

Zero-shot Audio Source Separation through Query-based Learning from Weakly-labeled Data

Ke Chen^{1*}, Xingjian Du^{2*}, Bilei Zhu², Zejun Ma², Taylor Berg-Kirkpatrick¹, Shlomo Dubnov¹

¹ University of California San Diego, CA, USA

² Bytedance AI Lab, Shanghai, China

{knutchen, sdubnov, tberg}@ucsd.edu, {duxingjian.real, zhubilei, mazejun}@bytedance.com

Abstract

Deep learning techniques for separating audio into different sound sources face several challenges. Standard architectures require training separate models for different types of audio sources. Although some universal separators employ a single model to target multiple sources, they have difficulty generalizing to unseen sources. In this paper, we propose a three-component pipeline to train a universal audio source separator from a large, but weakly-labeled dataset: AudioSet. First, we propose a transformer-based sound event detection system for processing weakly-labeled training data. Second, we devise a query-based audio separation model that leverages this data for model training. Third, we design a latent embedding processor to encode queries that specify audio targets for separation, allowing for zero-shot generalization. Our approach uses a single model for source separation of multiple sound types, and relies solely on weakly-labeled data for training. In addition, the proposed audio separator can be used in a zero-shot setting, learning to separate types of audio sources that were never seen in training. To evaluate the separation performance, we test our model on MUSDB18, while training on the disjoint AudioSet. We further verify the zero-shot performance by conducting another experiment on audio source types that are held-out from training. The model achieves comparable Source-to-Distortion Ratio (SDR) performance to current supervised models in both cases.

Introduction

Audio source separation is a core task in the field of audio processing using artificial intelligence. The goal is to separate one or more individual constituent sources from a single recording of a mixed audio piece. Audio source separation can be applied in various downstream tasks such as audio extraction, audio transcription, and music and speech enhancement. Although there are many successful backbone architectures (e.g. Wave-U-Net, TasNet, D3Net (Stoller, Ewert, and Dixon 2018; Luo and Mesgarani 2018; Takahashi and Mitsufuji 2020)), fundamental challenges and questions remain: How can the models be made to better generalize to multiple, or even unseen, types of audio sources when super-

vised training data is limited? Can large amounts of weakly-labeled data be used to increase generalization performance?

The first challenge is known as universal source separation, meaning that we only need a single model to separate as many sources as possible. Most models mentioned above require training a full set of model parameters for each target type of audio source. As a result, training these models is both time and memory intensive. There are several heuristic frameworks (Samuel, Ganeshan, and Naradowsky 2020) that leverage meta-learning to bypass this problem, but they have difficulty generalizing to diverse types of audio sources. In other words, these frameworks succeeded in combining several source separators into one model, but the number of sources is still limited.

One approach to overcome this challenge is to train a model with an audio separation dataset that contains a very large variety of sound sources. The more sound sources a model can see, the better it will generalize. However, the scarcity of the supervised separation datasets makes this process challenging. Most separation datasets contain only a few source types. For example, MUSDB18 (Raffi et al. 2017) and DSD100 (Liutkus et al. 2017) contain music tracks of only four source types (vocal, drum, bass, and other) with a total duration of 5-10 hours. MedleyDB (Bittner et al. 2014) contains 82 instrument classes but with a total duration of only 3 hours. There exists some large-scale datasets such as AudioSet (Gemmeke et al. 2017) and FUSS (Wisdom et al. 2021), but they contain only weakly-labeled data. AudioSet, for example, contains 2.1 million 10-sec audio samples with 527 sound events. However, only 5% of recordings in Audioset have a localized event label (Hershey et al. 2021). For the remaining 95% of recordings, the correct occurrence of each labeled sound event can be anywhere within the 10-sec sample. In order to leverage this large and diverse source of weakly-labeled data, we first need to localize the sound event in each audio sample, which is referred as an audio tagging task (Fonseca et al. 2018).

In this paper, as illustrated in Figure 1, we devise a pipeline¹ that comprises of three components: a transformer-based sound event detection system ST-SED for performing time-localization in weakly-labeled training data, a query-based U-Net source separator to be trained from this data,

*The first two authors have equal contribution, and this work was performed while Ke Chen interned at Bytedance.
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹The official code is available in <https://git.io/JDWQ5>

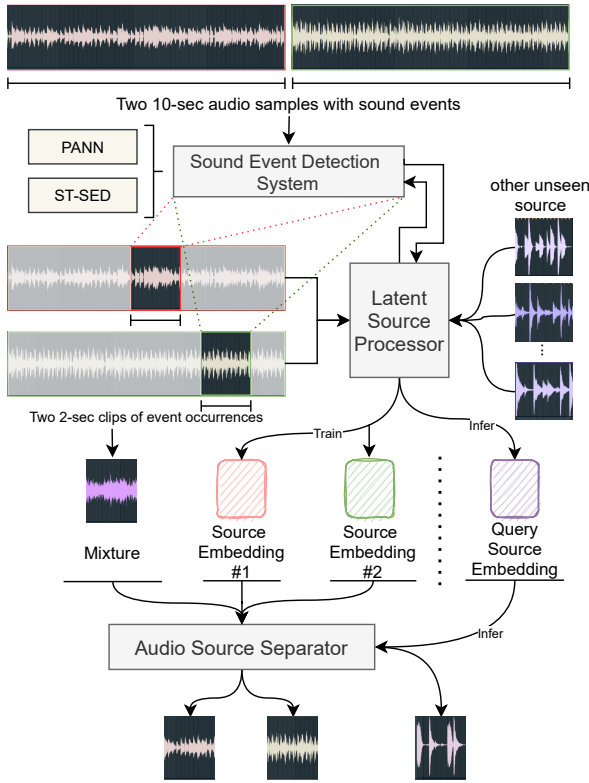


Figure 1: The architecture of our proposed zero-shot separation system.

