AUDIOCLIP: EXTENDING CLIP TO IMAGE, TEXT AND AUDIO

Andrey Guzhov[1,2], Federico Raue[1], Jörn Hees[1], Andreas Dengel[1,2]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
[2]TU Kaiserslautern
Kaiserslautern, Germany
firstname.lastname@dfki.de

ABSTRACT

The rapidly evolving field of sound classification has greatly benefited from the methods of other domains. Today, the trend is to fuse domain-specific tasks and approaches together, which provides the community with new outstanding models.

We present AudioCLIP – an extension of the CLIP model that handles audio in addition to text and images. Utilizing the AudioSet dataset, our proposed model incorporates the ESResNeXt audio-model into the CLIP framework, thus enabling it to perform multimodal classification and keeping CLIP's zero-shot capabilities.

AudioCLIP achieves new state-of-the-art results in the Environmental Sound Classification (ESC) task and out-performs others by reaching accuracies of $97.15\,\%$ on ESC-50 and $90.07\,\%$ on UrbanSound8K. Further, it sets new baselines in the zero-shot ESC-task on the same datasets ($69.40\,\%$ and $68.78\,\%$, respectively).

We also asses the influence of different training setups on the final performance of the proposed model. For the sake of reproducibility, our code is published.

Index Terms— Audio, multimodal, zero-shot, classification

1. INTRODUCTION

The latest advances of the sound classification community provided many powerful audio-domain models that demonstrated impressive results. Combination of widely known datasets (such as AudioSet [1], UrbanSound8K [2] and ESC-50 [3]) and domain-specific and inter-domain techniques conditioned the rapid development of audio-dedicated methods and approaches [4, 5, 6].

Previously, researchers were focusing mostly on the classification task using the audible modality exclusively. In the last years, however, popularity of multimodal approaches in application to audio-related tasks has been increasing [7, 8, 9]. Being applied to such tasks, this implied the use of either textual or visual modalities in addition to sound. While using an additional modality together with audio is not rare, combination of more than two modalities is still uncommon in the audio domain. However, the restricted amount of high quality labeled data is constraining the development of the field in both, uni- and multimodal directions. Such a lack of data has challenged the research and sparked a cautious growth of interest for zero- and few-shot learning approaches based on contrastive learning methods that rely on textual descriptions [10, 11].

In our work, we propose an approach to combine the high-performance audio model ESResNeXt [5] with a contrastive text-image model, namely CLIP [12], thus, obtaining a *tri-modal* hybrid architecture. The base CLIP model demonstrates impressive performance and strong domain adaptation capabilities that are referred as "zero-shot inference" in the original paper [12]. To keep consistency with the CLIP terminology, we use the term "zero-shot" in the sense defined in [12]. As we will see, the joint use of three modalities during training results in out-performance of previous models in environmental sound classification task and extends zero-shot capabilities of the base architecture to audio as well.

The remainder of this paper is organized as follows: In Section 2 we discuss the current approaches to handle audio standalone and jointly with additional modalities and provide an overview of architectures implementing those capabilities. Then, we propose and describe a hybrid architecture combining visual, textual and audible modalities in Section 3, its training and evaluation in Section 4 and the obtained results in Section 5. Finally, we summarize our work and highlight follow-up research directions in Section 6.

2. RELATED WORK

In this section, we provide an overview of multimodal- and audio-related tasks and approaches that are intersecting in our work. Describing zero-shot learning, multimodal processing and environmental sound classification, we illustrate them by enumerating particular models that implement the aforementioned approaches.

Zero-shot learning (ZSL): In the common supervised learning setup, a model has to learn associations between input samples and corresponding categories, and the latter are (usually) provided through manual labeling of datasets. Although the use of human-annotated datasets provides models with high quality data, the growth of the amount of necessary samples often turns manual labeling into an expensive problem. ZSL aims to partially overcome the need to manually annotate all possible categories. In contrast, ZSL employs latent semantics provided by, for example, natural language corpora. Here, a model has to learn associations between samples and class label(s) not directly, but rather by finding correspondences between samples and some vector representations. Earlier, such representation consisted of manually defined attributes [13], while recent studies rely on vector embeddings of textual descriptions [10, 11], similarly to our proposed model.

