SEMANTIC GROUPING NETWORK FOR AUDIO SOURCE SEPARATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio source separation is a typical and challenging problem that aims to separate individual sources from a mixture of audios. Recently, audio-visual separation approaches take advantage of the natural synchronization between the two modalities to boost separation performance. They extracted high-level semantics from visual inputs as the guidance to help disentangle sound representation for individual sources. Can we directly learn to disentangle the individual semantics from the sound itself? The dilemma is that multiple sound sources are mixed together in the original space. To tackle the difficulty, in this paper, we present a novel Semantic Grouping Network, termed as SGN, that can directly disentangle sound representations and extract high-level semantic information for each source from input audio mixture. Specifically, SGN aggregates category-wise source features through learnable class tokens of sounds. Then, the aggregated semantic features can be used as the guidance to separate the corresponding audio sources from the mixture. The proposed audio source separation framework is simple, flexible, and scalable. Comparing to the existing sound separation methods, our new framework can support audio separation from a flexible number of sources and is capable of generalizing to handle sound sources from different domains. We conducted extensive experiments on both music-only and universal sound separation benchmarks: MUSIC and FUSS. The results demonstrate that our SGN significantly outperforms previous audio-only methods and audio-visual models without utilizing additional visual cues.

1 INTRODUCTION

In our daily life, sound often appears to be a mixture of multiple sources, such as the sound in a classroom scene that consists of professors' speech and students whispering; the audio in a family reunion includes talking, laughing, and background music. Humans are capable of separating individual sources from these complex mixtures. For example, when we enjoy a beautiful melody, we are naturally aware of how many instruments are playing. This crucial human perception intelligence of auditory scenes attracts many researchers to explore the audio source separation problem.

Early audio-only works leveraged traditional machine learning approaches, such as hidden Markov models (Roweis, 2000; Mysore et al., 2010), Non-negative Matrix Factorization (Smaragdis & Brown, 2003; Virtanen, 2007; Cichocki et al., 2009) and Robust Principal Component Analysis (Huang et al., 2012), to solve the separation task. Benefiting from advances in deep learning, deep neural architectures, such as U-Net (Ulyanov et al., 2018) have been widely exploited to boost audio separation performance. Wave-U-Net (Stoller et al., 2018) adopted U-Net to the one-dimensional time domain to resample feature maps at multiple time scales for music source separation. In order to improve phase reconstruction, ResUNetDecouple+ (Kong et al., 2021) decoupled complex ideal ratio mask estimation into magnitude and phase estimation and proposed a residual U-Net architecture. Recently, some works (Wisdom et al., 2020; 2021) have started to explore source separation on universal audio mixtures in open domains. MixIT (Wisdom et al., 2020) developed an unsupervised learning algorithm with permutation invariant training (Yu et al., 2017) using a large-scale dataset: YFCC100M (Thomee et al., 2016). However, current audio-only methods can only handle fixed number of sources and they cannot learn compact representations for individual sources if just learning a direct mapping from mixture to individual sources. In contrast, we will solve them in our approach by extracting disentangled and compact representations as guidance for separation.

Since audio and visual contents are commonly matched and synchronized, audio-visual source separation methods (Hershey & Casey, 2001; Zhao et al., 2018; Ephrat et al., 2018; Gao et al., 2018; Xu et al., 2019; Gan et al., 2020b; Tian et al., 2021) are developed to improve separation. Unlike in audio space, sound sources are naturally separated in visual space. Thus, these methods can extract discriminative semantic information from visual inputs to help disentangle sound representation for reconstructing individual audio sources. SoP (Zhao et al., 2018) extracted pixel-level visual features from a dilated ResNet (He et al., 2016) to select the spectral components associated with the pixels. They then combined the magnitude of spectrogram with the phase of input spectrogram and applied inverse Short-Time Fourier Transform (STFT) to reconstruct the source waveform. A MinusPlus Net (Xu et al., 2019) was utilized to subtract salient sounds from the mixture and separate sounds recursively. A cyclic co-learning framework (Tian et al., 2021) was proposed to separate visual sounds sources with the help of sounding object visual grounding. While the methods achieve promising results on audio-visual source separation, they are extremely dependent on the ability of visual networks with large parameters. Without visual clues, their performance degrades significantly as observed in our experiments.

The success in audio-visual sound separation inspires us to ask a question: *can we directly disentangle the individual semantics from sound itself to guide separation*? The main challenge is that sounds are naturally mixed in the audio space. The individual semantics can be extracted from separated clean sources, but separation is the goal of our task. It is like a chicken or the egg problem. To tackle the dilemma, our key idea is to disentangle individual source representation using semantic-aware grouping to guide separation, which is different from the existing audio-only and audio-visual methods. During training, we aim to learn categorical codes to help disentangle and aggregate category-wise source features from the mixture for separation. The features will carry separated high-level semantic information for individual sources.

To this end, we propose a novel Semantic Grouping Network (SGN) that can learn to disentangle individual semantics from sound itself to guide source separation from audio mixtures. Specifically, we first learn class-aware features through category-aware grouping in terms of learnable source class tokens. Meanwhile, we use a U-Net to extract audio embeddeings from the input mixture. Our category-wise representations will serve as the semantic guidance to select their semantically corresponded audio features to reconstruct the corresponding audio sources. Our new framework is simple and flexible. During inference, it can separate different number of sources with respect to the aggregated category-specific semantics rather than a fixed number in existing audio-only separation methods[1]. In our implementation, audio waveforms are converted to spectrograms using Short-time Fourier transform (STFT) for model learning and the final waveforms of individual audio sources can be reconstructed by inverse STFT.

Experimental results on both MUSIC (Zhao et al., 2018) and FUSS (Wisdom et al., 2021) benchmarks can validate the superiority of our SGN against state-of-the-art audio source separation methods. Notably, without additional visual clues, our audio-only models can even achieve significant gains over audio-visual sound source separation methods. In addition, qualitative visualizations of separation results vividly showcase the effectiveness of our SGN in separating audio sources from mixtures. Extensive ablation studies also demonstrate the importance of category-aware grouping and learnable class tokens on learning compact representations for audio source separation.