and a latent source embedding processor that allows generalization to unseen types of audio sources. The ST-SED can localize the correct occurrences of sound events from weakly-labeled audio samples and encode them as latent source embeddings. The separator learns to separate out a target source from an audio mixture given a corresponding target source embedding query, which is produced by the embedding processor. Further, the embedding processor enables zero-shot generalization by forming queries for new audio source types that were unseen at training time. In the experiment, we find that our model can separate unseen types of audio sources, including musical instruments and held-out AudioSet’s sound classes, effectively by achieving the SDR performance on par with existing state-of-the-art (SOTA) models. Our contributions are specified as follows:

- We propose a complete pipeline to leverage weakly-labeled audio data in training audio source separation systems. The results show that our utilization of these data is effective.
- We design a transformer-based sound event detection system ST-SED. It outperforms the SOTA for sound event detection in AudioSet, while achieving a strong localization performance on the weakly-labeled data.
- We employ a single latent source separator for multiple types of audio sources, which saves training time and reduces the number of parameters. Moreover, we experimentally demonstrate that our approach can support zero-shot generalization to unseen types of sources.

Related Work

Sound Event Detection and Localization

The sound event detection task is to classify one or more target sound events in a given audio sample. The localization task, or the audio tagging, further requires the model to output the specific time-range of events on the audio timeline. Currently, the convolutional neural network (CNN) (LeCun et al. 1999) is being widely used to detect sound events. The Pretrained Audio Neural Networks (PANN) (Kong et al. 2020a) and the PSLA (Gong, Chung, and Glass 2021b) achieve the current CNN-based SOTA for the sound event detection, with their output featuremaps serving as an empirical probability map of events within the audio timeline. For the transformer-based structure, the latest audio spectrogram transformer (AST) (Gong, Chung, and Glass 2021a) re-purposes the visual transformer structure ViT (Dosovitskiy et al. 2021) and DeiT (Touvron et al. 2021) to use the transformer’s class-token to predict the sound event. It achieves the best performance on the sound event detection task in AudioSet. However, it cannot directly localize the events because it outputs only a class-token instead of a featuremap. In this paper, we propose a transformer-based model ST-SED to detect and localize the sound event. Moreover, we use the ST-SED to process the weakly-labeled data that is sent downstream into the following separator.

Universal Source Separation

Universal source separation attempts to employ a single model to separate different types of sources. Currently, the query-based model AQMSP (Lee, Choi, and Lee 2019) and the meta-learning model MetaTasNet (Samuel, Ganeshan, and Naradowsky 2020) can separate up to four sources in MUSDB18 dataset in the music source separation task. SuDoRM-RF (Tzinis, Wang, and Smaragdis 2020), the UniConvTasNet (Kavalerov et al. 2019), the PANN-based separator (Kong et al. 2020b), and MSI-DIS (Lin et al. 2021) extend the universal source separation to speech separation, environmental source separation, speech enhancement and music separation and synthesis tasks. However, most existing models require a separation dataset with clean sources and mixtures to train, and only support a limited number of sources that are seen in the training set. An ideal universal source separator should separate as many sources as possible even if they are unseen or not clearly defined in the training. In this paper, based on the architecture from (Kong et al. 2020b), we move further in this direction by proposing a pipeline that can use audio event samples for training a separator that generalizes to diverse and unseen sources.

Methodology and Pipeline

In this section, we introduce three components of our source separation model. The sound event detection system is established to refine the weakly-labeled data before it is used by the separation model for training. A query-based source separator is designed to separate audio into different sources. Then an embedding processor is proposed to connect the above two components and allows our model to perform separation on unseen types of audio sources.

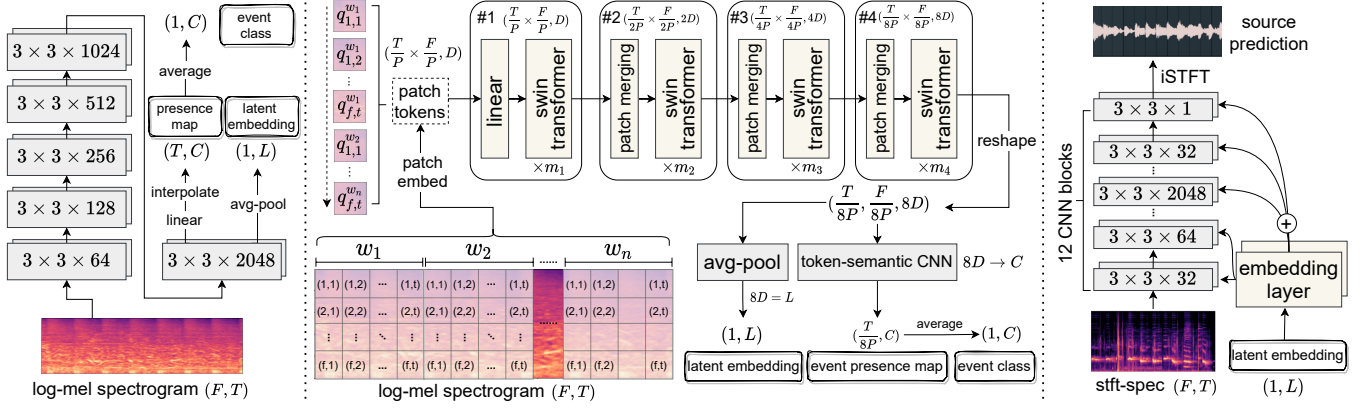


Figure 2: The network architecture of SED systems and the source separator. Left: PANN (Kong et al. 2020a); Middle: our proposed ST-SED; Right: the U-Net-based source separator. All CNNs are named as [2D-kernel size \times channel size].