Multimodal and contrastive learning: Unlike commonly occurring supervised learning setups, ZSL implies the use of an additional modality (usually textual) that serves to provide training targets, thus defining its multimodal nature implicitly. Going beyond zero-shot learning, combination of multiple modalities has steadily become more popular over the last years [14, 15]. One can observe that the joint use of more than one modality corresponds, in general, to video-related tasks, such as captioning [7], action recognition [9], retrieval [16], etc. Moreover, having different views of a same con-

This work was supported by the BMBF projects ExplAINN (Grant 01IS19074), EDeL (Grant 01IS19075), XAINES (Grant 01IW20005) and the TU Kaiserslautern PhD program.

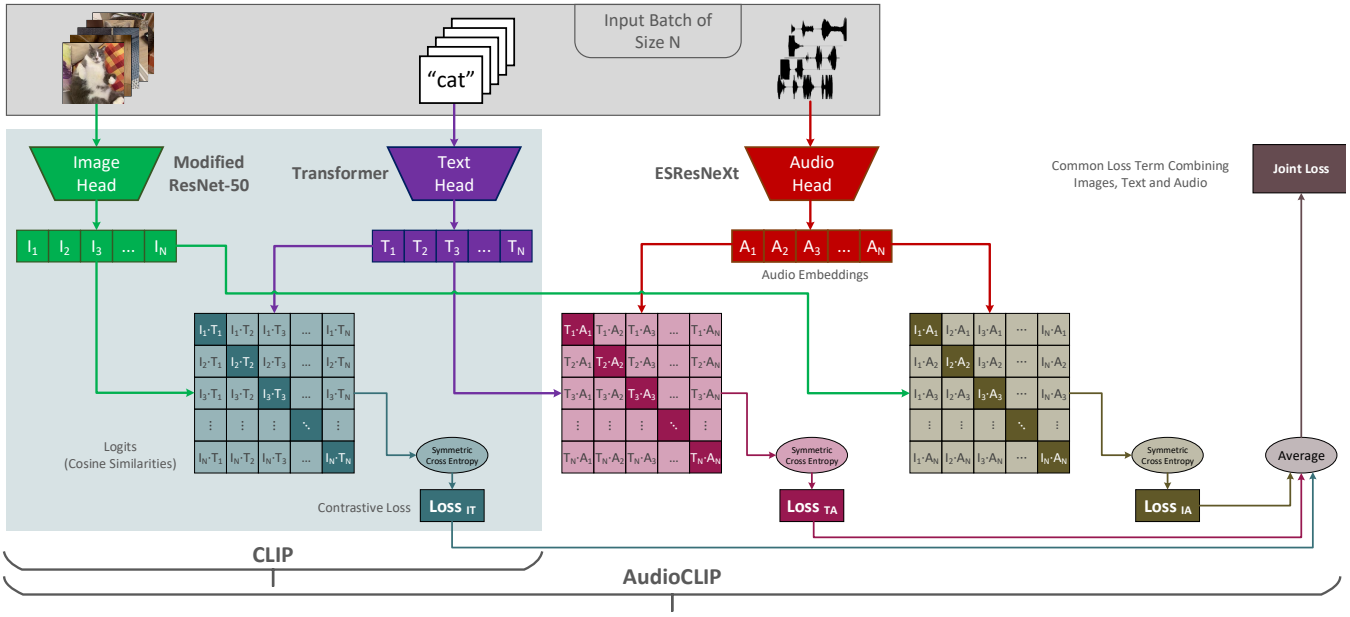

Fig. 1. Workflow of the proposed AudioCLIP model. On the left side we see the original CLIP model with Image and Text modalities. On the right we see our Audio-Head extension and its respective combinations with the Text (TA) and Image (IA) modalities. The alignment between two modalities is done by computing contrastive loss for the scaled dot-product across samples aiming to maximize diagonal values of the resulting similarity matrix using cross entropy function. The inter-modal loss terms are then averaged to come up with a joint *tri-modal* loss value.