Our main contributions can be summarized as follows:

- We present a novel Semantic Grouping Network, namely SGN, to disentangle the individual semantics from sound itself to guide source separation.
- We introduce learnable source class tokens in audio separation to aggregate category-wise source features with explicit high-level semantics.
- Extensive experiments on both music-only and universal separation datasets demonstrate the superiority of our SGN over state-of-the-art separation approaches.

[1]For past audio separation methods, a trained 2-sound separator will always produce 2 audio outputs. They cannot be used to handle mixtures with different number of sources.

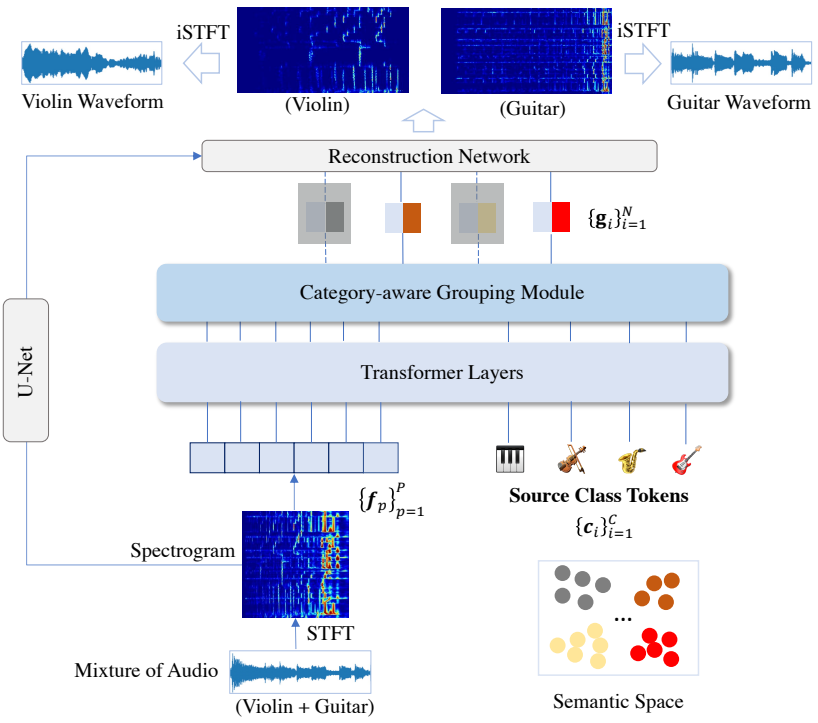

Figure 1: Illustration of our Semantic Grouping Network (SGN). The Category-aware Grouping module takes as input raw features of the input mixture spectrogram $\{\mathbf{f}_p\}_{p=1}^P$ and learnable class tokens $\{\mathbf{c}_i\}_{i=1}^C$ of for C categories in the semantic space to generate disentangled class-aware representations $\{\mathbf{g}_i\}_{i=1}^N$ for N sources. With the category-aware embeddings $\{\mathbf{g}_i\}_{i=1}^N$ and U-Net features of the mixture spectrogram, a light reconstruction network is used to reconstruct source spectrograms. Finally, inverse STFT is applied to recover the waveform of each source from spectrograms.

2 METHOD

Given a mixture of audios, our target is to separate all individual sound sources. We propose a novel Semantic Grouping Network named SGN for disentangling individual semantics from sound itself, which mainly consists of two modules, Source Class Tokens in Section 2.2 and Category-aware Grouping in Section 2.3.

2.1 PRELIMINARIES

In this section, we first describe the problem setup and notations, and then revisit an audio-visual separation method: Sound of Pixels (Zhao et al., 2018).

Problem Setup and Notations. Given a mixture of audio waveform, we apply Short-Time Fourier Transform (STFT) to extract a sound spectrogram with time T and frequency F from it. Our goal is to recover a matrix $\mathbf{W}$ of sound spectrograms with N sources, where $\mathbf{W} \in \mathbb{R}^{TF\times N}$. Note that T, F denotes the dimension of time and frequency of each source spectrogram.

Revisit Sound of Pixels. To solve the audio source separation problem, SoP (Zhao et al., 2018) extracted pixel-level visual features $\mathbf{P}$ to select the spectral components corresponding to the pixel, by leveraging the audio-visual synchronization. Since sounding sources are naturally separated in visual space, applying pixel-level features $\mathbf{P}$ on U-Net (Ulyanov et al., 2018) features $\mathbf{U}$ of the mixture spectrogram to recover the source matrix $\mathbf{W}$, which is denoted as

$$\mathbf{W} = \mathbf{UP} \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{TF\times D}, \mathbf{P} \in \mathbb{R}^{D\times N}$, and D is the dimension size of features. With the benefit of disentangled visual features, audio-visual separation methods achieved promising results on sound source separation.

However, those audio-visual approaches are extremely dependent on the capacity of the pre-trained visual networks to extract disentangled features. Without visual cues, their performance deteriorates significantly as shown in Sec 3.2. To address this issue, we are motivated by (Xu et al., 2022) and propose a novel Semantic Grouping Network that can learn to disentangle the individual semantics from sound itself to guide source separation from audio mixtures, as illustrated in Figure 1.

2.2 Source Class Tokens

In order to explicitly disentangle individual semantics from the mixed sound space, we introduce learnable source-specific class tokens $\{\mathbf{c}_i\}_{i=1}^{C}$ to help group semantic-aware information from raw features $\{\mathbf{f}_p\}_{p=1}^{P}$ that are extracted from the input mixture spectrogram via a convolution patch embedding layer (Dosovitskiy et al., 2020), where $\mathbf{c}_i \in \mathbb{R}^{1\times D}, \mathbf{f}_p \in \mathbb{R}^{1\times D}$, C is the total number of audio source classes, P is the total number of patches, and D is the dimension size.

