original datasets. Alternate approaches such as Kernel Inducing Points (KIP) [8] leverage kernel methods to enhance dataset distillation. The closest approach to our paper – *Efficient Dataset Distillation using Random Feature Approximation* (RFAD) [7] employs random feature approximation to achieve faster dataset distillation with reduced computational complexity, while maintaining the quality of distilled datasets. RFAD also strikes an optimal balance between data efficiency and performance, making it particularly suited for applications where computational speed is critical.

In the field of audio classification, the application of dataset distillation and active learning is particularly pertinent. Audio datasets, such as those used for keyword spotting and acoustic event detection, often consist of extensive, high-dimensional audio segments. Implementing dataset distillation in this context allows for the synthesis of concise, high-fidelity audio representations, significantly lowering the volume of data requiring annotation. While dataset distillation has been underexplored in audio classification, to the best of our knowledge, this is the first work to position it both as a substitute and an enhancement to active learning, with the goal of drastically reducing labeled data requirements and improving data efficiency. This makes it a highly viable solution for real-world, high-dimensional audio classification tasks. Leveraging dataset distillation as an efficient replacement for active learning offers a promising direction for developing scalable and efficient audio classification systems.

3 Our Approach

In this section, we discuss the problem formulation of utilizing dataset distillation – a data subset generation paradigm, as a highly data-efficient strategy in place of active learning. Figure 1 illustrates this distinction effectively, and the technical details of our proposed dataset distillation approach are outlined in detail in Algorithm 1.

In active learning, we typically select the most informative subset for training. In contrast, dataset distillation generates a compact, synthetic dataset that captures the knowledge of a much larger dataset. Leveraging dataset distillation as an alternative technique to active learning can be powerful for generating highly informative and efficient data subsets for training. To mathematically define these,

Data Subset Selection. This is the traditional approach in active learning where the goal is to *select* the most informative data subset S from the original dataset D for training. This can be described as, $S = \text{argmax}_{S \subseteq D, |S|=k} \mathcal{U}(S; \theta)$, where D is the entire dataset, S is the selected subset with size $|S| = k, \mathcal{U}(S; \theta)$ is a function to evaluate informativeness of S given model parameters θ.

Data Subset Generation via Dataset Distillation. In dataset distillation for active learning, the goal is to *synthesize* a distilled subset G that minimizes a loss function to represent the knowledge of the larger dataset D. This can be formulated as, $G = \text{argmin}_{G \in \mathcal{G}} \mathcal{L}(f(G; \theta), D)$, where G is the distilled subset, $\mathcal{G}$ is the space of all possible generated data subsets, $\mathcal{L}(f(G; \theta), D)$ how well the model $f(G; \theta)$ trained on G generalizes to the entire dataset D.

The core of our approach involves synthesizing a high-fidelity dataset that captures the essential features of the original data, providing a data-efficient method in place of active learning. One way to achieve this is inspired by advancements in dataset distillation and neural network Gaussian processes (NNGP), which leverage the principles of random feature approximation to generate a computationally efficient and informative distilled dataset using the RFAD method [7], as discussed in Section 2. Our proposed approach generates representative samples that encapsulate critical information required for audio classification, significantly reducing the annotation burden. The key steps of our approach are detailed below.

Initialization. Start with the original training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ and initialize a smaller coreset $\mathcal{C} = \{(x'_i, y'_i)\}_{i=1}^{M}$, where $M \ll N$.

Random Feature Generation. Generate random features for both the training set and coreset using multiple randomly initialized neural networks M. For a given input x, the random features are computed as,

$$\Phi(x) = \frac{1}{\sqrt{NM}} \begin{bmatrix} f_{\theta_1}(x) \\ \vdots \\ f_{\theta_N}(x) \end{bmatrix} \in \mathbb{R}^{NM}$$

where f_{θ_i} denotes the output of the i-th neural network, N denotes output dimension of M.

Kernel Matrix Computation. Construct the kernel matrices $\Phi(\mathcal{D})$ and $\Phi(\mathcal{C})$ for the training set and the coreset respectively. The kernel matrices are computed as,

$$K_{\mathcal{DC}} = \Phi(\mathcal{D})^\top \Phi(\mathcal{C})$$

$$K_{\mathcal{CC}} = \Phi(\mathcal{C})^\top \Phi(\mathcal{C})$$

Coreset Update. Using kernel ridge regression, update the coreset to minimize the loss function $\mathcal{L}$ defined as the difference between the true and predicted labels from the distilled dataset. The loss function and update rule are given by,

$$\mathcal{L} = \|y_{\mathcal{D}} - K_{\mathcal{DC}}(K_{\mathcal{CC}} + \lambda I)^{-1} y_{\mathcal{C}}\|^2$$

$$\mathcal{C} \leftarrow \mathcal{C} - \eta \nabla_{\mathcal{C}} \mathcal{L}$$

where η is the learning rate, λ is a regularization parameter.