Sound Event Detection System

In Audioset, each datum is a 10-sec audio sample with multiple sound events. The only accessible label is what sound events this sample contains (i.e., a multi-hot vector). However, we cannot get accurate start and end times for each sound event in a sample. This raises the problem of extracting a clip from a sample where one sound event most likely occurs (e.g., a 2-sec audio clip). As shown in the upper part of Figure 1, a pipeline is depicted by using a sound event detection (SED) system to process the weakly-labeled data. This system is designed to localize a 2-sec audio clip from a 10-sec sample, which will serve as an accurate sound event occurrence.

In this section, we will first briefly introduce an existing SOTA system: Pretrained Audio Neural Networks (PANN) (Left), which serves as the main model to compare in both sound event detection and localization experiments. Then we introduce our proposed system ST-SED (Middle) that leads to better performance than PANN.

Pretrained Audio Neural Networks As shown in the left of Figure 2, PANN contains VGG-like CNNs (Simonyan and Zisserman 2015) to convert an audio mel-spectrogram into a (T, C) featuremap, where T is the number of time frames and C is the number of sound event classes. The model averages the featuremap over the time axis to obtain a final probability vector $(1, C)$ and computes the binary cross-entropy loss between it and the groundtruth label. Since CNNs can capture the information in each time window, the featuremap (T, C) is empirically regarded as a presence probability map of each sound event at each time frame. When determining the latent source embedding for the following pipeline, the penultimate layer’s output (T, L) can be used to obtain its averaged vector $(1, L)$ as the latent source embedding.

Swin Token-Semantic Transformer for SED The transformer structure (Vaswani et al. 2017) and the token-semantic module (Gao et al. 2021) have been widely used in the image classification and segmentation task and achieve better performance. In this paper, we expect to bring similar improvements to the sound event detection and audio tag-

ging task, which then will contribute also to the separation task. As mentioned in the related work, the audio spectrogram transformer (AST) cannot be applied to audio tagging. Therefore, we refer to swin-transformer (Liu et al. 2021) in order to propose a swin token-semantic transformer for sound event detection (ST-SED). In the middle of Figure 2, a mel-spectrogram is cut into different patch tokens with a patch-embed CNN and sent into the transformer in order. We make the time and frequency lengths of the patch equal as $P \times P$. Further, to better capture the relationship between frequency bins of the same time frame, we first split the mel-spectrogram into windows w_1, w_2, \dots, w_n and then split the patches in each window. The order of tokens Q follows **time** \rightarrow **frequency** \rightarrow **window** as:

$$Q = \{q_{1,1}^{w_1}, q_{1,2}^{w_1}, \dots, q_{1,t}^{w_1}, q_{2,1}^{w_1}, q_{2,2}^{w_1}, \dots, q_{2,t}^{w_1}, \dots, q_{f,1}^{w_1}, q_{f,2}^{w_1}, \dots, q_{f,t}^{w_1}, q_{1,1}^{w_2}, q_{1,2}^{w_2}, \dots, q_{1,t}^{w_2}, q_{2,1}^{w_2}, q_{2,2}^{w_2}, \dots, q_{2,t}^{w_2}, \dots, q_{f,1}^{w_2}, q_{f,2}^{w_2}, \dots, q_{f,t}^{w_2}, \dots, q_{1,1}^{w_n}, q_{1,2}^{w_n}, \dots, q_{1,t}^{w_n}, q_{2,1}^{w_n}, q_{2,2}^{w_n}, \dots, q_{2,t}^{w_n}, \dots, q_{f,1}^{w_n}, q_{f,2}^{w_n}, \dots, q_{f,t}^{w_n}\}$$

Where $t = \frac{T}{P}$, $f = \frac{F}{P}$, n is the number of time windows, and $q_{i,j}^{w_k}$ denotes the patch in the position shown by Figure 2. The patch tokens pass through several network groups, each of which contains several transformer-encoder blocks. Between every two groups, we apply a patch-merge layer to reduce the number of tokens to construct a hierarchical representation. Each transformer-encoder block is a swin-transformer block with the shifted window attention module (Liu et al. 2021), a modified self-attention module to improve the training efficiency. As illustrated in Figure 2, the shape of the patch tokens is reduced by 8 times from $(\frac{T}{P} \times \frac{F}{P}, D)$ to $(\frac{T}{8P} \times \frac{F}{8P}, 8D)$ after 4 network groups.

We reshape the final block’s output to $(\frac{T}{8P}, \frac{F}{8P}, 8D)$. Then, we apply a token-semantic 2D-CNN (Gao et al. 2021) with kernel size $(3, \frac{F}{8P})$ and padding size $(1, 0)$ to integrate all frequency bins, meanwhile map the channel size $8D$ into the sound event classes C . The output $(\frac{T}{8P}, C)$ is regarded as a featuremap within time frames in a certain resolution. Finally, we average the featuremap as the final vector $(1, C)$ and compute the binary cross-entropy loss with the groundtruth label. Different from traditional visual transformers and AST, our proposed ST-SED does not use the

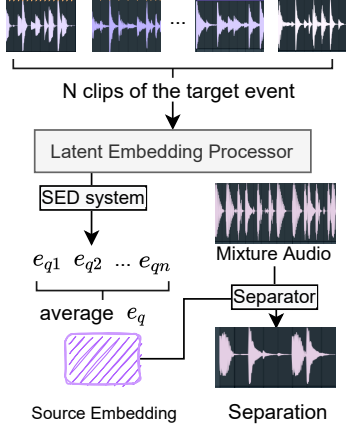


Figure 3: The mechanism to separate an audio into any given source. We collect N clean clips of the target event. Then we take the average of latent source embeddings as the query embedding e_q . The separator receives the embedding then performs the separation on the given audio.

class-token but the averaged final vector from the token-semantic layer to indicate the sound event. This makes the localization of sound events available in the output. In the practical scenario, we could use the featuremap $(\frac{T}{8P}, C')$ to localize sound events. And if we set $8D = L$, the averaged vector $(1, L)$ of the featuremap $(\frac{T}{8P}, L)$ can be used as the latent source embedding in line with PANN.