With the categorical token embeddings, we first apply self-attention transformers $\phi(\cdot)$ to aggregate temporal features from the raw input as

$$\{\hat{\mathbf{f}}_p\}_{p=1}^{P}, \{\hat{\mathbf{c}}_i\}_{i=1}^{C} = \{\phi(\mathbf{x}_j, \mathbf{X}, \mathbf{X})\}_{j=1}^{P+C}, \mathbf{X} = \{\mathbf{x}_j\}_{j=1}^{P+C} = [\{\mathbf{f}_p\}_{p=1}^{P}; \{\mathbf{c}_i\}_{i=1}^{C}] \tag{2}$$

where $[\ ;\ ]$ denotes the concatenation operator. $\mathbf{f}_p, \mathbf{c}_i, \mathbf{x}_j \in \mathbb{R}^{1\times D}$, and D is the dimension of embeddings. The self-attention operator $\phi(\cdot)$ is formulated as:

$$\phi(\mathbf{x}_j, \mathbf{X}, \mathbf{X}) = \text{Softmax}(\frac{\mathbf{x}_j\mathbf{X}^\top}{\sqrt{D}})\mathbf{X} \tag{3}$$

Then, to constrain the independence of each class token $\mathbf{c}_i$ in the semantic space, we exploit a fully-connected (FC) layer and add softmax operation to predict the class probability: $\mathbf{e}_i = \text{Softmax}(\text{FC}(\mathbf{c}_i))$. Each category probability is constrained by a cross-entropy loss $\sum_{i=1}^{C}\text{CE}(\mathbf{h}_i, \mathbf{e}_i)$, where $\text{CE}(\cdot)$ is cross-entropy loss; $\mathbf{h}_i$ denotes an one-hot encoding and only its element for the target category entry i is 1. With optimizing the loss, it will push the learned token embeddings to be discriminative and category-specific.

2.3 Category-aware Grouping

Benefiting from the class-constraint loss mentioned above, we propose a novel and explicit category-aware grouping module $g(\cdot)$ to take audio mixture features and class tokens as inputs to generate source category-aware representations as:

$$\{\mathbf{g}_i\}_{i=1}^{C} = \{g(\{\hat{\mathbf{f}}_p\}_{p=1}^{P}, \hat{\mathbf{c}}_i)\}_{i=1}^{C} \tag{4}$$

In the time of grouping, we conflate features from the same source category token into the new category-aware representation according to a similarity matrix $\mathbf{A}$ between features and category token, which is calculated as:

$$\mathbf{A}_{p,i} = \text{Softmax}(W_q\hat{\mathbf{f}}_p, W_k\hat{\mathbf{c}}_i) \tag{5}$$

where $W_q \in \mathbb{R}^{D\times D}, W_k \in \mathbb{R}^{D\times D}$ are learnable weights of linear projection layers for features and category tokens. With the similarity matrix, we compute the weighted sum of temporal features to generate the category-aware semantic representation for category i as:

$$\mathbf{g}_i = g(\{\hat{\mathbf{f}}_p\}_{p=1}^{P}, \hat{\mathbf{c}}_i) = \hat{\mathbf{c}}_i + W_o\frac{\sum_{p=1}^{P}\mathbf{A}_{p,i}W_v\hat{\mathbf{f}}_p}{\sum_{p=1}^{P}\mathbf{A}_{p,i}} \tag{6}$$

where $W_o \in \mathbb{R}^{D\times D}, W_v \in \mathbb{R}^{D\times D}$ are learnable weights of linear projection layers for output and value. Taken category-aware representations $\{\mathbf{g}_i\}_{i=1}^{C}$ as the inputs, we use a FC layer and sigmoid operation to predict the binary probability: $p_i = \text{Sigmoid}(\text{FC}(\mathbf{g}_i))$ for ith category. By applying audio source classes $\{y_i\}_{i=1}^{C}$ as the weak supervision and combining the class-constraint loss, we formulate a semantic-aware grouping loss as:

$$\mathcal{L}_{\text{group}} = \sum_{i=1}^{C}\{\text{CE}(\mathbf{h}_i, \mathbf{e}_i) + \text{BCE}(y_i, p_i)\}. \tag{7}$$

Table 1: Quantitative results of audio source separation on MUSIC dataset.

Method	Input	SDR	SIR	SAR
RPCA (Huang et al., 2012)	audio-only	-0.62	2.32	2.41
NMF (Virtanen, 2007)	audio-only	0.86	3.26	3.81
SoP (Zhao et al., 2018) (w/o visual)	audio-only	2.16	5.58	5.54
Wave-U-Net (Stoller et al., 2018)	audio-only	3.80	6.75	6.62
ResUNetDecouple+ (Kong et al., 2021)	audio-only	3.98	7.17	6.91
SoP (Zhao et al., 2018)	audio-visual	4.55	10.06	10.24
MP-Net (Xu et al., 2019)	audio-visual	4.82	10.19	10.56
SGN (ours)	audio-only	**5.20**	**10.81**	**10.67**

Since multiple audio sources could be in one mixture, we use binary cross-entropy loss: BCE$(\cdot)$ for each cateogry to handle this multi-label classification problem.

Similar to SoP (Zhao et al., 2018), we adopt a binary mask matrix $\mathbf{M} \in \mathbb{R}^{TF \times N}$ that separates the mixture spectrogram as the final target for stabilized training, that is, $\mathbf{M} = \{\mathbf{m}_i\}_{i=1}^{N}, \mathbf{m}_i \in \mathbb{R}^{TF}$. The binary mask $\mathbf{m}_i$ is computed by predicting whether the ith source is the dominant component in the mixed sound as $\mathbf{m}_i = \mathbf{w}_i > 0.5\mathbf{w}_{mix}$, where $\mathbf{w}_i, \mathbf{w}_{mix} \in \mathbb{R}^{TF}$ denotes the ith source and mixture spectrogram, respectively. With the category-aware embeddings $\{\mathbf{g}_i\}_{i=1}^{C}$ and U-Net features $\mathbf{U}$ of the mixture spectrogram as inputs, we adopt a light reconstruction network composed of an inner product operator and learnable bias parameters to reconstruct spectrogram masks $\{\hat{\mathbf{m}}_i\}_{i=1}^{N}$ with N sources. Finally, a reconstruction loss is formulated with the sum of binary cross-entropy for N sources as:

$$\mathcal{L}_{\text{rec}} = \sum_{i=1}^{N} \text{BCE}(\mathbf{m}_i, \hat{\mathbf{m}}_i) \tag{8}$$

where N denotes the number of mixture and m_i refers to the ground-truth mask. The overall objective of our model is simply optimized in an end-to-end manner as:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{group}} \tag{9}$$

