Improvement and cross domain evaluation of Slow-Fast-Networks

Ravi Kumar
Fraunhofer IDMT
Ilmenau, Germany
ravi.kumar@idmt.fraunhofer.de

Sascha Grollmisch
Fraunhofer IDMT
Ilmenau, Germany
goh@idmt.fraunhofer.de

Jakob Abeßer
Fraunhofer IDMT
Ilmenau, Germany
abr@idmt.fraunhofer.de

***Abstract*—Audio classification requires robust models which can capture both transient events and long-duration patterns. This poses a challenge for common single-stream convolutional neural networks (CNNs) that process short- and long-term events in a uniform fashion using the same filters. Inspired by the dual-stream processing of the human auditory system, SlowFast networks separate temporal and spectral analysis into parallel pathways. This study presents several enhancements to the SlowFast network, implementing uniform separable convolutions on both slow and fast streams to streamline the architecture and improve efficiency, while also introducing a lightweight version (SF-SC-small) with a 90% parameter reduction. We perform a comprehensive cross-domain evaluation of eight datasets that span speech, environmental sounds, industrial sounds, and bioacoustic sounds. The enhanced SlowFast models surpass the original SlowFast network as well as single-stream baselines like MobileNetV3 especially for single-label scenarios, while remaining competitive in multi-label tasks. The study highlights the benefits of dual-stream architectures for audio classification and underscores the importance of architectural design for audio classification.**

***Index Terms*—audio classification, dual-stream architecture, neural networks, separable convolutions**

I. Introduction

Audio classification is a fundamental component in many modern applications, including speech recognition, music genre identification, environmental sound monitoring, and industrial quality control. In real-world scenarios, multiple sound sources often overlap, and events can occur over diverse time scales, from short transient sounds, such as mechanical clicks in a factory, to long-lasting sounds, such as continuous machinery hum or extended vocalizations. Conventional single-stream convolutional neural networks (CNNs) typically apply rectangular filters uniformly throughout the spectrogram, limiting their ability to capture short- and long-duration patterns effectively.

Humans, on the contrary, navigate complex auditory scenes seamlessly. Neuroscience research identified two primary pathways in the human auditory cortex, often referred to as the dorsal (the "where") and ventral (the "what") pathways, which allow efficient processing of temporal dynamics and semantic content [1], [2]. Inspired by this dual-stream structure, SlowFast networks [3] have been proposed to handle multiple time scales within a single model. In the context of audio, one stream (the "slow" pathway) processes lower temporal resolution and captures broad spectral content over time, while the other stream (the "fast" pathway) operates at higher temporal resolution to emphasize transient features.

Despite advances in deep learning, audio classification still faces several key challenges. Temporal variability of sounds is a significant challenge, as audio signals exhibit rapid fluctuations over time, making pattern recognition on a single time-scale difficult. Sounds in different domains vary significantly in frequency content, duration, and structure further complicating this task. Audio recordings of real-world soundscapes contain overlapping sound events and background noises, which are often further affected by room-acoustic effects such as reverberation. Additionally, the spatial context of a recording plays a crucial role, as many sounds require interpretation based on their surrounding environment. For instance, a tonal beep sound may indicate an alarm in one setting and a simple button press in another. Previous work [4] has demonstrated the potential of SFNets for specific datasets such as EPIC-Kitchens [5] or VGG-Sound [6], but their general effectiveness across multiple audio domains ranging from speech to industrial sounds remains unexplored. Moreover, the computational demands of the proposed SFNets restrict their use on devices with limited resources.

To address the aforementioned challenges, we propose architectural refinements and a systematic cross-domain evaluation of SlowFast networks for audio classification. The original SFNet applies separable convolutions in all layers of the *fast pathway* but only in certain layers of the *slow pathway* without detailed reasoning. Therefore, we extend the use of separable convolutions across all layers of both pathways to simplify the model architecture. This modification improves computational efficiency while preserving critical spectral and temporal feature extraction. Furthermore, we introduce a lightweight variant of the SlowFast model that leverages depth-wise separable convolutions to significantly reduce the parameter footprint while improving or at least maintaining accuracy. We validate these modifications through an extensive cross-domain study on a diverse set of public datasets that span speech commands, environmental scenes, industrial sounds, and bioacoustic sounds. This broad evaluation allows us to assess how well a dual-stream network generalizes across disparate audio domains. To isolate the benefits of the dual-stream architecture, we benchmark our approach against the

single-stream MobileNetV3 architecture with model variants being trained from scratch or after an initial pre-training.