cept is naturally encouraging to design models enabled to employ self-supervised training routines, thus overcoming the lack of data issue mentioned previously. Beside artificial self-supervised task (e.g., re-ordering of visual patches, colorization) such training procedures are in practice often based on the application of contrastive learning. Here, a model has to learn correspondences between multiple modalities that describe same entities by maximizing a similarity measure for different representations of a same sample, thus minimizing similarities to others. This approach was shown to be successfully applied to bi- [17] and to tri-modal [8] setups. Important results based on contrastive learning and the joint use of visual and textual representations of same concepts that are semantically aligned in a shared embedding space were obtained using the CLIP model [12]. In this work, we derive and extend the contrastive learning-based multimodal approach provided by the CLIP framework. Conceptually, CLIP consists of two branches: text and image encoding heads. Both heads were pre-trained jointly under natural language supervision [12]. Such a training setup enabled the model to generalize the classification ability to image samples that belonged to previously unseen datasets according to the provided textual labels without any additional fine-tuning. Given an input batch (text-image pairs) of size N, both CLIP-heads produce the corresponding embeddings that are mapped linearly into a multimodal embedding space of size 1024 [12]. In such a setup, CLIP learns to maximize the cosine similarity between matching textual and visual representations, while minimizing it between incorrect ones using a symmetric cross entropy loss over similarity measures [12].

Environmental sound classification (ESC): This task implies an assignment of correct labels given samples belonging to sound classes that surround us in the everyday life (e.g., "alarm clock", "car horn", "jackhammer", "mouse clicking", "cat"). To successfully address this challenge, different architectures and strategies were proposed and their performance was assessed using standard datasets such as UrbanSound8K [2] and ESC-50 [3]. The quite rare architectural branch of audio-specific Convolutional Neural Networks (CNN) is represented by one-dimensional CNNs [18]. Although such a network architecture is a natural approach to handle time-domain signals, practical applications showed that their performance seem to have reached a saturation point. In contrast, the very commonly used two-dimensional CNNs demonstrated steady improvement of accuracy, which is to be attributed to architectural enhancements [19], availability of high-quality large-scale datasets (such as AudioSet [1]) and the use of advanced training techniques [4]. Particularly, the development of 2D-CNNs applied to the ESC-task includes the use of specialized architectures [20, 21], application of general-purpose visual models standalone [22] and in a transfer learning setup [23, 24] as well as in combination with audio-domain specific architectural extensions [25, 5]. The recently proposed high-performance audio classification model ESResNeXt [5] is based on the residual architecture and consists of a trainable time-frequency transformation. We incorporate this ESResNeXt model into our proposed model as an audio-dedicated branch. The model contains a moderate number of parameters to learn ($\sim$ 30 M), while performing competitive on a large-scale audio dataset, namely AudioSet [1], and providing state-of-the-art-level classification results on the UrbanSound8K [2] and ESC-50 [3] datasets. Additionally, ESResNeXt supports an implicit processing of a multi-channel audio input and provides improved robustness against additive white Gaussian noise and sample rate decrease [5].

3. MODEL

In this section, we describe the key components that comprise our proposed model, the way how it handles its input and the training procedures. On a high level, our hybrid architecture combines a ResNet-based CLIP model [12] for visual and textual modalities and an ESResNeXt model [5] for audio, as shown in Figure 1.

Architecture: From the network design point of view, the proposed AudioCLIP model inherits from the novel CLIP framework and naturally extends the existing architecture by introducing an additional

audible modality. We consider the newly added modality as equally important as the originally present. Thus, AudioCLIP consists of image-, text- and audio-heads depicted in Figure 1.

Similar to CLIP, our proposed model learns to align different representations of a sample in a pair-wise manner across modalities. Such an alignment is done by generating M-dimensional vector embeddings per modality ($M = 1\,024$) and the subsequent computing of pair-wise cosine similarities across samples, where the latter serve as inputs for the symmetric cross entropy loss function [12].