Iteration. Repeat the random feature generation, kernel matrix computation, and coreset update steps until convergence or a predefined stopping criterion is met.

Algorithm 1 Our Proposed Approach

```
  Input. Training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, Initial coreset $\mathcal{C} = \{(x'_i, y'_i)\}_{i=1}^M$, Number of random networks $N$, Output dimension of random networks $M$, Regularization parameter $\lambda$, Learning rate $\eta$
  while not converged do
    Sample a batch $\mathcal{B} \subset \mathcal{D}$
    Initialize $N$ random neural networks $\{f_{\theta_i}\}_{i=1}^N$
    for each $x \in \mathcal{B}$ do
      Compute random features $\Phi(x)$
    end for
    for each $x' \in \mathcal{C}$ do
      Compute random features $\Phi(x')$
    end for
    Compute kernel matrices $K_{\mathcal{BC}}$ and $K_{\mathcal{CC}}$
    Calculate predicted labels for the batch: $\hat{y}_{\mathcal{B}} = K_{\mathcal{BC}}(K_{\mathcal{CC}} + \lambda I)^{-1} y_{\mathcal{C}}$
    Compute loss: $\mathcal{L} = \|y_{\mathcal{B}} - \hat{y}_{\mathcal{B}}\|^2$
    Update coreset using gradient descent: $\mathcal{C} \leftarrow \mathcal{C} - \eta \nabla_{\mathcal{C}} \mathcal{L}$
  end while
Ensure: Distilled coreset $\mathcal{C}$
```

Our method is particularly well-suited for high-dimensional audio classification tasks like keyword spotting and acoustic event detection. By distilling the original dataset into a compact and representative coreset, we significantly reduce labeling requirements, making active learning more scalable. In addition, by focusing on generating rather than selecting data, our approach offers greater flexibility and potentially better generalization by retaining the essential features of the original dataset ensuring strong performance while drastically minimizing computational and storage demands.

4 Experiments and Results

We evaluate our proposed dataset distillation approach using three popular audio classification datasets: Google Speech Commands (GSC) [18], UrbanSound8K (US8K) [12], and ESC-50 [10]. For GSC, we select 10 keywords: *Yes, No, Up, Down, Left, Right, On, Off, Stop, Go* as performed in [18, 19, 3] and extract Mel-frequency Cepstral Coefficients (MFCC) at 16 kHz, dividing the data into windows of 50 ms each. US8K comprises 10 environmental sound classes: *air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, street music*, with preprocessing similar to that of GSC, and is evaluated using five-fold cross-validation. ESC-50 includes 2,000 five-second audio clips across 50 environmental sound classes, ranging from chirping birds to car horns, sampled at 44.1 kHz and downsampled to 16 kHz. The dataset is evaluated using five-fold cross-validation, and similar preprocessing steps are applied. These datasets offer diverse and challenging keyword spotting and acoustic event detection tasks for effective evaluation of our proposed distillation approach.

We evaluate the performance of our dataset distillation approach using two models: ResNet-18 and a 4-layer convolutional neural network (CNN). ResNet-18 is widely used due to its effectiveness in extracting hierarchical features from complex data, particularly in image and audio classification tasks [5, 16]. The 4-layer CNN is designed to capture local patterns in audio signals and has been successfully applied in small-footprint keyword spotting models [11, 9]. Our implementation utilizes the PyTorch framework, with a learning rate of $2e^{-4}$ and an Adam optimizer. Both models are trained using 7 random networks (M) for random feature approximation and significantly distilled coresets, achieving competitive accuracy with drastically reduced data requirements.

4.1 Baselines for Active Learning

For a given classification model M, data D, and input x, we evaluate the following acquisition functions popularly used in literature [2, 4].

Table 1: Comparison of test classification accuracy vs % of training samples for baseline methods and our proposed approach across all three datasets using a ResNet-18 model.