The remainder of this paper is organized as follows. Section II reviews related work for audio classification using multi-stream strategies. Section III details the proposed modifications to SlowFast networks. Section IV describes our experimental setup and datasets before Section V reports and analyzes the results. Finally, Section VI concludes the paper and discusses directions for future research.

II. Related Work

Previous work in audio classification has focused mainly on single-stream convolutional neural networks (CNNs) operating on time-frequency spectrogram input [7]. One common example that has been successfully applied to audio classification [8] is the VGG-style architecture [9], which consists of up to 19 convolutional layers with a filter size of 3×3 with pooling and fully connected classification layers. Deeper CNN variants such as ResNet [10] introduced residual connections to alleviate vanishing gradients in training, allowing networks to grow in depth and maintain stable gradient flow. ResNet-based architectures have been evaluated for various audio tasks, such as music instrument classification, keyword spotting, and industrial sound analysis [11], [12]. Typically, these architectures use two-dimensional convolutions with square filters, treating both time and frequency dimensions in a uniform way. However, this symmetry may be suboptimal because temporal and frequency information in audio have inherently different characteristics, such as rapid temporal transients or stable tones.

One line of improvement has been to modify CNN filter shapes and factorization to handle the non-uniform time-frequency patterns. For example, rectangular $k \times m$ filters (where $k \neq m$) have been explored as a way to separately tune frequency and temporal resolution within a single-stream model [13]. Previous studies found that the use of vertical or horizontal elongated filters can capture frequency-specific or temporal motifs more effectively than square filters [7]. Another related strategy is the use of separable convolutions that split a 2D convolution into consecutive $1\times k$ and $k\times1$ convolutions. This factorization reduces computational costs while still enabling the network to learn frequency- and time-specific features. For example, MobileNetV3 [14] employs depth-wise separable convolutions [15] throughout the network to achieve efficiency and performance for audio tasks. However, these single-stream models do not explicitly treat temporal dynamics differently from spectral information.

To address this limitation, multi-stream architectures process the same audio input in parallel streams, each specializing in different aspects of the signal. The SlowFast network is a dual-stream architecture originally introduced for video action recognition [3] and later adapted to audio [4]. SlowFast for audio employs dual convolutional streams operating on the same log-Mel spectrogram input. The Slow stream focuses on spectral details, as it uses a larger number of channels, processes the input at a reduced frame rate using strided sampling, and thereby prioritizes a detailed frequency analysis. In contrast, the Fast stream targets temporal patterns, as it operates at the full temporal resolution for capturing fast-changing audio events using fewer channels for efficiency. This dual-pathway architecture disentangles the processing of time and frequency components and allows each stream to specialize on a separate task. Both streams are fused at multiple stages using lateral connections, where intermediate outputs of the Fast stream are periodically merged into the Slow stream after scaling to match the temporal strides, see Figure 1. This multi-level fusion ensures that the Slow stream (frequency-focused) is enriched with temporal cues from the Fast stream before the final classification layer combines their outputs. Notably, the original SlowFast design uses separable convolutions in the Fast stream (splitting each 3×3 filter into 3×1 followed by 1×3) to better isolate temporal features, while the Slow stream mostly uses regular 2D convolutions without further reasoning.