During inference, we multiply the predicted N individual masks: $\{\hat{\mathbf{m}}_i\}_{i=1}^{N} \in \mathbb{R}^{TF \times N}$ by the mixture spectrogram $\mathbf{W}_{mix} \in \mathbb{R}^{TF}$ to generate the audio spectrograms for separated sources, that is, $\{\hat{\mathbf{w}}_i\}_{i=1}^{N} = \{\mathbf{W}_{mix} \odot \hat{\mathbf{m}}_i\}_{i=1}^{N}$, where $\odot$ denotes element-wise multiplication. $\hat{\mathbf{w}}_i$ is reshaped to a $T \times F$ spectrogram and inverse STFT is applied to recover the waveform of each audio source.

Relation to Sound of Pixels. Recall that SoP (Zhao et al., 2018) separated U-Net features $\mathbf{U} \in \mathbb{R}^{TF \times D}$ into source matrix $\mathbf{W}$ and visual features $\mathbf{P}$ in Eq. 1. In contrast to their solutions, the main motivation of our SGN is to separate U-Net features $\mathbf{U}$ into source matrix $\mathbf{W}$ with respect to disentangled category-aware representations $\mathbf{G}$, which is formulated as:

$$\mathbf{W} = \mathbf{U}\mathbf{G} \tag{10}$$

where $\mathbf{G} \in \mathbb{R}^{D \times N}$ with N source class-specific embeddings. Different from SoP (Zhao et al., 2018) using visual cues, the proposed SGN enables learning disentangled semantics from audio itself to guide sound source separation from mixtures.

3 Experiments

3.1 Experimental Setup

Datasets. MUSIC[2] (Zhao et al., 2018) contains 448 untrimmed YouTube music videos of solos and duets from 11 instrument categories. 358 solo videos are applied for training, and 90 solo videos for evaluation. FUSS (Wisdom et al., 2021) is a universal sound dataset with 10 second clips from FSD50K (Fonseca et al., 2017) with labels from the AudioSet Ontology, which includes between 1

[2]Since many videos are no longer publicly available, the used dataset is smaller than the original MUSIC dataset. For a fair comparison, we trained all models on the same training data.

Table 2: Quantitative results of audio-only universal sound separation on FUSS dataset.

Method	Input	SDR	SIR	SAR
RPCA (Huang et al., 2012)	audio-only	-1.16	0.28	1.24
NMF (Virtanen, 2007)	audio-only	-0.49	0.56	2.80
SoP (Zhao et al., 2018) (w/o visual)	audio-only	1.65	2.98	3.25
Wave-U-Net (Stoller et al., 2018)	audio-only	2.36	5.12	4.89
ResUNetDecouple+ (Kong et al., 2021)	audio-only	2.57	5.63	5.38
TDCN++ (Wisdom et al., 2021)	audio-only	3.93	7.21	7.66
SGN (ours)	audio-only	**4.56**	**8.02**	**8.47**

and 4 sound sources. The number of available catogories is 286. We use 20000 mixture clips for training, 1000 mixture clips for validation, and 1000 mixture clips for testing.

Evaluation Metrics. Following previous work (Zhao et al., 2018; Stoller et al., 2018; Kong et al., 2021), we use Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR) to evaluate the separation performance. The open-source mir_eval (Raffel et al., 2014) library is utilized for computing the results to report.

Implementation. We follow the prior work and the audio signal is sub-sampled to 11kHz. An STFT with a window size of 1022 and a hop length of 256 is further applied on to generate a 512 × 256 Time-Frequency representation of the audio, which is resampled to a log-frequency scale with size of 256 × 256 as the input spectrogram. A SGD optimizer with momentum 0.9 is used to train the model. The learning rate is 0.001. We train the model with a batch size of 80 for 100 epochs. The depth of self-attention transformers is 6, and the dimension size D is 256. The kernel and stride size of the convolution patch embedding layer is 16 and 16. The total number of patches P is 256. In experiments, the number of sources N for training is 2.

3.2 Comparison to Prior Work

In this work, we propose a novel and effective framework for audio source separation. In order to validate the effectiveness of the proposed SGN, we fully compare it to previous audio-only and audio-visual baselines: 1) NMF (Virtanen, 2007): a traditional signal processing baseline with non-negative matrix factorization to separate each testing source spectrogram directly; 2) RPCA (Huang et al., 2012): a machine learning method without parameters to separate each source via robust principal component analysis; 3) SoP (w/o visual) (Zhao et al., 2018): an audio-visual baseline without the visual network to do separation by applying U-Net on audio only; 4) WAVE U-NET (Stoller et al., 2018): an audio-only approach with multi-scale feature maps from one-dimensional waveform; 5) SoP (Zhao et al., 2018): the first audio-visual model on MUSIC dataset by leveraging pixel-level visual features to match the spectral components to recover magnitude and phase of input spectrogram; 6) MP-Net (Xu et al., 2019): an audio-visual baseline with recursive separation from the mixture; 7) ResUNetDecouple+ (Kong et al., 2021): a recent method by applying a residual U-Net on spectrogram to decouple the estimation of complex ideal ratio mask into phase and magnitude; 8) TDCN++ (Wisdom et al., 2021): a typical audio-only baseline for universal sound source separation on FUSS benchmark.