The use of the third (audible) modality also introduces two new inter-modal alignment terms: text-audio (TA) and image-audio (IA), as shown in Figure 1 . Thus, AudioCLIP fuses the original CLIP's image-text contrastive learning setup (filled rectangle, on the left of Figure 1) and its audible modality-based extensions in order to enable the simultaneous processing of the three basic modalities in arbitrary combinations.

Training: Similarly to [12], AudioCLIP is trained using a contrastive learning approach that is based on the maximization of cosine similarities between multimodal embeddings of data samples using symmetric cross-entropy loss. The training of AudioCLIP consists of up to three stages.

For each experiment, modality-specific heads of AudioCLIP were pre-initialized (**Stage 1**). Image- and text-heads, which build up vanilla CLIP, used weights acquired by the authors of [12] on the composite CLIP dataset (Section 4). The audio-head, namely ESResNeXt, in all cases used pre-loaded ImageNet-weights and then was fitted by the authors of [5] to the AudioSet dataset, as shown to be beneficial for its initialization. We utilized snapshots of such trained CLIP and ESResNeXt as initial states for the branches of AudioCLIP.

Due to its nature, the AudioSet dataset also allows us to connect the different modalities (see description in Section 4). When trained on AudioSet, audio recordings, their textual labels and the corresponding video frames served as inputs for the audio-, text- and image-heads of AudioCLIP, respectively. In particular, audio tracks and the respective class labels were used to perform image-to-audio transfer learning for the ESResNeXt model (**Stage 2**), and then, the extracted frames in addition to audio and class names served as an input for the hybrid AudioCLIP model. During the training part, ten equally distant frames were extracted from a video, and one of them was picked randomly ($\sim \mathcal{U}$) and passed through the AudioCLIP model. In the evaluation phase, the same extraction procedure held, with the difference that only a central frame was presented to the model.

We performed the multimodal training based on the AudioSet dataset (**Stage 2**). For that, two different paths are available: either to keep the shared embedding space compatible to the one of vanilla CLIP (filled rectangle on the left of Figure 1) or to fit AudioCLIP better to the distribution of audio. The first option implies training of the audio-head while keeping image- and text-heads frozen ("frozen: IH, TH"), resulting in audio embeddings trying to align to the frozen CLIP model's embeddings. The second option, allows the full AudioCLIP to follow the audio distribution closer, as in this case a joint training of all three heads was performed (and improves accuracy on the target datasets).

Independent of the chosen AudioSet training strategy, we can also introduce a final fine-tuning (**Stage 3**, optional) to fit AudioCLIP to the target datasets, namely UrbanSound8K and ESC-50. We point out, that due to the nature of our model, we can do zero-shot inference on these datasets even without this traditional (and still beneficial) fine-tuning step, as will be evaluated in Table 2. In case of fine-tuning, it was performed for the audio-head exclusively, under the supervision of the rest of the AudioCLIP model (as the target datasets only consist of the audio modality). Freezing of the image- and text-heads implies exclusion of the corresponding network parameters from the backpropagation during the optimization step, thus keeping values of these parameters equal to the snapshot. That allows us to compare the performance of AudioCLIP relying on a vanilla CLIP initialization to the audio domain-adapted one.

Table 1. Evaluation results in audio classification task, accuracy (%).

Model	Source	Target	
		US8K	ESC
Human (2015)	[3]	–	81.30
Piczak-CNN (2015)	[20]	73.70	64.50
SB-CNN (2017)	[21]	79.00	–
ESResNet (2020)	[23]	85.42	91.50
WEANET N^4 (2020)	[4]	–	94.10
DenseNet-201×5 (2020)	[24]	87.42	92.89
ESResNeXt (2021)	[5]	89.14	95.20
AST (2021)	[19]	–	95.60
ERANN (2021)	[6]	–	96.10
AudioCLIP (frozen: IH, TH)		89.95	96.65
AudioCLIP		**90.07**	**97.15**

US8K: UrbanSound8K [2]; ESC: ESC-50 [3]; IH: Image-Head; TH: Text-Head.

4. EXPERIMENTAL SETUP

In this section, we describe the datasets that were used, the data augmentation methods we applied, the training's hyper-parameters and the performance evaluation methods.