The SlowFast model has demonstrated state-of-the-art performance on the VGG-Sound dataset [4] and outperformed the Temporal Segment Networks (TSN) [16] baseline on the EPIC-KITCHENS-100 audio set [5]. Ablation studies verified a clear complementary specialization of both streams. The Slow stream alone yielded around 45% accuracy, focusing on classes with steady or slowly evolving spectral content such as “mosquito buzzing” or “sea waves”, while the Fast stream alone achieved approximately 41% accuracy, excelling at rapidly changing or percussive sounds such as “woodpecker pecking” or “playing drum kit”. These results underscore the benefit of separate spectral and temporal processing of audio data. However, the original SlowFast network (SFNet) was evaluated only on those two datasets, leaving its generalizability to other audio domains an open question.

III. Proposed Modifications

We propose two modifications to the Slow-Fast network to simplify its architecture without sacrificing performance. The first modification, SFNet-SeparableConv (SF-SC), implements separable convolutions uniformly across all layers of both the Slow and Fast streams. This modification streamlines the architecture while improving the computational efficiency.

The second modification, SFNet-SeparableConv Small (SF-SC-small), builds upon the above depth-wise separable convolution design and introduces a more compact architecture. It employs a [2,2,3,2] residual block configuration (representing the number of blocks in each stage of the network), which substantially reduces the number of parameters compared to the original SF model, see Table II . Despite this reduction in model size, SFNet-Small preserves the dual-stream processing capability. The goal of this change is to balance efficiency and performance, creating a faster and lighter model while maintaining the benefits of the Slow-Fast dual-stream architecture.

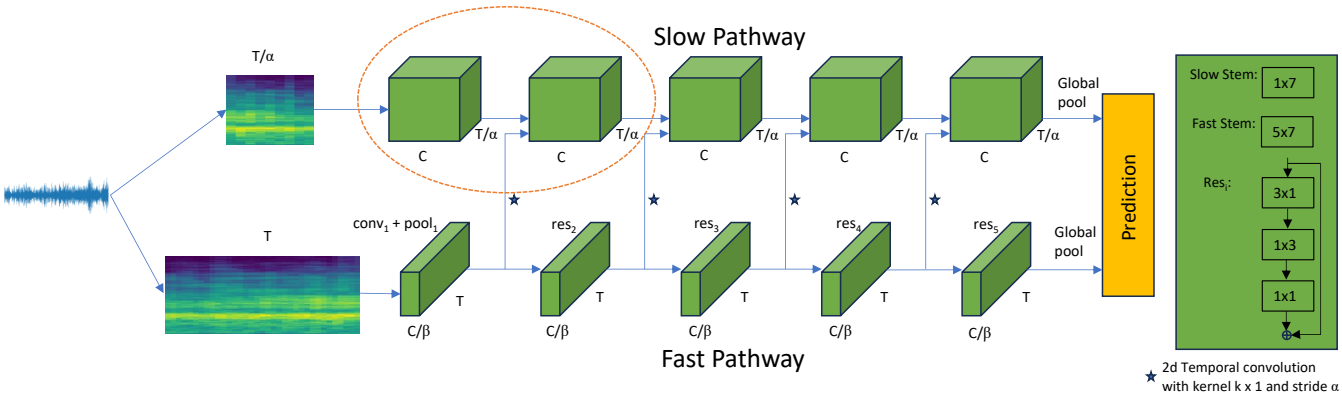

Fig. 1. Audio data processing flow in the SFNet architecture: Audio data is converted into a spectrogram, which is temporally downsampled using a striding operation in the Slow pathway (top) to process low temporal-resolution and input with the full-resolution into the Fast pathway (bottom) operating at high temporal resolution. Lateral connections (arrows) fuse Fast stream features into the Slow stream at multiple stages. The proposed modification uses separable convolutions (3×1 followed by 1×3 filters) across both pathways. **Note:** The star symbol ($\star$) indicates the 2D temporal convolution (kernel $k \times 1$, stride α). Dashed circles in the figure highlight where the new separable convolutions are integrated within each residual block.

TABLE I

OVERVIEW OF THE EIGHT AUDIO DATASETS USED FOR CROSS-DOMAIN EVALUATION, INCLUDING NUMBER OF CLASSES, PRE-PROCESSING DETAILS (SAMPLING RATE IN HZ, FFT SIZE AND HOP LENGTH IN SAMPLES), PATCH LENGTH (PL) IN SECONDS, FEATURE TYPE (MEL-SPECTROGRAM, MFCC), AND AUGMENTATION STRATEGIES (RANDAUGMENT, MIXUP, CUTOUT).