For music-only sound source, we report the quantitative comparisons in Table 1. As can be seen, we achieve the best performance in terms of all metrics compared to previous both audio-only and audio-visual baselines. In particular, the proposed SGN significantly outperforms ResUNetDecouple+ (Kong et al., 2021), the current state-of-the-art audio-only model, where we achieve the performance gains of 1.22 SDR, 3.64 SIR, and 3.76 SAR. Without addition visual cues, the performance of SoP (Zhao et al., 2018) drops a lot, which implies the importance of extracting high-level semantics from visual inputs as the guidance for audio-visual approaches. Meanwhile, our SGN achieves comparable even better results against those audio-visual baselines. These improvements demonstrate the effectiveness of our method in disentangling individual semantics from the audio itself.

In addition, significant gains on universal audio source separation can be observed in Table 2. Compared to ResUNetDecouple+ (Kong et al., 2021) with decoupling estimations of phase and magnitude, our SGN achieves the results gains of 1.99 SDR, 2.39 SIR, and 3.09 SAR. Furthermore, the

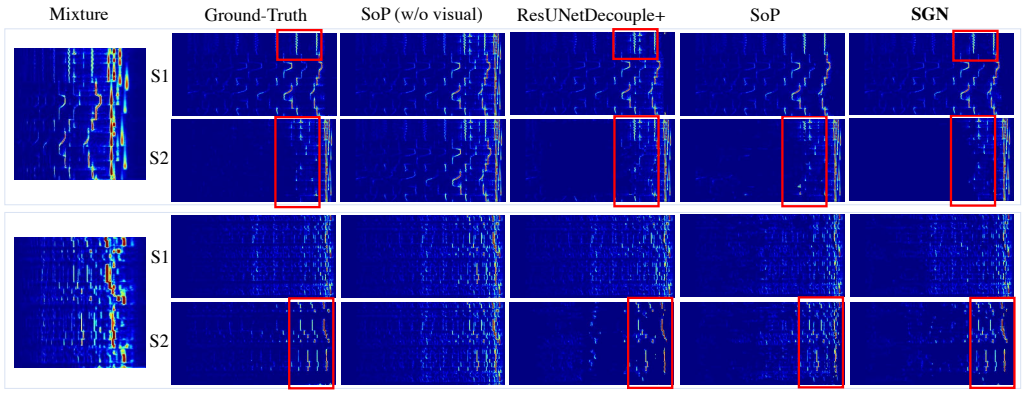

Figure 2: Qualitative comparisons with audio-visual and audio-only baselines (Zhao et al., 2018; Kong et al., 2021). The proposed SGN achieves much better separation performance in terms of the quality of reconstructed source spectrograms. To better illustrate the superiority, some regions are highlighted by red boxes.

Table 3: Ablation studies on Source Class Tokens (SCT) and Category-aware Grouping (CAG).

SCT	CAG	SDR	SIR	SAR
✗	✗	1.51	5.92	6.21
✓	✗	4.57	8.36	9.55
✗	✓	3.32	7.73	8.56
✓	✓	**5.20**	**10.81**	**10.67**

proposed approach outperforms the TDCN++ baseline (Wisdom et al., 2021) by 0.63 SDR, 0.81 SIR, and 0.81 SAR. These results validate the superiority of our method in sound separation.

In order to qualitatively evaluate reconstructed source spectrograms, we compare the proposed SGN with audio-visual and audio-only baselines (Zhao et al., 2018; Kong et al., 2021) in Figure 2. From comparisons, three main observations can be derived: 1) removing visual cues from SoP (Zhao et al., 2018) makes it hard to do separation from the sound itself, where two recovered source spectrograms are similar among each other; 2) the quality of spectrograms generated by our method is much better than the audio-only baseline (Kong et al., 2021); 3) the proposed SGN achieves competitive even better results on reconstructed spectrograms against the audio-visual approach (Zhao et al., 2018) by leveraging visual cues. These visualizations further showcase the advantage of our simple SGN in directly learning category-aware representations from the sound itself to guide source separation.

3.3 Experiment Analysis

In this section, we performed ablation studies on the demonstrate the benefit of introducing Source Class Tokens and Category-aware Grouping module. We also conducted extensive experiments to explore flexible number of audio sources separation and learned sound representations.

Source Class Tokens & Category-aware Grouping. In order to validate the effectiveness of the introduced source class tokens (SCT) and category-aware grouping (CAG), we ablate the necessity of each module and report the quantitative results in Table 3. We can observe that adding learnable SCT highly improves the vanilla baseline by 3.06 SDR, 2.44 SIR, and 3.34 SAR, which shows the benefit of class tokens in guiding source separation. Meanwhile, introducing only CAG in the baseline also increases the separation results. More importantly, incorporating SCT and CAG together into the baseline significantly raises the performance by 3.69 SDR, 4.89 SIR, and 4.46 SAR. These results demonstrate the importance of source class tokens and category-aware grouping on extracting disentangled semantics from the audio itself for source separation.

Generalizing to Flexible Number of Sources. In order to demonstrate the generalizability of the proposed SGN to flexible number of sources, we directly transfer the model without additional training to inference a mixture of 3 sources. We achieve satisfactory performance of 3.18 SDR, 6.12 SIR, 5.98 SAR, and still outperform SoP (Zhao et al., 2018) (2.57 SDR, 5.56 SIR, 5.63 SAR). These

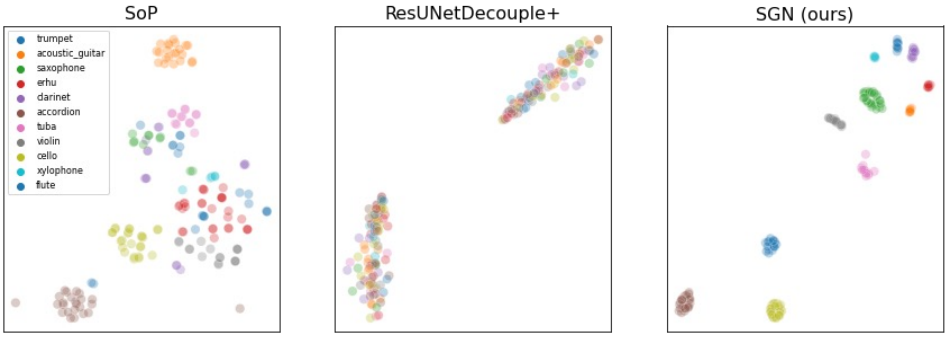

Figure 3: Qualitative comparisons of representations learned by SoP, ResUNetDecouple+, and the proposed SGN. Note that each spot denotes the feature of one source sound, and each color refers to one category, such as "acoustic_guitar" in yellow and "erhu" in red.

results indicate that our SGN can support to separate a flexible number of sources, which differs from existing audio-only separation methods (Stoller et al., 2018; Kong et al., 2021).