Datasets: In this work, five image, audio and mixed datasets were used: two indirectly (CLIP dataset, ImageNet) as weight initializers and three directly (AudioSet, UrbanSound8K, ESC-50) for training and / or evaluation. The description of the aforementioned datasets is provided below.

1) *Composite CLIP dataset (init):* In order to train CLIP, a new dataset was constructed by its authors. The dataset consisted of roughly $400\,\mathrm{M}$ text-image pairs based on a set of $\sim 500\,\mathrm{k}$ text-based queries, and each query covered at least $\sim 20\,\mathrm{k}$ pairs [12].

2) *ImageNet (init):* ImageNet is a large-scale visual dataset described in [26] containing more than $1\,\mathrm{M}$ images across $1\,000$ classes.

3) *AudioSet:* Being proposed in [1], the AudioSet dataset provides a large-scale collection ($\sim 1.8\,\mathrm{M}$ / $\sim 20\,\mathrm{k}$, training / evaluation set) of audible data organized into 527 classes in a non-exclusive way. Each sample is a snippet up to $10\,\mathrm{s}$ long from a YouTube-video, defined by the corresponding ID and timings. For this work, we acquired video frames in addition to audio tracks. Thus, the AudioSet dataset became the glue between the vanilla CLIP framework and our tri-modal extension on top of it.

4) *UrbanSound8K:* The UrbanSound8K dataset provides $8\,732$ mono- and binaural audio tracks sampled at frequencies in the range $16 - 48\,\mathrm{kHz}$ (each track $\leq 4\,\mathrm{s}$) organized into 10 classes. To ensure correctness during the evaluation phase, the UrbanSound8K dataset was split by its authors into 10 non-overlapping folds [2] that we used in this work.

5) *ESC-50:* The ESC-50 dataset provides $2\,000$ single-channel $5\,\mathrm{s}$ long audio tracks sampled at $44.1\,\mathrm{kHz}$. As the name suggests, the dataset consists of 50 classes that can be divided into 5 major groups according to their nature. To ensure correctness during the evaluation phase, the ESC-50 dataset was split by its author into 5 non-overlapping folds [3] that we used in this work.

Data augmentation: In comparison to the composite CLIP dataset, the audio datasets provide two orders of magnitude less training samples, which makes overfitting an issue, especially for the

Table 2. Zero-Shot Comparison, accuracy (%). The audio head of each AudioCLIP variant is initialized with weights from ImageNet, then fine-tuned on AudioSet. In case of Zero-Shot inference there is no further fine-tuning to the target dataset (US8K / ESC here).

Model	Zero-Shot	Target	
		US8K	ESC
VGGish (2019) [10]	✓	–	26.00
VGGish (2021) [11]	✓	–	33.00
AudioCLIP (frozen: IH, TH)	✓	65.31 89.95	**69.40** 96.65
AudioCLIP	✓	**68.78** **90.07**	68.60 **97.15**

US8K: UrbanSound8K [2]; ESC: ESC-50 [3]; IH: Image-Head; TH: Text-Head.

UrbanSound8K and ESC-50 datasets. To address this challenge, we applied the following data augmentations.

1) *Time scaling:* Simultaneous change of track duration and its pitch is achieved using random scaling along the time axis. This kind of augmentation combines two computationally expensive ones, namely time stretching and pitch shift. Being a faster alternative to the combination of the aforementioned techniques, the time scaling in the range of random factors $[-1.5, 1.5]$, $\sim \mathcal{U}$ provides a lightweight though powerful method to fight overfitting [23].

2) *Time inversion:* Inversion of a track along its time axis relates to the random flip of an image, which is an augmentation technique that is widely used in the visual domain. The random time inversion (probability of 0.5) was applied to the training data similarly to [5].

3-4) *Random crop and padding:* Due to the requirement to align track duration before the processing through the model we applied random cropping or padding to the samples that were longer or shorter than the longest track in a dataset, respectively. During the evaluation, the random operation was replaced by the center one.

