Composing and Validating Large-Scale Datasets for Training Open Foundation Models for Audio

Marianna Nezhurina[1,2,4] **Ke Chen**[3]
Yusong Wu[4,5] **Tianyu Zhang**[4] **Yuchen Hui**[5] **Haohe Liu**[6]
Taylor Berg-Kirkpatrick[3] **Shlomo Dubnov**[3] **Jenia Jitsev**[1,2]
[1]Juelich Supercomputing Center, Research Center Juelich
[2]LAION
[3]University of California San Diego
[4]Mila, Quebec Artificial Intelligence Institute
[5] Université de Montréal
[6]Centre for Vision Speech and Signal Processing, University of Surrey

Abstract

Obtaining strong reproducible foundation language-audio models require open datasets of sufficient scale and quality. To pre-train contrastive language-audio model we compose large-scale sound effects dataset with detailed text descriptions for each sample. Generating music, as a special type of audio, presents further challenges due to limited availability of music-text pairs with expressive enough captions. We show here how we combine various composed datasets to pre-train a large-scale audio-language contrastive model (CLAP). Then we train, on music samples we collected, a state-of-the-art text-to-music model, MusicLDM, that adapts AudioLDM, which is based on Stable Diffusion architecture, to the music domain, by utilizing pre-trained CLAP model and the Hifi-GAN vocoder, as components of MusicLDM. The modelling work validates thus composed text-audio and text-music datasets as strong basis for further studies on language-rooted foundation models for audio at larger scales.

1 Introduction

Language-audio modeling is a relatively new, but rapidly growing area that receives strong impulses from progress in language-vision modeling. While contrastive language-audio models like CLAP inherit representation learning ideas from CLIP [1], the goal in text-to-audio generation is to generate audio pieces, such as sound events, sound effects, and music, based on textual descriptions. Diffusion models have shown superior performance in these types of cross-modal generation tasks, with approaches for text-to-image generation like DALLE-2 [2] and Stable Diffusion [3]; and AudioGen [4], AudioLDM [5], and Make-an-Audio [6] for text-to-audio. Compared to other modalities such as text-to-image, there are relatively few text-music pairs with both high-quality audio and text available, making it difficult to train a high-quality conditional generative model.

In this paper, we focus on work done along both these lines. We report on datasets composed for training contrastive language-audio models and for training generative text-conditioned music models, using either sound effect focused or music focused data with detailed text descriptions. We describe our efforts to provide open-source versions of contrastive language-audio and text-to-music generative models pretrained on the composed open datasets.

We have developed a pipeline for contrastive audio-language pretraining using the collected language-sound effects dataset [7], performing extensive evaluation of pretrained models on wide range of

Preprint. Under review.

standardized benchmarking tasks such as text-to-audio retrieval, zero-shot audio classification, and supervised audio classification. Using the composed music focused dataset, we train a state-of-the-art text-to-music generation model, MusicLDM, adapting AudioLDM [5] architecture, which itself borrows design from the Stable Diffusion [3], to the music domain.

AudioLDM follows the same design idea as Stable Diffusion. While Stable Diffusion re-uses CLIP [1] and VQ-GAN [8], a pre-trained contrastive language-vision model and a vision encoder-decoder, AudioLDM relies on CLAP [7], pre-trained contrastive language-audio model that we have studied, and HiFi-GAN [9] as vocoder. Both Stable Diffusion and AudioLDM crucially benefit from open-source availability of corresponding model components. In our work, we open-source both dataset and model pipelines, allowing the community to replicate the results and build on the top of our work.

2 Methods

LAION-603k-Audio dataset We introduce LAION-Audio-630K, an extensive audio-text dataset with 633,526 pairs, totaling 4,325.39 hours of content. This dataset comprises a diverse range of audio content, including human activities, natural and animal sounds, and audio effects, sourced from 8 publicly available websites 1. Each audio sample has either text caption, class label, tags, filename that describes the audio or a combination of both. When working with datasets that only provide tags or labels, we expand these labels into captions by using a pre-trained language model T5 [10] to make captions on top of these keywords. Some audio samples have relative text (i.e. user comments) which we also integrated into our constructed captions.