	Google Speech Commands		**UrbanSound8K**		**ESC-50**	
Method	**% Samples**	**Accuracy**	**% Samples**	**Accuracy**	**% Samples**	**Accuracy**
Total Training Data	100%	89.92	100%	79.27	100%	69.36
Max Entropy	60%	72.2	60%	62.15	60%	50.2
	40%	69.6	40%	57.45	40%	45.12
	30%	65.1	30%	53.36	30%	41.25
	20%	57.25	20%	45.18	20%	36.75
	0.029%	9.15	0.063%	7.56	0.15%	5.12
Variation Ratios	**60%**	**73.36**	60%	63.02	**60%**	**51.25**
	40%	70.85	40%	58.75	40%	45.36
	30%	65.3	30%	54.12	30%	41.58
	20%	58.72	20%	46.78	20%	36.24
	0.029%	9.24	0.063%	7.15	0.15%	5.84
BALD	60%	73.15	**60%**	**63.12**	60%	50.95
	40%	70.5	40%	58.95	40%	44.78
	30%	65.1	30%	54.08	30%	40.75
	20%	58.55	20%	46.42	20%	35.16
	0.029%	9.38	0.063%	7.39	0.15%	5.5
Random	60%	73.15	60%	62.87	60%	50.67
	40%	70.25	40%	59.08	40%	44.95
	30%	65.36	30%	54.45	30%	39.25
	20%	58.48	20%	46.92	20%	33.18
	0.029%	9.27	0.063%	7.72	0.15%	4.95
Proposed Method	**0.029%**	**72.24**	**0.063%**	**61.67**	**0.15%**	**49.65**
	0.017%	61.13	0.038%	50.24	0.09%	31.25
	0.012%	51.68	0.025%	37.85	0.0625%	17.96

Max Entropy: Pool points are selected to maximize predictive entropy [15].

$$\mathbb{H}[y|x, D] := -\sum_{c} p(y = c|x, D) \log p(y = c|x, D)$$

Variation Ratios (VR): The LC (Least Confident) method for uncertainty-based pool sampling is performed here [1].

$$variation - ratio[x] := 1 - \max_{y} p(y|x, D)$$

Bayesian Active Learning by Disagreement (BALD): Pool points are selected to maximize mutual information between predictions and model posterior [6]. These points reflect the model's average uncertainty, where parameter disagreement on the outcome is the highest.

$$\mathbb{I}[y, \omega|x, D] = \mathbb{H}[y|x, D] - E_{p(\omega|D)}\big[\mathbb{H}[y|x, \omega]\big]$$

where $\mathbb{H}[y|x, \omega]$ is the entropy of y, given model weights ω.

Random Sampling: Data points are selected uniformly at random from the given pool.

4.2 Discussion on Results

We evaluate our proposed dataset distillation approach by showcasing the percentage of training samples used and corresponding accuracy on three audio classification datasets, using both ResNet-18 model (Table 1) and a 4-layer CNN (Table 2). The number of audio samples per class (AS/C) generated by our approach is also provided in Table 3.

Our proposed approach, using the ResNet-18 model, achieves a test classification accuracy of 72.24% on GSC with only 0.029% of the training data (50 samples per class), closely matching Variation Ratios (73.36%), which requires 60% of the data, resulting in a more than 2,000x reduction in labeled data. On US8K, our approach reaches 61.67% accuracy with only 0.063% of the data, compared

Table 2: Comparison of test classification accuracy vs % of training samples for baseline methods and our proposed approach across all three datasets using a 4-layer CNN model.

	Google Speech Commands		**UrbanSound8K**		**ESC-50**	
Method	**% Samples**	**Accuracy**	**% Samples**	**Accuracy**	**% Samples**	**Accuracy**
Total Training Data	100%	87.45	100%	77.24	100%	67.62
Max Entropy	60%	69.92	60%	59.36	**60%**	**48.56**
	40%	66.25	40%	53.15	40%	44.78
	30%	63.55	30%	50.65	30%	39.05
	20%	56.18	20%	43.55	20%	34.56
	0.029%	8.27	0.063%	6.25	0.15%	4.85
Variation Ratios	60%	70.48	**60%**	**59.75**	60%	48.21
	40%	66.95	40%	53.5	40%	44.35
	30%	63.15	30%	49.87	30%	38.67
	20%	55.86	20%	43.78	20%	34.95
	0.029%	8.75	0.063%	5.92	0.15%	4.72
BALD	60%	70.27	60%	59.25	60%	47.75
	40%	66.18	40%	53.15	40%	43.85
	30%	62.67	30%	49.33	30%	38.75
	20%	56.25	20%	42.95	20%	33.48
	0.029%	8.35	0.063%	6.04	0.15%	4.3
Random	**60%**	**70.67**	60%	59.27	60%	48.05
	40%	67.05	40%	52.92	40%	44.72
	30%	62.18	30%	49.45	30%	38.02
	20%	56.67	20%	43.15	20%	35.05
	0.029%	8.96	0.063%	6.36	0.15%	4.67
Proposed Method	**0.029%**	**69.18**	**0.063%**	**58.52**	**0.15%**	**46.92**
	0.017%	57.45	0.038%	48.15	0.09%	28.05
	0.012%	45.67	0.025%	35.75	0.0625%	15.67

Table 3: Number of Audio Samples per Class (AS/C) across all three datasets for our proposed method.