Query-based Source Separator

By SED systems, we can localize the most possible occurrence of a given sound event in an audio sample. Then, as shown in the Figure 1, suppose that we want to localize the sound event s_1 in the sample x_1 and another event s_2 in x_2 , we feed x_1, x_2 into the SED system to obtain two featuremaps m_1, m_2 . From m_1, m_2 we can find the time frame t_1, t_2 of the maximum probability on s_1, s_2 , respectively. Finally, we could get two 2-sec clips c_1, c_2 as the most possible occurrences of s_1, s_2 by assigning t_1, t_2 as center frames on two clips, respectively.

Subsequently, we resend two clips c_1, c_2 into the SED system to obtain two source embeddings e_1, e_2 . Each latent source embedding $(1, L)$ is incorporated into the source separation model to specify which source needs to be separated. The incorporation mechanism will be introduced in detail in the following paragraphs.

After we collect c_1, c_2, e_1, e_2 , we mix two clips as $c = c_1 + c_2$ with energy normalization. Then we send two training triplets $(c, c_1, e_1), (c, c_2, e_2)$ into the separator f , respectively. We let the separator to learn the following regression:

$$f(c_1 + c_2, e_j) \mapsto c_j, j \in \{1, 2\}. \quad (1)$$

As shown in the right of Figure 2, we base on U-Net (Ronneberger, Fischer, and Brox 2015) to construct our source separator, which contains a stack of downsampling and up-sampling CNNs. The mixture clip c is converted into the spectrogram by Short-time Fourier Transform (STFT). In each CNN block, the latent source embedding e_j is in-

corporated by two embedding layers producing two featuremaps and added into the audio featuremaps before passing through the next block. Therefore, the network will learn the relationship between the source embedding and the mixture, and adjust its weights to adapt to the separation of different sources. The output spectrogram of the final CNN block is converted into the separate waveform c' by inverse STFT (iSTFT). Suppose that we have n training triplets $\{(c^1, c_j^1, e_j^1), (c^2, c_j^2, e_j^2), \dots, (c^n, c_j^n, e_j^n)\}$, we apply the Mean Absolute Error (MAE) to compute the loss between separate waveforms $C' = \{c^{1'}, c^{2'}, \dots, c^{n'}\}$ and the target source clips $C_j = \{c_j^1, c_j^2, \dots, c_j^n\}$:

$$MAE(C_j, C') = \frac{1}{n} \sum_{i=0}^n |c_j^i - c^{i'}| \quad (2)$$

Combining these two components together, we could utilize more datasets (i.e. containing sufficient audio samples but without separation data) in the source separation task. Indeed, it also indicates that we no longer require clean sources and mixtures for the source separation task (Kong et al. 2020b, 2021) if we succeed in using these datasets to achieve a good performance.

Zero-shot Learning via Latent Source Embeddings

The third component, the embedding processor, serves as a communicator between the SED system and the source separator. As shown in Figure 1, during the training, the function of the latent source embedding processor is to obtain the latent source embedding e of given clips c from the SED system, and send the embedding into the separator. And in the inference stage, we enable the processor to utilize this model to separate more sources that are unseen or undefined in the training set.

Formally, suppose that we need to separate an audio x_q according to a query source s_q . In order to get the latent source embedding e_q , we first need to collect N clean clips of this source $\{c_{q1}, c_{q2}, \dots, c_{qN}\}$. Then we feed them into the SED system to obtain the latent embeddings $\{e_{q1}, e_{q2}, \dots, e_{qN}\}$. The e_q is obtained by taking the average of them:

$$e_q = \frac{1}{N} \sum_{i=1}^N e_{qi} \quad (3)$$

Then, we use e_q as the query for the source s_q and separate x_q into the target track $f(x_q, e_q)$. A visualization of this process is depicted in Figure 3.

The 527 classes of Audioset are ranged from ambient natural sounds to human activity sounds. Most of them are not clean sources as they contain other backgrounds and event sounds. After training our model in Audioset, we find that the model is able to achieve a good performance on separating unseen sources. According to (Wang et al. 2019), we declare that this follows a Class-Transductive Instance-Inductive (CTII) setting of zero-shot learning (Wang et al. 2019) as we train the separation model by certain types of sources and use unseen queries to let the model separate unseen sources.

Model	mAP
AudioSet Baseline (2017)	0.314
DeepRes. (2019)	0.392
PANN. (2020a)	0.434
PSLA. (2021b)	0.443
AST. (single) w/o. pretrain (2021a)	0.368
AST. (single) (2021a)	0.459
768-d ST-SED	0.467
768-d ST-SED w/o. pretrain	0.458
2048-d ST-SED w/o. pretrain	0.459

Table 1: The mAP results in Audioset evaluation set.

Experiment

There are two experimental stages for us to train a zero-shot audio source separator. First, we need to train a SED system as the first component. Then, we train an audio source separator as the second component based on the processed data from the SED system. In the following subsections, we will introduce the experiments in these two stages.

Sound Event Detection

Dataset and Training Details We choose AudioSet to train our sound event detection system ST-SED. It is a large-scale collection of over 2 million 10-sec audio samples and labeled with sound events from a set of 527 labels. Following the same training pipeline with (Gong, Chung, and Glass 2021a), we use AudioSet’s full-train set (2M samples) for training the ST-SED model and its evaluation set (22K samples) for evaluation. To further evaluate the localization performance, we use DESED test set (Serizel et al. 2020), which contains 692 10-sec audio samples with strong labels (time boundaries) of 2765 events in total. All labels in DESED are the subset (10 classes) of AudioSet’s sound event classes. In that, we can directly map AudioSet’s classes into DESED’s classes. There is no overlap between AudioSet’s full-train set and DESED test set. And there is no need to use DESED training set because AudioSet’s full-train set contains more training data.