Dataset	Classes	FFT	Hop	SR (kHz)	PL(s)	PCEN/Log Scaling	Augmentation
TUT2017	15	2048	1024	44.1	3	Mel-Spec (PCEN)	Mixup, RA, Cutout
GSCv2	35	640 [17]	320	16	1	MFCC	Mixup, RA, Cutout
FMA	8	2048	710	44.1	3	STFT	Mixup, RA, Cutout
MB	3	2048	1024	32	1	STFT (log)	Mixup, RA, Cutout
Pucks	5	2048	1024	32	1	STFT (log)	No Mixup, BG Noise, Cutout
FSD50K	200	1024	441	44.1	1.28	Mel-Spec (log)	Mixup, RA, Cutout
USMv2	26	1024	441	44.1	5	Mel-Spec (log)	Multi-aug (grid dist, spec, erasing, brightness)
Bioacoustic	18	2048	1024	16	60	Mel-Spec (log)	Mixup, RA, Cutout

IV. EXPERIMENTS

A. Datasets

To evaluate SFNet including its modifications on a wide variety of audio signals, we use public datasets that cover everyday sounds, music, speech, and industrial sounds. Table I provides an overview of the key characteristics and processing setup for all datasets, including the number of classes, the pre-processing parameters (FFT size, hop length, sampling rate), the patch length (PL) used for the model input, the feature type and scaling (e.g., mel-spectrogram with PCEN or logarithmic scaling, MFCC), and the data augmentation techniques applied during training. For all datasets, we use the published splits into training, validation, and test sets. If no validation data was provided, we use a random 10% subset of the training data. Furthermore, we specifically analyze the performance for both single- and multi-label classification datasets. Single-label classification assumes that each audio segment contains a single dominant sound event, thus mapping one label to each instance. Conversely, multi-label classification addresses overlapping sound events, requiring the prediction of multi-hot labels for a single audio clip. This distinction becomes increasingly relevant in real-world scenarios, where simultaneous sound sources often complicate the labeling process.

Acoustic scene classification (TUT2017) The TUT Acoustic scenes 2017 dataset [18], [19] contains recordings of 15 different acoustic scene classes (e.g., café, park, metro station). This is a single-label environmental sound classification task in which each 10-second recording is labeled with exactly one scene class. The dataset comprises 13 hours of stereo audio, with 4,680 segments for training and 1,620 for testing.

Google Speech Commands v2 (GSCv2) [20] The Google Speech Commands v2 dataset is a collection of spoken commands designed for voice interface applications, featuring 35 distinct classes that include common keywords (e.g., up, down, and stop). This is a single-label classification task and the complete dataset contains 105,829 utterances recorded from 2,618 different speakers.

Free Music Archive (FMA) [21] The Free Music Archive dataset offers a single-label classification challenge where each recording is assigned to one musical genre. We focus on the "FMA Small" subset containing 25,000 30-second segments with 8 genres such as Pop, Folk, and Rock.

IDMT ISA Metal Balls Dataset (MB) [22] The IDMT ISA Metal Balls dataset [22] contains sounds of metal balls with three types of coatings (coated, eloxed, and broken) rolling down a metal slide. This single-label dataset simulates wear inside ball bearings, comprising 2,832 files (18.87 minutes

total).

IDMT-ISA-PUCKS Dataset (Pucks) [23] The IDMT-ISA-PUCKS dataset [23] features recordings for industrial material analysis, which contain sounds of air hockey pucks manufactured from four materials (Factory, ABS, PA2200, PA12) recorded under three different background noise levels. We train on the lowest noise level and test on the highest noise level to evaluate the performance against unknown background sounds – a common scenario in industrial sound analysis. This dataset consists of 260 one-minute recordings.