Google Speech Commands		**UrbanSound8K**		**ESC-50**	
% Samples	**AS/C**	**% Samples**	**AS/C**	**% Samples**	**AS/C**
0.029%	50	0.063%	50	0.15%	5
0.017%	30	0.038%	30	0.09%	3
0.012%	20	0.025%	20	0.0625%	2

to BALD's 63.12% with 60% of the data, representing a nearly 1,000x reduction in labeled data. Similarly, on ESC-50, our method achieves 49.65% accuracy with only 0.15% of the data, closely aligning the performance of Variation Ratios (51.25%) which requires 60% of the training data. In addition, it is also evident that all baseline active learning methods exhibit poor performance across all three datasets when using a negligible number of training samples, comparable to the sample size in dataset distillation.

We observe similar trends with the 4-layer CNN model. Our approach achieves 69.18% accuracy on GSC using 0.029% of the training data, slightly below Random, which achieves 70.67%. On US8K, our method reaches 58.52% accuracy with 0.063% of the data, compared to Variation Ratios at 59.75% using 60% of the training data. For ESC-50, our method achieves 46.92% accuracy with 0.15% of the data, closely aligning with Max Entropy at 48.56%.

Additionally, as shown in Tables 1, 2, and 3, increasing the number of audio samples per class (AS/C) leads to noticeable improvements in accuracy. The trade-off between accuracy and data efficiency is evident with varying AS/C, and we hypothesize that further increases in AS/C would yield even better accuracy. Moreover, AS/C is not necessarily applicable to baseline active learning methods, as data points are typically not chosen based on class-wise selection.

Overall, our approach achieves up to a $\sim$3,000x reduction in labeled data while maintaining performance comparable to other active learning techniques. This balance demonstrates the potential of dataset distillation to significantly lower annotation costs with minimal impact on performance, making it a promising and efficient solution for resource-constrained audio classification tasks in real-world applications.

5 Conclusion and Future Work

In this work, we introduce dataset distillation as an efficient strategy to largely reduce the labeling burden for large-scale audio classification tasks, offering a complementary approach to active learning. By distilling high-dimensional audio data into compact, representative coresets, our approach preserves strong classification performance while significantly reducing the amount of audio data required for training. Experiments on three widely used public datasets show that our method achieves competitive results with traditional and modern active learning techniques, using up to ∼3,000x fewer data points. Leveraging dataset distillation as an alternative active learning technique not only enhances data efficiency but also ensures scalability, making it a promising solution for on-device training and resource-constrained environments. In future work, we aim to extend this approach to more complex architectures and tasks, incorporating human-in-the-loop studies, and exploring real-world deployment with human interaction.

References

[1] Linton C Freeman. *Elementary Applied Statistics: For Students in Behavioral Science*. John Wiley & Sons, 1965.

[2] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192, 2017.

[3] Gautham Krishna Gudur and Satheesh Kumar Perepu. Zero-shot federated learning with new classes for audio classification. In *Proc. Interspeech 2021*, pages 1579–1583, 2021.

[4] Gautham Krishna Gudur, Prahalathan Sundaramoorthy, and Venkatesh Umaashankar. Activeharnet: Towards on-device deep bayesian active learning for human activity recognition. In *The 3rd International Workshop on Deep Learning for Mobile Systems and Applications*, pages 7–12, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[7] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. In *Advances in Neural Information Processing Systems*, volume 35, pages 13877–13891, 2022.

[8] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *International Conference on Learning Representations*, 2021.

[9] Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. Rethinking cnn models for audio classification. *arXiv preprint arXiv:2007.11154*, 2020.

[10] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[11] Tara N Sainath and Carolina Parada. Convolutional neural networks for small-footprint keyword spotting. In *Interspeech*, pages 1478–1482, 2015.

[12] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 1041–1044, 2014.

[13] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[14] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2012.

[15] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[16] Raphael Tang and Jimmy Lin. Deep residual learning for small-footprint keyword spotting. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5484–5488, 2018.

[17] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[18] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[19] Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra. Hello edge: Keyword spotting on microcontrollers. *arXiv preprint arXiv:1711.07128*, 2017.

[20] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pages 12674–12685, 2021.

[21] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023.

[22] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021.