For the pre-processing of audio, all samples are converted to mono as 1 channel by 32kHz sampling rate. To compute STFTs and mel-spectrograms, we use 1024 window size and 320 hop size. As a result, each frame is $\frac{320}{32000} = 0.01$ sec. The number of mel-frequency bins is $F = 64$. Each 10-sec sample constructs 1000 time frames and we pad them with 24 zero-frames ($T = 1024$). The shape of the output featuremap is (1024, 527) ($C = 527$). The patch size is 4×4 and the time window is 256 frames in length. We propose two settings for the ST-SED with a latent dimension size L of 768 or 2048. We adopt the 768-d model to make use of the swin-transformer ImageNet-pretrained model for achieving a potential best result. And we adopt the 2048-d model in the following separation experiment because it shares the consistent latent dimension size with PANN’s. We set 4 network groups in the ST-SED, containing 2,2,6, and 2 swin-transformer blocks respectively.

Validation Set: AudioSet Evaluation Set			
Metric-SDR: dB	mixture	clean	silence
527-d PANN-SEP (2020b)	7.38	8.89	11.00
2048-d PANN-SEP	9.42	13.96	15.89
2048-d ST-SED-SEP	10.55	27.83	16.64

Table 2: The SDR performance of different models with different source embeddings in the validation set.

We implement the ST-SED in PyTorch², train it with a batch size of 128 and the AdamW optimizer ($\beta_1=0.9$, $\beta_2=0.999$, $\text{eps}=1\text{e-}8$, $\text{decay}=0.05$) (Kingma and Ba 2015) in 8 NVIDIA Tesla V-100 GPUs in parallel. We adopt a warm-up schedule by setting the learning rate as 0.05, 0.1, 0.2 in the first three epochs, then the learning rate is halved every ten epochs until it returns to 0.05.

AudioSet Results Following the standard evaluation pipeline, we use the mean average precision (mAP) to verify the classification performance on Audioset’s evaluation set. In Table 1, we compare the ST-SED with previous SOTAs including the latest PANN, PSLA, and AST. Among all models, PSLA, AST, and our 768-d ST-SED apply the ImageNet-pretrained models. Specifically, PSLA uses the pretrained EfficientNet (Tan and Le 2019); AST uses the pretrained DeiT; and 768-d ST-SED uses the pretrained swin-transformer in Swin-T/C24 setting³. We also provide the mAP result of the 768-d ST-SED without pretraining for comparison. For the 2048-d ST-SED, we train it from zero because there is no pretrained model. For the AST, we compare our model with its single model’s report instead of the ensemble one to ensure the fairness of the experiment. All ST-SEDs are converged around 30-40 epochs in about 20 hours’ training.

From Table 1, we find that the 768-d pretrained ST-SED achieves a new mAP SOTA as 0.467 in Audioset. Moreover, our 768-d and 2048-d ST-SEDs without pretraining can also achieve the pre-SOTA mAP as 0.458 and 0.459, while the AST without pretraining could only achieve a low mAP as 0.368. This indicates that the ST-SED is not limited to the pretraining parameters of the computer vision model, and can be used more flexibly in audio tasks.

DESED Results We conduct an experiment on DESED test set to evaluate the localization performance of PANN and the 2048-d ST-SED. We do not include AST and PSLA since AST does not directly support the event localization and the PSLA’s code is not published. We use the event-based F1-score on each class as the evaluation metric, implemented by a Python library `psds_eval`⁴.

The F1-scores on all 10 classes in DESED by two models are shown in Table 3. We find that the 2048-d ST-SED achieves better F1-scores on 8 classes and a better average F1-score than PANN. A large increment is on the Frying class as increasing the F1-score by 40.92. However, we

²<https://pytorch.org/>

³<https://github.com/microsoft/Swin-Transformer>

⁴https://github.com/audioanalytic/psds_eval

Model	Alarm	Blender	Cat	Dishes	Dog	Shaver	Frying	Water	Speech	Cleaner	Average
PANN	34.33	42.35	36.31	17.60	35.82	23.81	9.30	30.58	69.68	51.01	35.08
ST-SED	44.66	52.23	69.98	27.35	49.93	43.90	50.22	42.76	45.11	41.55	46.77

Table 3: The F1-score results on each class of two models in DESED test set.

also notice that the F1-scores on Speech class and Cleaner class are dropped when using ST-SED, indicating that there are still some improvements for a better localization performance.

From the above experiments, we can conclude that the ST-SED achieves the best sound event detection results and the superior results on localization performance in AudioSet and DESED. These results are sufficient for us to use **the 2048-d ST-SED model** to conduct the following separation experiments. It is better to evaluate the ST-SED on datasets. Due to the page limit, we leave these as future work.

Audio Source Separation

Dataset and Training Details We train our audio separator in AudioSet full-train set, validate it in Audioset evaluation set, and evaluate it in MUSDB18 test set as following the 6th community-based Signal Separation Evaluation Campaign (SiSEC 2018). MUSDB18 contains 150 songs with a total duration of 3.5 hours in different genres. Each song provides a mixture track and four original stems: vocal, drum, bass, and other. All SOTAs are trained with MUSDB18 training set (100 songs) and evaluated in its test set (50 songs). Different from these SOTAs, we train our model only with Audioset full-train set other than MUSDB and directly evaluate it in MUSDB18 test set.

Since Audioset is not a natural separation dataset (i.e., no mixture data), to construct the training set and the validation set, during each training step, we sample two classes from 527 classes and randomly take each sample x_1, x_2 from two classes in the full-train set. We implement a balanced sampler that all classes will be sampled equally during the whole training. During the validation stage, we follow the same sampling paradigm to construct 5096 audio pairs from Audioset evaluation set and fix these pairs. By setting a fixed random seed, all models will face the same training data and the validation data.