FSD50K Freesound Dataset (FSD50K) [24] The FSD50K Freesound dataset [24] represents a comprehensive collection of 51,197 audio clips that encompass 200 diverse sound classes (e.g., urban sounds such as car horn, musical instruments such as violin, and climate sounds such as rainfall), where each clip can contain multiple sound events. This curated dataset contains over 100 hours of audio divided into 40,966 training and 10,231 test clips with variable durations (0.3–30 seconds), all verified by human annotators.

Urban Sound Monitoring v2 (USMV2) The Urban Sound Monitoring v2 dataset [25] focuses on urban environmental sound event detection, featuring 26 distinct sound categories that encompass urban infrastructure (such as cars, buses, and trucks) and environmental events (for example, rain, thunderstorm, and wind). This represents a multi-label classification task where multiple sound events can occur simultaneously within each recording. The dataset contains 24,000 labeled sound clips totaling approximately 12 hours of audio, with recordings 5 seconds in duration.

Bioacoustics (DCASE2021 T5) [26] This bioacoustics dataset from Detection and Classification of Acoustic Scenes and Events Challenge 2021 Task 5 comprises several datasets of bird vocalizations in complex soundscapes. It is divided into training (14 hours 20 minutes), validation (5 hours), and evaluation sets, comprising a total of 4,996 labeled events across 19 different bird species, creating a multi-class classification problem with significant class imbalance. The evaluation set consists of 46 audio files acquired from different bioacoustic sources with varying sampling rates from 6 to 44.1 kHz, all resampled to 16 kHz.

B. Evaluation metrics

We use two key metrics to evaluate the model's performance: file-wise accuracy (acc.) and Mean Average Precision (mAP). For single-label tasks, where each instance is associated with just one category, file-wise accuracy is calculated as the fraction of instances that are correctly classified out of the total number of instances. This metric provides a straightforward measure of the overall effectiveness of the classification.

In contrast, mAP is fitting for multi-label scenarios, where instances can belong to multiple categories. It is calculated as the weighted average precision across all relevant labels, considering both precision (the fraction of true positives among predicted positives) and recall (the fraction of true positives among actual positives).

C. Processing Pipeline

For all datasets, we use the same processing pipeline to ensure comparability. This includes feature extraction, as detailed in Table I, as well as normalization using zero mean and unit variance per feature. Normalization values are calculated on the training set and applied to the validation and test set.

During training, in previous work various data augmentation strategies have been applied [11], [17], [23], [25]. These are listed in Table I. RandAugment (RA), as described in [11], applies two randomly selected image augmentations with random magnitude, such as brightness/contrast adjustments and affine distortions, to each sample. We also incorporate Mixup [27], which blends between random pairs of samples, and Cutout [28], which masks rectangular regions in the input feature. All models are trained until convergence on the validation with a batch size of 32 and Adam optimizer with a learning rate of 0.001. It is important to note that different feature extraction parameters and augmentation strategies might lead to different results. However, the focus of this work is to compare architectures and not the best results for each dataset.

V. Results

All results can be found in Table II. One main finding is that the performance of the single-stream baseline model in form of the MobileNetV3 (MV3) strongly depends on the applied pre-training strategy. The version trained from scratch (MV3U) consistently underperformed the pre-trained version (MV3P), even though it was trained only on image data.[1] This validates that pre-training can be beneficial in general and is in line with findings for other neural network architectures. For example, Patchout faSt Spectrogram Transformer (PaSST) required an image-based pre-training before further training iterations on audio data for best results [29]. Furthermore, different hyperparameters such as learning rate could have improved the results for MV3U on datasets such as TUT2017, however, for comparability, we decided to keep this the same. Due to the large performance difference and easy availability, we decide to use the MV3P as the baseline, but need to emphasize that all SFNets are trained from scratch and pre-training could have a beneficial effect as well.

Although SFNet performed better than MV3U on all datasets, it is on par with MV3P on single- and label multi-label tasks, and the advantage is dataset dependent. This shows that the SFNet can perform well across different audio tasks and domains. However, it is not necessarily superior to single-stream architectures, and there is no clear pattern on which kind of data benefits from the multi-stream approach.