Data Source	Number of Samples	Duration	Data Type
BBC sound effects	15973	463.48hrs	1 caption per audio, audio
Free To Use Sounds	6370	175.73hrs	Filename as caption, audio
Sonniss Game effects	5049	84.6hrs	Filename as caption, audio
We Sound Effects	488	12.00hrs	Filename as caption, audio
Paramount Motion Sound Effects	4420	19.49hrs	Filename as caption, audio
Audiostock	10000	46.30hrs	1 caption per audio, audio
Freesound [11]	515581	3003.38rs	1-2 captions per audio, audio
Epidemic Sound	75645	220.41hrs	2 captions per audio, audio

Table 1: LAION-Audio-630k Datasets

LAION-Audio-630K stands out as the largest publicly accessible audio-text dataset, surpassing previous datasets by a significant margin 2. All audio files have been converted to FLAC format with a mono channel at a 48kHz sample rate. This approach allows us to incorporate additional data into the training of the contrastive language-audio pretraining model. By combining all the datasets, we significantly augment the total number of audio samples paired with text captions, reaching a total of 2.5 million. Despite audio samples from datasets mentioned in Table 2 coming from different sources, we cannot exclude the overlap between them and LAION-Audio-630K. However, due to much larger scale of LAION-Audio-630K the overlap can be considered negligible.

Dataset	Pairs	Audio Durations (hrs)
Clotho[12]	5,929	37.00
SoundDescs[13]	32,979	1060.40
AudioCaps[14]	52,904	144.94
LAION-Audio-630K	633,526	4325.39

Table 2: LAION-Audio-630K compared with existing audio caption datasets.

Music focused dataset The original CLAP model trained on sound effect datasets. To work with music, we trained a CLAP model on 20 000 hours of music data in addition to its original training data, allowing it to better understand the relation between music and textual descriptions. For MusicLDM, we acquired the Audiostock dataset through web crawling of the Audiostock website, capturing URLs of audio samples along with associated descriptions and other metadata offered by the platform. The Audiostock dataset contains 9000 music tracks for training and 1000 tracks for testing. The total duration is 455.6 hours. It provides a correct textual description of each music track.

CLAP Architecture of CLAP is similar to one of the CLIP [1]: it has two decoders to separately process audio and text data. The model is trained using the contrastive learning approach, wherein it learns to align audio and text embeddings in pairs, employing the identical loss function as described

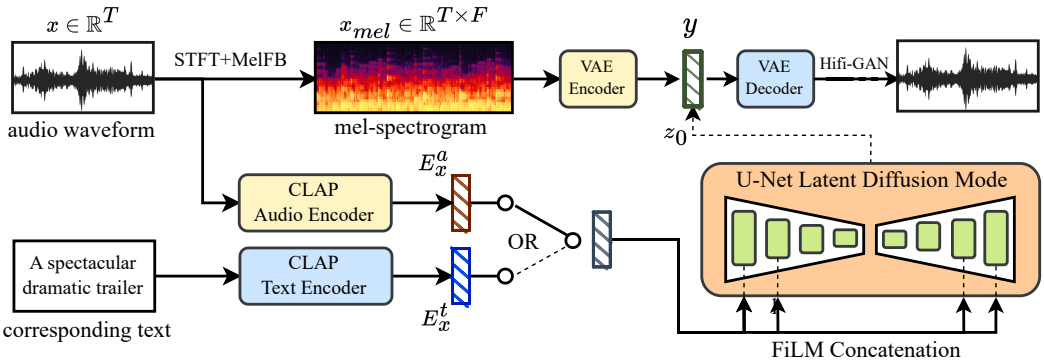

Figure 1: The architecture of MusicLDM, which contains a contrastive language-audio pretraining (CLAP) model, an audio latent diffusion model with VAE, and a Hifi-GAN vocoder.

in [1]. We select two models, PANN [15] and HTSAT [16], to construct the audio encoder. We select three models, CLIP transformer [1] (text encoder of CLIP), BERT [17], and RoBERTa [18], to construct the text encoder.

MusicLDM As illustrated in Figure 1, MusicLDM has similar architecture as AudioLDM: a contrastive language-audio pretraining (CLAP) model [19], an audio latent diffusion model [5] with a pretrained variational auto-encoder (VAE) [20], and a Hifi-GAN neural vocoder [9]. For MusicLDM, we make two changes from the original AudioLDM to enhance its performance on text-to-music generation. Since the original contrastive language-audio pretraining (CLAP) model is pretrained on text-audio pair datasets dominated by sound events, sound effects and natural sounds, we retrained the CLAP on text-music pair datasets (details in Supp. B) to improve its understanding of music data and corresponding texts. We also retrained the Hifi-GAN vocoder to ensure high-quality transforms from mel-spectrograms to music waveforms. Second, in the original AudioLDM, the model is only fed with audio embeddings as the condition during the training process, i.e., $\boldsymbol{E}_x = \boldsymbol{E}_x^a$; and it is fed with text embeddings to perform the text-to-audio generation, i.e., $\boldsymbol{E}_x = \boldsymbol{E}_x^t$. This approach leverages the alignment of text and audio embeddings inside CLAP to train the latent diffusion model with more audio data without texts.