For the model design, our SED system has two choices: PANN or ST-SED. And the separator we apply comprises 6 encoder blocks and 6 decoder blocks. In encoder blocks, the numbers of channels are namely 32, 64, 128, 256, 512, 1024. In decoder blocks, they are reversed (i.e., from 1024 to 32). There is a final convolution kernel that converts 32 channels into the output audio channel. Batch normalization (Ioffe and Szegedy 2015) and ReLU non-linearity (Agarap 2018) are used in each block. The final output is a spectrogram, which can be converted into the final separate audio c by iSTFT. Similarly, we implement our separator in PyTorch and train it with the Adam optimizer ($\beta_1=0.9, \beta_2=0.999, \text{eps}=1e-8, \text{decay}=0$), the learning rate 0.001 and the batch size of 64 in 8 NVIDIA Tesla V-100 GPUs in parallel.

Standard State-of-the-art Model				
Metric - Median SDR	vocal	drum	bass	other
WaveNet (2019)	3.25	4.22	3.21	2.25
WK (2014)	3.76	4.00	2.94	2.43
RGT1 (2018)	3.85	3.44	2.70	2.63
Spec-U-Net (2018)	5.74	4.66	3.67	3.40
UHL2 (2017)	5.93	5.92	5.03	4.19
MMDenseLSTM (2018)	6.60	6.41	5.16	4.15
Open Unmix (2019)	6.32	5.73	5.23	4.02
Demucs (2019)	6.21	6.50	6.21	3.80
Query-based Model w/. MUSDB18 Training				
Metric - Median SDR	vocal	drum	bass	other
AQMSP-Mean (2019)	4.90	4.34	3.09	3.16
Meta-TasNet (2020)	6.40	5.91	5.58	4.19
Zero-shot Model w/o. MUSDB18 Training				
Metric - Median SDR	vocal	drum	bass	other
527-d PANN-SEP (2020b)	4.16	0.95	-0.86	-2.65
2048-d PANN-SEP	6.06	5.00	3.38	2.86
2048-d ST-SED-SEP	6.24	5.39	4.03	3.07

Table 4: The SDR performance in MUSDB18 test set. All models are categorized into three slots.

Evaluation Metrics We use source-to-distortion ratio (SDR) as the metric to evaluate our separator. For the validation set, we compute three SDR metrics between the prediction and the groundtruth in different separation targets:

- mixture-SDR’s target: $f(c_1 + c_2, e_j) \mapsto c_j$
- clean-SDR’s target: $f(c_j, e_j) \mapsto c_j$
- silence-SDR’s target: $f(c_{\neg j}, e_j) \mapsto \mathbf{0}$

Where the symbol $\neg j$ denotes any clip which does not share the same class with the j -th clip. In our setting, $\neg 1 = 2$ and $\neg 2 = 1$. The clean SDR is to verify if the model can maintain the clean source given the self latent source embedding. The silence SDR is to verify if the model can separate nothing if there is no target source in the given audio. These help us understand if the model can be generalized to more general separation scenarios only by using the mixture training. For the testing, we only compute the mixture SDR between each stem and each original song in MUSDB18 test set. Each song is divided into 1-sec clips. The song’s SDR is the median SDR over all clips. And the final SDR is the median SDR over all songs.

The Choice of Source Embeddings We choose three source embeddings for our separator: (1) the 527-d presence probability vector from PANN, referring to (Kong

Class	Conversation	Whispering	Clapping	Cat	Orchestra	Aircraft	Medium Engine	Pour	Scratch	Creak	Average
Mixture-SDR	9.08	8.04	9.67	9.49	9.18	8.47	8.31	7.92	8.42	6.56	8.52
Clean-SDR	17.44	10.50	17.78	15.01	10.06	13.09	14.85	14.28	15.52	13.79	14.23
Silence-SDR	14.05	13.86	14.45	17.63	12.08	11.97	11.56	12.76	13.95	13.61	13.59

Table 5: The SDR performance of the 2048-d ST-SED-SEP in the zero-shot verification experiment.

et al. 2020b); (2) the 2048-d latent embedding from PANN’s penultimate layer; and (3) the 2048-d latent embedding from ST-SED. This helps to verify if the latent source embedding can perform a better representation for separation, and if the embedding from ST-SED is better than that from PANN.

In the training and validation stage, we get each latent source embedding directly from each 2-sec clip according to the pipeline in Figure 1. After picking the best model in the validation set, we follow Figure 3 to get the query source embeddings in MUSDB18. Specifically, we collect all separate tracks in the highlight version of MUSDB10 training set (30 secs in each song, 100 songs in total) and take the average of their embeddings on each source as four queries: vocal, drum, bass, and other.

Separation Results Table 2⁵ shows the SDRs of two models in the validation set. We could clearly figure out that when using the 2048-d latent source embedding, PANN achieves better performance in increasing three types of SDR by 2-4 dB. A potential reason is that the extra capacity of the 2048-d embedding space helped the model better capture the feature of the sound comparing to the 527-d probability embedding. In that, the model can receive more discriminative embeddings and perform a more accurate separation.

Then we pick the best models of 527-d PANN-SEP, 2048-d PANN-SEP, 2048-d ST-SED-SEP and evaluate them in MUSDB18. As shown in Table 4, there are three categories of models: (1) Standard Model: these models can only separate one source, in that they need to train 4 models to separate each source in MUSDB18. (2) Query-based Model: these models can separate four sources in one model. Both models in (1) and (2) require the training data in MUSDB training set and cannot generalize to separate other sources. And (3) Zero-shot Model: our proposed models can separate four sources in one model without any MUSDB18 training data. Additionally, they can even separate more sources.