5) *Random noise:* The addition of random noise was shown to be helpful to overcome overfitting in visual-related tasks [27]. Also, the robustness evaluation of the ESResNeXt model suggested the improved sustainability of the chosen audio encoding model against the additive white Gaussian noise (AWGN) [5]. In this work, we extended the set of data augmentation techniques using AWGN, whose sound-to-noise ratio varied randomly ($\sim \mathcal{U}$) from 10.0dB to 120dB. The probability of the presence of the noise was set to 0.25.

Hyper-parameters: In this work, we trained our model on AudioSet, UrbanSound8K and ESC-50. In all training phases, the model parameters were optimized using Stochastic Gradient Descent [28] optimizer with Nesterov's momentum [29] of 0.9, weight decay of $5 \cdot 10^{-4}$ and batch size of 64. The learning rate value decreased exponentially, varying its value η and the decrease factor γ from 10^{-4} and 0.95, respectively, during the standalone pre-training of the audio-head to $5 \cdot 10^{-5}$ and 0.98 during the fine-tuning of AudioCLIP. The number of epochs was set to 30 for the AudioSet-based training, and to 50 for the fine-tuning to the downstream tasks.

Performance evaluation: Assessment of the AudioCLIP's performance was done based on the environmental sound classification task. Accuracy served to score classification capabilities of the model. Performance measurements were done on two datasets (UrbanSound8K and ESC-50) with common supervised and without fine-tuning to the target (*zero-shot inference*).

As described in Section 3, AudioCLIP, like vanilla CLIP [12], embeds multimodal inputs into a shared vector space and assesses similarities between different representations of same concepts. This allows us to define actual target classes *on-the-fly*, since the network's architecture does not consist of a dedicated classification layer, those outputs can be mapped onto the classes. Due to that, one has to perform an additional step preceding the actual classification, namely obtaining vector embeddings of the textual labels from the AudioCLIP's text-head. Once acquired, the subsequent calculation of cosine similarities between them and visual and / or audio embeddings can be done in order to identify most similar classes.

We describe the performance evaluation results in Section 5.

5. RESULTS

Audio classification: In the ESC-task, our proposed AudioCLIP model demonstrated benefits of the simultaneous use of multiple modalities during the training by achieving highest accuracy on the UrbanSound8K and ESC-50 datasets. In the audio classification task, AudioCLIP took advantage of the joint multimodal pre-training on AudioSet, as shown in Table 1. Specifically, the training of the audio-head under the supervision of the frozen text- and image-heads already allows to out-perform current state-of-the-art results on the UrbanSound8K and ESC-50 datasets by achieving accuracy of $89.95\,\%$ and $96.65\,\%$, respectively. Finally, the simultaneous training of image-, text- and audio-heads provides further performance improvements in comparison to the training of the audio-head exclusively. Such a trained AudioCLIP model sets the new state-of-the-art classification accuracy on the UrbanSound8K and ESC-50 datasets ($90.07\,\%$ and $97.15\,\%$, respectively).

Zero-shot inference: The composite CLIP dataset provides image- and text-heads with good initialization, which allows either to run inference in a zero-shot manner [12] or to achieve new state-of-the-art results by performing fine-tuning. The ImageNet dataset [26] that serves as a basic initialization for the audio-head was shown to be a good starting point for the cross-domain transfer learning [23, 24], and its combination with AudioSet [1] enables to build models that demonstrate state-of-the-art-level audio classification capabilities [5]. In particular, initialization of the image- and text-heads through the composite CLIP dataset is already enough to successfully train the audio-head under the multimodal supervision. Such a training (without fine-tuning to the actual targets—UrbanSound8K and ESC-50) results in a model that achieves highest *zero-shot inference* results on the UrbanSound8K and ESC-50 datasets—$65.31\,\%$ and $69.40\,\%$ (new state-of-the-art), respectively (Table 2). Furthermore, performing training of the image- and text-heads on AudioSet together with the audio-head can give an additional performance increase on the downstream tasks. Such a trained AudioCLIP sets new *zero-shot inference* baseline for UrbanSound8K ($68.78\,\%$, Table 2). Finally, the fine-tuning to the targets provides a valuable improvement resulting into state-of-the-art audio classification accuracy ($90.07\,\%$ @ UrbanSound8K, $97.15\,\%$ @ ESC-50, Table 2).