3 Results.

CLAP performance The CLAP has undergone extensive evaluation across various audio tasks, including Zero-Shot and supervised audio classification (as depicted in Table 3), as well as text-audio retrieval. As evidenced by the results in Table 3, our CLAP model, pretrained on LAION-630k, exhibits a substantial performance boost compared to a model with a similar architecture [21]. This outcome underscores the superiority of LAION-630k as a dataset for foundational model pretraining.

Model	Audio Classification Dataset & Setting				
	ESC-50	US8K	VGGSound		FSD50K
	ZS.	ZS.	ZS.	SV.	SV.
Wav2CLIP[22]	41.4	40.4	10.0	46.6	43.1
AudioClip[23]	69.4	65.3	-	-	-
CLAP[24]	82.6	73.2	-	-	58.6
Ours	89.1	73.2	29.1	**75.4**	64.9
Ours+Fusion	88.0	75.8	26.3	75.3	64.4
Our+K2C Aug.	**91.0**	**77.0**	**46.2**	75.3	59.7
SoTA*	82.6[24]	73.2[24]	10.0[25]	64.1[26]	**65.6**[27]

Table 3: The zero-shot (ZS.) and supervised (SV.) audio classification results. The SoTA of each dataset/setting is denoted by the reference after the number.

We use HTSAT-RoBERTa as our best model to conduct the text-to-audio retrieval experiments as a comprehensive evaluation in Table 4, adopting the same metrics in [13] to compute recall scores at different ranks in this task. We gradually increase the scale of the dataset. We find that scaling up the dataset from "AudioCaps + Clotho" to "LA." does not improve the result on AudioCaps evaluation set but gets better performance on Clotho evaluation set, which can be explained by relative audio

Model	Training Set	AudioCaps Eval.						Clotho Eval.					
		T-A Retrieval			A-T Retrieval			T-A Retrieval			A-T Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
MMT [13]	AudioCaps or Clotho	36.1	**72.0**	**84.5**	39.6	76.8	86.7	6.7	21.6	33.2	7.0	22.7	34.6
ML-ACT [28]	AudioCaps or Clotho	33.9	69.7	82.6	39.4	72.0	83.9	14.4	36.6	49.9	16.2	37.6	50.2
CLAP-HTSAT [24]	AudioCaps + Clotho + WT5K	34.6	70.2	82.0	41.9	73.1	84.6	16.7	41.1	54.1	20.0	44.9	58.7
HTSAT-RoBERTa	AudioCaps + Clotho	**36.7**	70.9	83.2	45.3	78.0	87.7	12.0	31.6	43.9	15.7	36.9	51.3
HTSAT-RoBERTa	AudioCaps + Clotho + LA.	32.7	68.0	81.2	43.9	77.7	87.6	15.6	38.6	52.3	23.7	48.9	59.9
HTSAT-RoBERTa	ACaps. + Clotho + LA. + AudioSet (template)	34.7	70.5	83.2	45.3	79.5	89.2	16.4	39.0	51.0	21.8	44.6	60.1
HTSAT-RoBERTa	ACaps. + Clotho + LA. + AudioSet (K2C aug.)	36.1	71.8	83.9	**46.8**	**82.9**	**90.7**	16.1	38.3	51.1	22.7	48.5	60.8

Table 4: The text-to-audio retrieval performance on AudioCaps and Clotho datasets, where "LA." refers to LAION-Audio-630K, "template" refers to the text prompting by templates, "K2C aug." refers to the keyword-to-caption augmentation.

similirity of sample from AudioCaps and AudioSet on which the audio encoder's loaded checkpoint was pretrained. When we add more data from other sources to the training, we observe the increase in model's generalization.

MusicLDM performance Following evaluation techniques used in past work on audio generation [5], we use Frechet distance (FD), inception score (IS), and Kullback-Leibler (KL) divergence to evaluate the quality of generated outputs. Frechet distance evaluates the audio quality by using an audio embedding model to measure the similarity between the embedding space of generations and that of targets. In this paper, we use two standard audio embedding models: VGGish [29] and PANN [15]. The resulting distances we denote as FD_{vgg} and FD_{pann}, respectively. Inception score measures the diversity and the quality of the full set of audio outputs, while KL divergence is measured on individual pairs of generated and ground truth audio samples and averaged. We use the `audioldm_eval` library[1] to evaluate all the metrics mentioned above, comparing the ground truth audio from the Audiostock 1000-track test set. Table in Supplementary A 5 presents the FD, IS, and KL results for our models in comparison with baseline models (see details in Supplementary).

Table 5: The evaluation of generation quality among MusicLDMs and baselines. AA-Train. and TA-Train. refer to the audio-audio training scheme and the text-audio traning scheme.