From Table 4 our proposed model 2048-d ST-SED-SEP outperforms PANN-SEP models in all SDRs (6.24, 5.39, 4.03, 3.07). The SDRs in vocal, drum, and bass are compatible with standard and query-based SOTAs. However, we observe a relatively low SDR in the “other” source. One possible reason is that we compute a wrong “other” embedding for separation by averaging over all source embedding of “other” in MUSDB18 training set. But this “other” embedding might not be a general embedding because “other” denotes different instruments and timbres in different tracks. This sub-topic can be further researched in the future.

In summary, the most novel and surprising observation is that our proposed audio separator succeeds in separating

4 sources in MUSDB18 test set without any of its training data but only Audioset. The model performs as a zero-shot separator by using any latent source embedding collected from accessible data, to separator any source it faces.

Zero-Shot Verification

In this section, we conduct another experiment to let the model separate sources that are held-out from training. We first select 10 sound event classes in Audioset. Then during the training, we remove all data of these 10 classes. The model only learns how to separate clips mixed by the left 517 classes. During the evaluation, we construct 1000 (100×10) mixture samples in Audioset evaluation set whose constituents only belong to these 10 classes. Then we calculate the mixture SDR, the clean SDR, and the silence SDR of them.

Table 5 shows the results by the 2048-d ST-SED model. We can find that the model can still separate the held-out sources well by achieving the average mixture SDR, clean SDR, and silence SDR as 8.52 dB, 14.23 dB, and 13.59 dB. The detailed SDR distribution of these 1000 samples is depicted in the open source repository. The intrinsic reason for this good performance is that the SED system captures many features of 517 sound classes in its latent space. And it generalizes to regions of the embedding space it never saw during training, which the unseen 10 classes lie in. Finally, the separator utilizes these features in the embedding to separate the target source. The zero-shot setting of our model is essentially built by a solid feature extraction mechanism and a latent source separator.

Conclusion and Future Work

In this paper, we propose a zero-shot audio source separator that can utilize weakly-labeled data to train, target different sources to separate, and support more unseen sources. We train our model in Audioset while evaluating it in MUSDB18 test set. The experimental results show that our model outperforms the query-based SOTAs, meanwhile achieves a compatible result with standard supervised models. We further verify our model in a complete zero-shot setting to prove its generalization ability. With our model, more weakly-labeled audio data can be trained for the source separation problem. And more sources can be separated via one model. In future work, since audio embeddings have been widely used in other audio tasks such as music recommendation (Chen et al. 2021) and music generation (Chen et al. 2019; Chen, Xia, and Dubnov 2020; Chen 2021; Dong et al. 2020; Chen et al. 2020), we expect to use these audio embeddings as source queries to see if they can capture different audio features and lead to better separation performance.

⁵To avoid being out of bounds, we use `shortcite` allowed by AAAI Guide in the Table 2, and the Table 4.

References

- Agarap, A. F. 2018. Deep Learning using Rectified Linear Units (ReLU). arXiv:1803.08375.
- Bittner, R. M.; Salamon, J.; Tierney, M.; Mauch, M.; Cannam, C.; and Bello, J. P. 2014. MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014*, 155–160.
- Chen, K. 2021. Controllable Monophonic Music Generation Via Latent Variable Disentanglement. *Master Thesis Archive*.
- Chen, K.; Liang, B.; Ma, X.; and Gu, M. 2021. Learning Audio Embeddings with User Listening Data for Content-Based Music Recommendation. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021*, 3015–3019. IEEE.
- Chen, K.; Wang, C.; Berg-Kirkpatrick, T.; and Dubnov, S. 2020. Music SketchNet: Controllable Music Generation via Factorized Representations of Pitch and Rhythm. In *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020*, 77–84.
- Chen, K.; Xia, G.; and Dubnov, S. 2020. Continuous Melody Generation via Disentangled Short-Term Representations and Structural Conditions. In *IEEE 14th International Conference on Semantic Computing, ICSC 2020*, 128–135. IEEE.
- Chen, K.; Zhang, W.; Dubnov, S.; Xia, G.; and Li, W. 2019. The Effect of Explicit Structure Encoding of Deep Neural Networks for Symbolic Music Generation. In *International Workshop on Multilayer Music Representation and Processing, MMRP 2019*. IEEE.
- Défosssez, A.; Usunier, N.; Bottou, L.; and Bach, F. R. 2019. Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed. arXiv:1909.01174.
- Dong, H.; Chen, K.; McAuley, J. J.; and Berg-Kirkpatrick, T. 2020. MusPy: A Toolkit for Symbolic Music Generation. In *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020*, 101–108.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net.
- Fonseca, E.; Plakal, M.; Font, F.; Ellis, D. P. W.; Favory, X.; Pons, J.; and Serra, X. 2018. General-purpose Tagging of Freesound Audio with AudioSet Labels: Task Description, Dataset, and Baseline. arXiv:1807.09902.
- Ford, L.; Tang, H.; Grondin, F.; and Glass, J. R. 2019. A Deep Residual Network for Large-Scale Acoustic Scene Analysis. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019*, 2568–2572. ISCA.
- Gao, W.; Wan, F.; Pan, X.; Peng, Z.; Tian, Q.; Han, Z.; Zhou, B.; and Ye, Q. 2021. TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization. In *Premier International Computer Vision Event, ICCV 2021*.
- Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017*, 776–780. IEEE.
- Gong, Y.; Chung, Y.-A.; and Glass, J. 2021a. AST: Audio Spectrogram Transformer. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021*. ISCA.
- Gong, Y.; Chung, Y.-A.; and Glass, J. 2021b. PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation. *IEEE ACM Trans. Audio Speech Lang. Process. TASLP 2021*.
- Hershey, S.; Ellis, D. P. W.; Fonseca, E.; Jansen, A.; Liu, C.; Moore, R. C.; and Plakal, M. 2021. The Benefit of Temporally-Strong Labels in Audio Event Classification. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021*, 366–370. IEEE.
- Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, 448–456. JMLR.org.
- Kavalerov, I.; Wisdom, S.; Erdogan, H.; Patton, B.; Wilson, K. W.; Roux, J. L.; and Hershey, J. R. 2019. Universal Sound Separation. In *Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2019*, 175–179. IEEE.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.
- Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020a. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE ACM Trans. Audio Speech Lang. Process. TASLP 2020*, 28: 2880–2894.
- Kong, Q.; Liu, H.; Du, X.; Chen, L.; Xia, R.; and Wang, Y. 2021. Speech enhancement with weakly labelled data from AudioSet. arXiv:2102.09971.
- Kong, Q.; Wang, Y.; Song, X.; Cao, Y.; Wang, W.; and Plumbley, M. D. 2020b. Source Separation with Weakly Labelled Data: an Approach to Computational Auditory Scene Analysis. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020*, 101–105. IEEE.
- LeCun, Y.; Haffner, P.; Bottou, L.; and Bengio, Y. 1999. Object Recognition with Gradient-Based Learning. In *Shape, Contour and Grouping in Computer Vision*, volume 1681 of *Lecture Notes in Computer Science*, 319. Springer.
- Lee, J. H.; Choi, H.; and Lee, K. 2019. Audio Query-based Music Source Separation. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, 878–885.
- Lin, L.; Xia, G.; Kong, Q.; and Jiang, J. 2021. A unified model for zero-shot music source separation, transcription and synthesis. In *Proceedings of the 22nd International*