Learned Disentangled Representations. Learning disentangled semantic representations is essential for us to separate the sound source from a mixture. To better evaluate the quality of learned category-aware features, we visualize the learned sound representations of 11 categories in MUSIC by t-SNE (van der Maaten & Hinton, 2008), as shown in Figure 3. It is noted that each color represents one category of source sound, such as "acoustic guitar" in yellow and "erhu" in red. As can be seen in the last column, features extracted by the proposed SGN are both intra-class compact and inter-class separable. In contrast to our disentangled embeddings in the semantic space, ResUNetDecouple+ (Kong et al., 2021) with audio as input failed to learn category-aware features, but instead learned two separate representations for magnitude and phase estimations. With the benefit of visual cues, SoP (Zhao et al., 2018) can extract separate features on some classes, *e.g.*, "accoustic_guitar" in yellow and "accordion" in khaki. However, SoP features from most classes are not separable as they did not use explicit category-aware grouping mechanism proposed in our SGN. These meaningful visualization results further demonstrate that our SGN successfully extracts disentangled and compact representations for audio source separation.

3.4 Limitation

Although the proposed SGN achieves superior results on music and universal sound source separation, the gains of our approach over audio-visual models are not significant. One possible solution is to incorporate the visual modality with audio features together for multi-modal grouping. Meanwhile, we notice that if we transfer our model directly without additional training, it would be hard to separate all unseen classes as we do not have learned unseen class tokens to guide source separation. The primary cause is that we need to pre-define a set of category types for training. Therefore, the future work is potentially to learn enough large number of sound source tokens or to explore continual learning when it comes to new classes of audio.

4 Related Work

Audio Representation Learning. Audio representation learning has been addressed in many previous works (Aytar et al., 2016; Owens et al., 2016; Arandjelovic & Zisserman, 2017; Korbar et al., 2018; Senocak et al., 2018; Zhao et al., 2018; 2019; Gan et al., 2020b; Morgado et al., 2020; 2021a;b) to learn discriminative representations of waveform and spectrogram from audios. Such learnable features are beneficial for many audio-relevant tasks, such as audio-event localization (Tian et al., 2018; Lin et al., 2019; Wu et al., 2019; Lin & Wang, 2020), audio-visual parsing (Tian et al., 2020; Wu & Yang, 2021; Lin et al., 2021), audio-visual spatialization (Morgado et al., 2018; Gao & Grauman, 2019; Chen et al., 2020; Morgado et al., 2020), and sound source localization (Senocak et al., 2018; Hu et al., 2019; Afouras et al., 2020; Qian et al., 2020; Chen et al., 2021; Mo & Morgado, 2022a;b). In this work, our main focus is to learn compact audio representations for source separation from sound mixtures, which is more challenging than those tasks listed above.

Audio-Visual Source Separation. Audio-visual source separation aims at separating sound sources from the audio mixture given the image of sources, such as music source separation from a picture of orchestra playing. In the recent years, researchers (Hershey & Casey, 2001; Zhao et al., 2018; Ephrat et al., 2018; Gao et al., 2018; Xu et al., 2019; Tian et al., 2021; Tzinis et al., 2020) have proposed diverse pipelines to learn better visual representations from images. Zhao *et al.* (Zhao et al., 2018) first leveraged visual features to learn the correspondance between the spectral components and pixels for recovering magnitude and phase of input spectrogram. MP-Net (Xu et al., 2019) applied a recursive MinusPlus Net to separate salient sounds from the mixture. Tian *et al.* (Tian et al., 2021) leveraged sounding object visual grounding to separate visual sounds sources in a cyclic co-learning framework. To make full use of visual clues, more modalities are introduced to learn better visual represenations, such as motion in SoM (Zhao et al., 2019), gesture consisting of pose and keypoints in MG (Gan et al., 2020a), and spatio-temporal visual scene graphs in AVSGS (Chatterjee et al., 2021). While the aforementioned audio-visual approaches achieve promising performance on source separation, they are extremely dependent on the ability of visual networks with exhaustive parameters to learn discriminative representations. Without visual features as clues, their performance degrades significantly as observed in our experiments in Sec. 3.2. In this work, we aim to get rid of the dependency of visual sounds separation on visual branches, by introducing learnable source class tokens to learn compact audio representations for separation.

Audio Source Separation. Audio source separation is a challenging problem that separates the individual audio source from a mixture without any visual clues. Early methods applied classical hidden Markov models (Roweis, 2000; Mysore et al., 2010), Non-negative Matrix Factorization (Smaragdis & Brown, 2003; Virtanen, 2007; Cichocki et al., 2009) and Robust Principal Component Analysis (Huang et al., 2012) to separate the source component directly from the mixture. With the development of deep neural networks, U-Net (Ulyanov et al., 2018) with learnable parameters was introduced to learn the spectrogram representations of audio mixtures. Wave U-Net (Stoller et al., 2018) proposed to learn multi-scale feature maps from the one-dimensional time domain of waveform. Recently, ResUNetDecouple+ (Kong et al., 2021) used a residual U-Net to decouple the estimation of complex ideal ratio masks into magnitude and phase estimations for music sources. Beyond music source separation, MixIT (Wisdom et al., 2020) adapted permutation invariant training (Yu et al., 2017) to do separation on simulated mixtures in the large-scale YFCC100M (Thomee et al., 2016) videos. More recently, a task of single channel distance-based sound separation (Patterson et al., 2022) was proposed to separate near sounds from far sounds in synthetic reverberant mixtures. Different from audio source separation baselines, we develop a fully novel framework to aggregate compact category-wise audio source representations with explicit learnable source class tokens. To the best of our knowledge, we are the first to leverage explicit grouping mechanism for audio source separation. In addition, we do not need the unsupervised learning on large-scale simulated audio data with expensive training costs. Our experiments in Sec. 3.2 also demonstrate the effectiveness of SGN in source separation on both music-only and universal sounds.