The proposed modification of the SFNet (SF-SC) outperforms SFNet on all single-labels tasks by 7 percentage points on average accuracy. Especially the relatively small industrial sound datasets (MB and Pucks) benefit from this simplified model architecture. However, on multi-label datasets, it reduces the mAP by 1 percentage point. This gap is mostly

[1]We use the version provided in keras, which was pre-trained on ImageNet: https://www.tensorflow.org/api_docs/python/tf/keras/applications/MobileNetV3Large, last accessed 2025-03-06.

TABLE II
RESULTS FOR SINGLE-LABEL TASKS AND MULTI-LABEL DATASETS. MODEL VARIANTS INCLUDE THE ORIGINAL SFNET, SF-SC (UNIFORM SEPARABLE CONVOLUTIONS), SF-SC-SMALL (LIGHTWEIGHT VARIANT), AND MOBILENETV3 UNTRAINED (MV3U) & PRE-TRAINED (MV3P).

	Single label (file-wise accuracy)						Multi-label (mAP)				Params
Model	**TUT2017**	**GSCv2**	**FMA**	**MB**	**PUCKS**	**Avg.**	**FSD50K**	**USMv2**	**DCASE2021 T5**	**Avg.**	
SFNet	71.30	88.34	50.25	47.17	79.56	67.32	44.00	38.10	**46.00**	42.70	26M
SF-SC	**71.85**	88.70	50.38	**79.14**	88.89	**75.79**	43.50	38.40	42.00	41.30	26M
SF-SC-Small	65.62	73.74	48.25	77.00	**95.11**	71.94	37.30	**37.10**	43.70	39.37	**2M**
MV3(U)	27.90	89.93	48.75	33.33	20.00	43.98	44.90	14.80	15.40	25.03	3M
MV3(P)	67.04	**91.88**	**53.62**	54.97	65.78	66.66	**48.60**	36.30	43.30	**42.73**	3M

caused by the DCASE2021 T5 dataset. The reason for this difference needs to be investigated in future work. All in all, SF-SC is an efficient model that integrates the architecture, thus enhancing its classification and computational efficiency.

The main aim of the SF-SC-Small contains only about 2 million parameters, roughly a 90% reduction in size compared to SFNet. Although the average accuracy is reduced by 3 percentage points compared to SF-SC, it still performs better than the original SFNet and the larger MobileNetV3 trained from scratch and even pre-trained, indicating that the network design can be more important than the number of trainable parameters. For the multi-label tasks, SF-SC-Small performs better than the SF-SC on the DCASE2021 T5 dataset, however, the mAP on FSD50k is reduced from 43.5% to 37.3% which leads to a slight decease of 2 percentage points in the average mAP over all multi-label datasets. Overall, SF-SC-Small is an efficient alternative to the original SFNet, the SF-SC, and the MobileNetV3, especially for on-the-edge computing, since it uses far fewer parameters while maintaining most of the initial classification performance. Furthermore, these results show that the current results cannot be solely attributed to the number of trainable parameters but rather to the architectural choices made.

VI. CONCLUSION

Neural network architectures for automatic audio analysis are a key element of current machine listening research. Multi-stream architectures model the temporal variability of audio content using separate streams. In this paper, we simplified the dual-stream SlowFast network (SFNet) by applying uniform separable convolutions throughout both streams in the SFNet-SeparableConv (SF-SC). Furthermore, we investigated the effect of the model size on the performance by reducing the trainable parameters to 10% of the initial size in the SF-SC-small. Evaluation experiments on eight datasets from multiple audio domains show that all SFNet versions perform better than the single-stream MobileNetV3 trained from scratch. SF-SC serves as a valid and useful simplification of the original SFNet design, while SF-SC-Small demonstrates the feasibility of multi-stream architectures for on-the-edge audio classification.

Nevertheless, pre-training the MobileNetV3 closed the gap to the SFNets, suggesting that good initial weights may be as crucial as network design itself. Thus, our future work will address pre-training SFNet versions to unlock their full potential, for example, by applying self-supervised learning approaches on large audio corpora [30]. By combining strong initial representations with a multi-stream architecture, we aim to further enhance generalization and efficiency in diverse audio classification scenarios.