6. CONCLUSION

In this work, we extended CLIP [12] from textual and visual modalities to audio using an effective sound classification model [5]. The proposed AudioCLIP model achieves new state-of-the-art classification results on UrbanSound8K ($90.07\,\%$) and ESC-50 ($97.15\,\%$). To ease reproducibility, the details on hyper-parameters and implementation as well as weights of the trained models are made available for the community[1]. Additionally, for the zero-shot inference, our model out-performs previous approaches on the ESC-50 dataset with a large gap (from $33.00\,\%$ to $69.40\,\%$) and sets a baseline for UrbanSound8K ($68.78\,\%$). In the future, we would like to further investigate the performance of AudioCLIP on a wider variety of datasets and tasks. Also, changing the backbones of image- and audio-heads to more powerful networks could further improve the model performance.

[1] https://github.com/AndreyGuzhov/AudioCLIP

7. REFERENCES

[1] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[2] 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*, 2014, pp. 1041–1044.

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

[4] Anurag Kumar and Vamsi Ithapu, "A sequential self teaching approach for improving generalization in sound event recognition," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5447–5457.

[5] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, "Esresne(x)t-fbsp: Learning robust time-frequency transformation of audio," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021.

[6] Sergey Verbitskiy and Viacheslav Vyshegorodtsev, "Eranns: Efficient residual audio neural networks for audio pattern recognition," *arXiv preprint arXiv:2106.01621*, 2021.

[7] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[8] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman, "Self-supervised multimodal versatile networks," *arXiv preprint arXiv:2006.16228*, 2020.

[9] Jingran Zhang, Xing Xu, Fumin Shen, Huimin Lu, Xin Liu, and Heng Tao Shen, "Enhancing audio-visual association with self-supervised curriculum learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 3351–3359.

[10] Huang Xie and Tuomas Virtanen, "Zero-shot audio classification based on class label embeddings," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 264–267.

[11] Huang Xie and Tuomas Virtanen, "Zero-shot audio classification via semantic embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1233–1242, 2021.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[13] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 3, pp. 453–465, 2013.

[14] Relja Arandjelovic and Andrew Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[15] Relja Arandjelovic and Andrew Zisserman, "Objects that sound," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 435–451.

[16] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko, "Mdmmt: Multidomain multimodal transformer for video retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3354–3363.

[17] Luyu Wang, Pauline Luc, Adria Recasens, Jean-Baptiste Alayrac, and Aaron van den Oord, "Multimodal self-supervised learning of general audio representations," *arXiv preprint arXiv:2104.12807*, 2021.

[18] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada, "Learning from between-class examples for deep sound recognition," *arXiv preprint arXiv:1711.10282*, 2017.

[19] Yuan Gong, Yu-An Chung, and James Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[20] Karol J Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[21] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[22] Zhichao Zhang, Shugong Xu, Shunqing Zhang, Tianhao Qiao, and Shan Cao, "Learning attentive representations for environmental sound classification," *IEEE Access*, vol. 7, pp. 130327–130339, 2019.

[23] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, "Esresnet: Environmental sound classification based on visual domain models," in *25th International Conference on Pattern Recognition (ICPR)*, January 2021, pp. 4933–4940.

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

[25] Hardik B Sailor, Dharmesh M Agrawal, and Hemant A Patil, "Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification.," in *INTERSPEECH*, 2017, pp. 3107–3111.

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[27] Zeshan Hussain, Francisco Gimenez, Darvin Yi, and Daniel Rubin, "Differential data augmentation techniques for medical imaging classification tasks," in *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2017, vol. 2017, p. 979.

[28] Boris T Polyak and Anatoli B Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM journal on control and optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[29] Yu Nesterov, "A method of solving a convex programming problem with convergence rate o (1/kˆ 2) o (1/k2)," in *Sov. Math. Dokl*, 1983, vol. 27.