- Society for Music Information Retrieval Conference, ISMIR 2021*, 381–388.
- Liu, J.; and Yang, Y. 2018. Denoising Auto-Encoder with Recurrent Skip Connections and Residual Regression for Music Source Separation. In *17th International Conference on Machine Learning and Applications, ICMLA 2018*, 773–778. IEEE.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030.
- Liutkus, A.; Stöter, F.-R.; Rafii, Z.; Kitamura, D.; Rivet, B.; Ito, N.; Ono, N.; and Fontecave, J. 2017. The 2016 Signal Separation Evaluation Campaign. In *Latent Variable Analysis and Signal Separation - 12th International Conference, LVA/ICA 2015*, 323–332. Springer International Publishing.
- Lluís, F.; Pons, J.; and Serra, X. 2019. End-to-End Music Source Separation: Is it Possible in the Waveform Domain? In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019*, 4619–4623. ISCA.
- Luo, Y.; and Mesgarani, N. 2018. TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018*, 696–700. IEEE.
- Rafii, Z.; Liutkus, A.; Stöter, F.-R.; Mimitakis, S. I.; and Bitner, R. 2017. The MUSDB18 corpus for music separation.
- Roma, G.; Green, O.; and Tremblay, P. A. 2018. Improving Single-Network Single-Channel Separation of Musical Audio with Convolutional Layers. In *Latent Variable Analysis and Signal Separation - 14th International Conference, LVA/ICA 2018*, volume 10891 of *Lecture Notes in Computer Science*, 306–315. Springer.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, 234–241. Springer.
- Samuel, D.; Ganeshan, A.; and Naradowsky, J. 2020. Meta-Learning Extractors for Music Source Separation. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020*, 816–820. IEEE.
- Serizel, R.; Turpault, N.; Shah, A. P.; and Salamon, J. 2020. Sound Event Detection in Synthetic Domestic Environments. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020*, 86–90. IEEE.
- Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015*.
- Stoller, D.; Ewert, S.; and Dixon, S. 2018. Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, 334–340.
- Stöter, F.; Uhlich, S.; Liutkus, A.; and Mitsufuji, Y. 2019. Open-Unmix - A Reference Implementation for Music Source Separation. *J. Open Source Softw.*, 4(41): 1667.
- Takahashi, N.; Goswami, N.; and Mitsufuji, Y. 2018. Mmdenselstm: An Efficient Combination of Convolutional and Recurrent Neural Networks for Audio Source Separation. In *16th International Workshop on Acoustic Signal Enhancement, IWAENC 2018*, 106–110. IEEE.
- Takahashi, N.; and Mitsufuji, Y. 2020. D3Net: Densely connected multidilated DenseNet for music source separation. arXiv:2010.01733.
- Tan, M.; and Le, Q. V. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, 6105–6114. PMLR.
- Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, 10347–10357. PMLR.
- Tzinis, E.; Wang, Z.; and Smaragdis, P. 2020. Sudo RM -RF: Efficient Networks for Universal Audio Source Separation. In *30th International Workshop on Machine Learning for Signal Processing, MLSP 2020*, 1–6. IEEE.
- Uhlich, S.; Porcu, M.; Giron, F.; Enenkl, M.; Kemp, T.; Takahashi, N.; and Mitsufuji, Y. 2017. Improving music source separation based on deep neural networks through data augmentation and network blending. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017*, 261–265. IEEE.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 5998–6008.
- Wang, W.; Zheng, V. W.; Yu, H.; and Miao, C. 2019. A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Trans. Intell. Syst. Technol.*, 10(2).
- Weninger, F.; Hershey, J. R.; Roux, J. L.; and Schuller, B. W. 2014. Discriminatively trained recurrent neural networks for single-channel speech separation. In *2014 Global Conference on Signal and Information Processing, GlobalSIP 2014*, 577–581. IEEE.
- Wisdom, S.; Erdogan, H.; Ellis, D. P. W.; Serizel, R.; Turpault, N.; Fonseca, E.; Salamon, J.; Seetharaman, P.; and Hershey, J. R. 2021. What’s all the Fuss about Free Universal Sound Separation Data? In *International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021*, 186–190. IEEE.