REFERENCES

[1] R. Santoro, M. Moerel, F. Di Martino, R. Goebel, K. Ugurbil, E. Yacoub, and E. Formisano, "Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex," *PLOS Computational Biology*, vol. 10, no. 1, pp. 1–14, 2014.

[2] I. Zulfiqar, M. Moerel, and E. Formisano, "Spectro-temporal processing in a two-stream computational model of auditory cortex," *Frontiers in Computational Neuroscience*, vol. 13, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:210840257

[3] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," 2019. [Online]. Available: https://arxiv.org/abs/1812.03982

[4] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Slow-fast auditory streams for audio recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 855–859.

[5] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Rescaling egocentric vision," *CoRR*, vol. abs/2006.13256, 2020. [Online]. Available: https://arxiv.org/abs/2006.13256

[6] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," 2020. [Online]. Available: https://arxiv.org/abs/2004.14368

[7] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep Learning for Audio Signal Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[8] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," 2017. [Online]. Available: https://arxiv.org/abs/1609.09430

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[11] S. Grollmisch and E. Cano, "Improving Semi-Supervised Learning for audio classification with FixMatch," *Electronics*, vol. 10, no. 15, 2021.

[12] B. Kim, S. Chang, J. Lee, and D. Sung, "Broadcasted Residual Learning for Efficient Keyword Spotting," *arXiv preprint arXiv:2106.04140*, 2021.

[13] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1706.07156

[14] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019, pp. 1314–1324.

[15] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2017. [Online]. Available: https://arxiv.org/abs/1610.02357

[16] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," 2016. [Online]. Available: https://arxiv.org/abs/1608.00859

[17] P. C. Wallbott, S. Grollmisch, and T. Köllmer, "Examining speaker and keyword uniqueness: Partitioning keyword spotting datasets for federated learning with the largest differencing method," in *Artificial Intelligence and Machine Learning*, T. Calders, C. Vens, J. Lijffijt, and B. Goethals, Eds. Cham: Springer Nature Switzerland, 2023, pp. 167–177.

[18] A. Mesaros, T. Heittola, and T. Virtanen, "TUT Acoustic scenes 2017, development dataset," Zenodo. https://zenodo.org/record/400515, last accessed 23/08/2022, 2017.

[19] ——, "TUT Acoustic scenes 2017, evaluation dataset," Zenodo. https://zenodo.org/record/1040168, last accessed 23/08/2022, 2017.

[20] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *ArXiv e-prints*, Apr. 2018. [Online]. Available: https://arxiv.org/abs/1804.03209

[21] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," 2016.

[22] S. Grollmisch, J. Abeßer, J. Liebetrau, and H. Lukashevich, "Sounding industry: Challenges and datasets for Industrial Sound Analysis," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, 2019, pp. 1–5.

[23] S. Grollmisch, D. Johnson, T. Krüger, and J. Liebetrau, "Plastic material classification using neural network based audio signal analysis," in *Sensor and Measurement Science International (SMSI)*, Online, 2020, pp. 337–338.

[24] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[25] J. Abeßer, "Classifying Sounds in Polyphonic Urban Sound Scenes," in *Proceedings of the 152nd Audio Engineering Society (AES) Convention*, Online, 2022.

[26] V. Morfi, D. Stowell, V. Lostanlen, A. Strandburg-Peshkin, L. Gill, H. Pamula, D. Benvent, I. Nolasco, S. Singh, S. Sridhar, M. Duteil, and A. Farnsworth, "Dcase 2021 task 5: Few-shot bioacoustic event detection development set (2.0)," Data set, 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5412896

[27] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 2018.

[28] T. DeVries and G. W. Taylor, "Improved regularization of Convolutional Neural Networks with Cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[29] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, Incheon, Korea, 2022, pp. 2753–2757.

[30] S. Liu, A. Mallol-Ragolta, E. Parada-Cabeleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, "Audio Self-supervised Learning: A Survey," *arXiv preprint arXiv:2203.01205*, 2022.