5 Conclusion

In this work, we propose SGN, a novel Semantic Grouping Network that directly disentangles audio representations and extracts high-level semantic information for each sound source from input mixtures. We introduce learnable class tokens of sounds to aggregate category-wise source features. Then, we use the aggregated semantic features as the guidance to separate the corresponding audio sources from the mixture. Different from previous sound separation approaches, the proposed SGN can support separation from a flexible number of sound sources and is capable of generalizing to tackle sources from different audio domains. Empirical experiments on both music-only and universal sound separation benchmarks demonstrate the significant advantage of our SGN against audio-only methods and audio-visual models without additional visual cues involved.

Broader Impact. The proposed approach separates the mixture sounds from manually-collected music and universal datasets, which might cause the model to learn internal biases in the data. For instance, the model could fail to separate rare but crucial sound sources. Therefore, these issues should be addressed for the deployment of real applications.

REPRODUCIBILITY

The proposed SGN is simple to implement, and results in experiments are reproducible. We will open release the code and pre-trained models of our implementation.

REFERENCES

Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 208–224, 2020. 8

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 609–617, 2017. 8

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 8

Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1204–1213, 2021. 9

Changan Chen, Unnat Jain, Carl Schissler, S. V. A. Garí, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 17–36, 2020. 8

Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16867–16876, 2021. 8

Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009. 1, 9

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 2, 9

Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. Freesound datasets: a platform for the creation of open audio datasets. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 486–493, 2017. 5

Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020a. 9

Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10478–10487, 2020b. 2, 8

Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 324–333, 2019. 8

Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 35–53, 2018. 2, 9

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. 2

John Hershey and Michael Casey. Audio-visual sound separation via hidden markov models. *Advances in Neural Information Processing Systems*, 14, 2001. 2, 9

Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9248–9257, 2019. 8

Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 57–60, 2012. 1, 5, 6, 9

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017. 15

Qiuqiang Kong, Yin Cao, Haohe Liu, Keunwoo Choi, and Yuxuan Wang. Decoupling magnitude and phase estimation with deep resunet for music source separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2021. 1, 5, 6, 7, 8, 9

Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 8

Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 8

Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2002–2006, 2019. 8

Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 8

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017. 15

Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*, 2022a. 8

Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022b. 8

Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360°video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 8

Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4733–4744, 2020. 8

Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12934–12945, 2021a. 8

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12475–12486, June 2021b. 8

Gautham J Mysore, Paris Smaragdis, and Bhiksha Raj. Non-negative hidden markov modeling of audio with application to source separation. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 140–148. Springer, 2010. 1, 9

Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–816, 2016. 8

Katharine Patterson, Kevin Wilson, Scott Wisdom, and John R. Hershey. Distance-based sound separation. In *Proceedings of Interspeech*, 2022. 9

Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 292–308, 2020. 8

Colin Raffel, Brian Mcfee, Eric Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel Ellis. mir_eval: A transparent implementation of common mir metrics. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014. 6

Sam Roweis. One microphone source separation. *Advances in neural information processing systems*, 13, 2000. 1, 9

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4358–4366, 2018. 8

P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177–180, 2003. 1, 9

Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 334–340, 2018. 1, 5, 6, 8, 9

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, 2016. 1, 9

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018. 8

Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 436–454, 2020. 8

Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2745–2754, 2021. 2, 9

Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*, 2020. 9

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3, 9

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 8

Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007. 1, 5, 6, 9

Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron Weiss, Kevin Wilson, and John Hershey. Unsupervised sound separation using mixture invariant training. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 3846–3857, 2020. 1, 9

Scott Wisdom, Hakan Erdogan, Daniel P. W. Ellis, Romain Serizel, Nicolas Turpault, Eduardo Fonseca, Justin Salamon, Prem Seetharaman, and John R. Hershey. What's all the fuss about free universal sound separation data? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 186–190, 2021. 1, 2, 5, 6, 7

Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1326–1335, 2021. 8

Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 6291–6299, 2019. 8

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. *arXiv preprint arXiv:2202.11094*, 2022. 4, 14

Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 5, 6, 9

Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, 2017. 1, 9

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 570–586, 2018. 2, 3, 5, 6, 7, 8, 9, 14

Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1735–1744, 2019. 8, 9

Table 4: Exploration studies on the depth of transformer layers and grouping strategies in Category-aware Grouping (CAG) module.

Depth	CAG	SDR	SIR	SAR
1	Softmax	4.32	8.18	8.07
3	Softmax	5.14	10.80	10.51
6	Softmax	**5.20**	**10.81**	**10.67**
12	Softmax	5.18	10.65	10.59
6	Hard-Softmax	3.25	6.53	6.36

A Appendix

In this appendix, we provide the detailed network structures, differences from between our SGN and the recent work, GroupViT (Xu et al., 2022), more experiments on the depth of transformer layers and grouping strategies. In addition, we validate the effectiveness of learnable class tokens in learning disentangled embeddings and report qualitative visualization results of separated spectrograms.

A.1 Network Structures

In this section, we report detailed network structures and parameters used in the proposed SGN. First, we use a Convolution2D layer to extract raw features from the input mixture spectrogram with a shape of 256×256, where the kernel size is 16 and stride size is 16. The dimension size of embeddings is 256. Then we adopt 6 self-attention transformer layers to extract audio representations 256×256 and class embeddings 11×256. In Category-aware Grouping module, we utilize class embeddings to aggregate new category-aware representations 11×256. Meanwhile, the U-Net architecture with 7 downsampling blocks and 7 upsampling blocks in SoP (Zhao et al., 2018) is applied to extract audio mixture features 65536×256. In the end, we leverage a light reconstruction network composed of inner product on new category-aware embeddings and audio mixture features to generate category-aware spectrograms 11×65536. For each source reconstruction, we keep the corresponding class index from 11×65536 to use the output 65536 as the final spectrogram (reshaped as 256×256).