Model	AA-Train.	TA-Train.	$FD_{pann}\downarrow$	$FD_{vgg}\downarrow$	Inception Score ↑	KL Div. ↓
Riffusion [30]	✗	✓	68.95	10.77	1.34	5.00
MuBERT [31]	—	—	31.70	19.04	1.51	4.69
AudioLDM	✓	✗	38.92	3.08	1.67	3.65
MusicLDM	✓	✗	26.67	2.40	**1.81**	3.80
MusicLDM (Only TA-Training)	✗	✓	32.40	2.51	1.49	3.96
MusicLDM w/. Text-Finetune	✓	✓	27.81	1.75	1.76	3.60

4 Conclusion & Outlook

In this paper, we demonstrate validation of open language-audio datasets by training strong baselines for two important audio model classes, contrastive language-audio (CLAP) and generative text-to-music model (musicLDM). We evaluate training on standardized benchmarks and show that both CLAP and musicLDM pre-trained on the composed datasets match strong results, either on zero-shot and fine-tuned supervised audio classification and text-audio retrieval tasks (CLAP), or on text-to-image generation tasks (musicLDM). MusicLDM incorporates CLAP, VAE, Hifi-GAN, and latent diffusion models as open-source components, showing power of re-usability and extendability given by open-source artefacts. We assess MusicLDM generated samples using objective and subjective metric as well as text-music relevance.

All datasets we composed in this work, including LAION-630k-audio [7] utilized in CLAP and other related tasks, are shown in our experiments to produce strong representational and generative language-audio models. The experimental results validate the composed datasets as suitable basis for pretraining various foundation audio models and open perspective for pretraining experiments on larger scales, including scaling law derivation. Our future research aims to explore methods for combining these datasets and other vastly available data in systematic manner to establish foundation audio reference dataset for training of robust foundation language-audio models that excel across various audio task types. This effort is similar to DataComp [32], an experimental community-driven testbed and challenge centered around finding the best possible dataset while fixing model architecture and training procedure.

[1] https://github.com/haoheliu/audioldm_eval

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint:2204.06125*, 2022.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pages 10684–10695, 2022.

[4] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually guided audio generation. *Proc. ICLR*, 2022.

[5] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. *Proc. ICML*, 2023.

[6] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint:2301.12661*, 2023.

[7] Authors will be shown upon submission. Title will be shown upon submission.

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[9] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HifiGAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Proc. NeurIPS*, 33:17022–17033, 2020.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[11] Frederic Font Corbera, Gerard Roma Trepat, and Xavier Serra. Freesound technical demo. In *Proc. ACM Multimed.*, 2013.

[12] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset, 2019.

[13] A. Sophia Koepke, Andreea-Maria Oncescu, Joã o F. Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25:2675–2685, 2023.

[14] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019.

[15] Qiuqiang Kong, Yin Cao, and Turab Iqbal et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE Trans. Audio, Speech, Lang. Process.*, 2020.

[16] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. In *Proc. ICASSP*, 2022.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, 2019.

[18] Yinhan Liu, Myle Ott, and Naman Goyal et al. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[19] Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proc. ICASSP*, pages 1–5. IEEE, 2023.

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *Proc. ICLR*, 2013.

[21] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: learning audio concepts from natural language supervision. *CoRR*, abs/2206.04769, 2022.

[22] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *Proc. ICASSP*, pages 4563–4567. IEEE, 2022.

[23] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *Proc. ICASSP*, 2022.

[24] Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. Audio retrieval with wavtext5k and CLAP training. *CoRR*, abs/2209.14275, 2022.

[25] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *Proc. ICASSP*, 2022.

[26] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Proc. NeurIPS*, 2021.

[27] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Proc. Interspeech*, 2022.

[28] Xinhao Mei, Xubo Liu, and Jianyuan Sun et al. On metric learning for audio-text cross-modal retrieval. In *Proc. Interspeech*, 2022.

[29] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *Proc. ICASSP*, pages 131–135. IEEE, 2017.

[30] Seth Forsgren and Hayk Martiros. Riffusion - Stable diffusion for real-time music generation. 2022.

[31] MubertAI. Mubert: A simple notebook demonstrating prompt-based music generation.

[32] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

[33] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 1998.

[34] Ashish Vaswani, Noam Shazeer, and Niki Parmar et al. Attention is all you need. In *Proc. NeurIPS*, 2017.

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *Proc. ICLR*, 2020.

[36] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[37] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *Proc. ICASSP*, pages 976–980. IEEE, 2022.

[38] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *Proc. ICASSP*, pages 1–5. IEEE, 2023.

[39] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *Proc. ICASSP*, pages 646–650. IEEE, 2022.

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint:1907.11692*, 2019.

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, 2014.

[42] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[43] Karol J. Piczak. ESC: dataset for environmental sound classification. In *Proc. ACM Multimed.*, 2015.

[44] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proc. ACM Multimed.*, 2014.

[45] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *Proc. ICASSP*, 2020.

[46] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.