A.2 Different from GroupViT and our SGN

When compared to GroupViT (Xu et al., 2022) on image segmentation, there are three main distinct characteristics of our SGN for addressing audio source separation, which are highlighted as follows:

1) **Disentangled Constraint on Class Tokens**. The most significant difference is that we have learned disentangled class tokens for each sound source, *e.g.*, 11 tokens for 11 categories in MUSIC dataset. With the disentangled constraint, each class token does not have overlapping information to learn during training, where we apply a cross-entropy loss $\sum_{i=1}^{C} \text{CE}(\mathbf{h}_i, \mathbf{e}_i)$, with an one-hot encoding target $\mathbf{h}_i$ to constraint each category probability $\mathbf{e}_i$. In contrast, the number of group tokens used in GroupViT is indeed a hyper-parameter and they do not apply any constraint on them.

2) **Category-aware Grouping**. We introduce the category-aware grouping module for extracting individual semantics from audio spectrogram in the original mixed space. However, GroupViT utilized the grouping mechanism on visual patches without class-aware tokens involved. Thus, GroupViT can not be directly applied on sound spectrogram for solving source separation. In addition, they leveraged multiple grouping stages during training and the number of grouping stage is a hyper-parameter. In our case, only one category-aware stage with meaningful class tokens is enough to learn distinguished representations in the semantic space.

3) **Class as Weak Supervision**. We apply the source class as weak supervision to address sound separation problem, but GroupViT used a contrastive loss to match the global visual representations with text embeddings. Therefore, GroupViT required a large batch size for self-supervised training on large-scale data. In this work, we do not need the unsupervised learning on large-scale simulated sound data with expensive training costs.

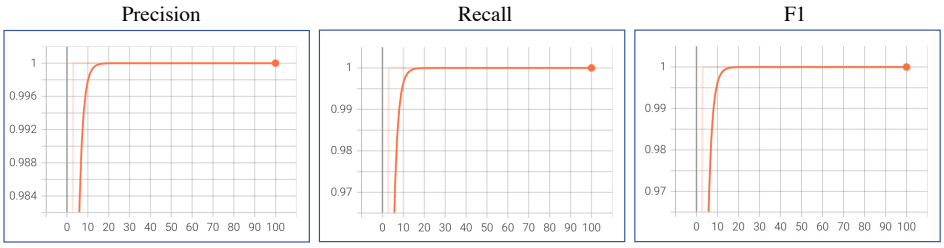

Figure 4: Quantitative results (Precision, Recall, and F1 score) of learned source class tokens.

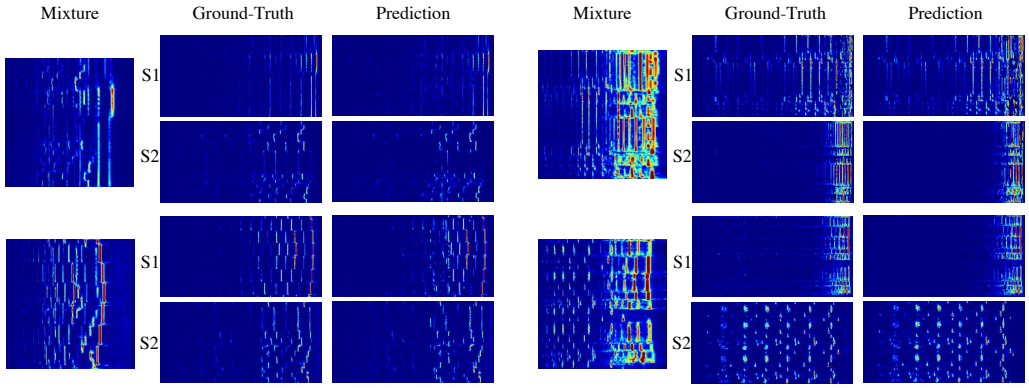

Figure 5: Visualization results of separated source spectrograms.

A.3 Depth of Transformer Layers & Grouping Strategy

The depth of transformer layers and grouping strategies in CAG affect the extracted and grouped representations for audio source separation. In order to explore such effect more comprehensively, we varied the depth of transformer layers from $\{1, 3, 6, 12\}$ and ablated the grouping strategy using Hard-Softmax. To make Hard-Softmax differentiable during training, we applied the Gumbel-Softmax (Jang et al., 2017; Maddison et al., 2017) as the alternative.

The comparison results of the separation performance are shown in Table 4. When the depth of transformer layers is 6 and using Softmax in CAG, the proposed SGN achieves the best results in terms of all metrics. With the increase of the depth, we achieve consistently raising performance as we extract better audio representations from the mixture. However, increasing the depth to 12 will not continually improve the result since 6 transformer layers might be enough to extract the learned embedding for category-aware grouping. In addition, replacing Softmax with Hard-Softmax significantly deteriorates the separation performance of our approach, which implies the importance of the proposed CAG in extracting disentangled semantics for separation.

A.4 Quantitative Validation on Class Tokens

Learnable Class Tokens are essential to aggregate category-aware representations from the sound mixture. In order to quantitatively validate the rationality of learned class token embeddings, we calculate the Precision, Recall, and F1 score of classification using these embeddings across training iterations. We report the quantitative results in Figure 4. It can be seen that all metrics raise to 1 at epoch 20, which implies the learned class tokens have category-aware semantics. These results also demonstrate the effectiveness of class tokens in category-aware grouping for extracting disentangled representations from audio mixtures.

A.5 QUALITATIVE VISUALIZATION ON SEPARATION

In order to qualitatively demonstrate the effectiveness of our method, we report more visualization results in Figure 5. We can observe that the proposed SGN achieves decent separation performance in terms of reconstructing the spectrogram for each sound source.