Independency Adversarial Learning for Cross-Modal Sound Separation

Zhenkai Lin[1, 2], **Yanli Ji**[1, 2*], **Yang Yang**[1, 3]

[1]School of Computer Science and Engineering, UESTC, China
[2]Shenzhen Institute for Advanced Study, UESTC, China
[3]Institute of Electronic and Information Engineering of UESTC in Guangdong, China

Abstract

The sound mixture separation is still challenging due to heavy sound overlapping and disturbance from noise. Unsupervised separation would significantly increase the difficulty. As sound overlapping always hinders accurate sound separation, we propose an Independency Adversarial Learning based Cross-Modal Sound Separation (IAL-CMS) approach, where IAL employs adversarial learning to minimize the correlation of separated sound elements, exploring high sound independence; CMS performs cross-modal sound separation, incorporating audio-visual consistent feature learning and interactive cross-attention learning to emphasize the semantic consistency among cross-modal features. Both audio-visual consistency and audio consistency are kept to guarantee accurate separation. The consistency and sound independence ensure the decomposition of overlapping mixtures into unrelated and distinguishable sound elements. The proposed approach is evaluated on MUSIC, VGGSound, and AudioSet. Extensive experiments certify that our approach outperforms existing approaches in supervised and unsupervised scenarios.

Introduction

Sound mixture separation is a challenging task due to significant sound overlap. In recent years, researchers have observed that humans use multi-modal perception to understand complex activities and endeavor to mimic this skill in sound mixture separation. Specifically, visual information is adopted to guide the sound separation (Zhao et al. 2018; Tzinis et al. 2022; Tian, Hu, and Xu 2021; Zhou et al. 2023), resulting in improved separation quality. Most existing frameworks require ground truths of single-source sounds to support separation model training in a supervised setting (Zhao et al. 2018), but the limitation is that single-source sounds are not always available. The discussion on unsupervised sound mixture separation is much more challenging but also very attractive.

We put our attention on the task of unsupervised sound separation where clean single-source sounds are unknown. This presents a challenge: separating desired sounds from heavily overlapping mixtures without ground truths of target sounds. How to define suitable restraint conditions to accurately extract individual sounds? The Co-Separation framework (Gao and Grauman 2019; Chatterjee et al. 2021a; Ma et al. 2021) employs category information to assist in separating sound elements from mixtures by referencing other single-source audios that belong to the same class. However, the category can not provide sufficient information, being unable to guarantee a complete and clear separation.

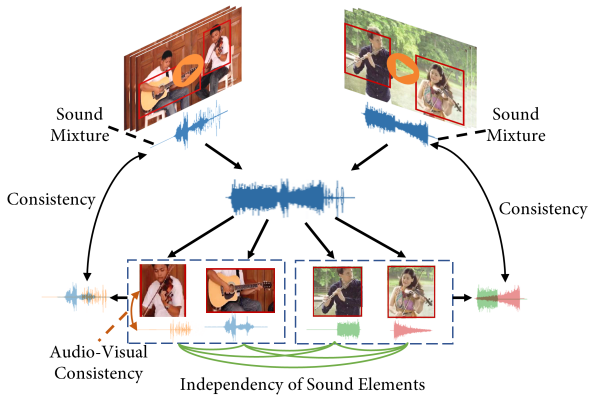


Figure 1: Motivation of our proposed approach. To enable unsupervised sound separation, we keep cross-modal semantic consistency and sound mixture consistency and design our cross-modal sound separation (CMS) solution. To reduce the entanglement of separated sound elements, Independency Adversarial Learning (IAL) is proposed to minimize the correlation between them for obtaining accurate separation.

To address the issue, as illustrated in Figure 1, we design an unsupervised sound mixture separation architecture that makes sufficient use of the intra-modal and inter-modal consistency as restraint conditions for model training. Instead of manually mixing two single-source sounds as mixture sources (Zhao et al. 2018), we directly mix two mixtures that contain two or more sound sources and use the mixed results as separation inputs. We guide the accurate separation process by maintaining two types of consistency: audio-visual cross-modal semantic consistency between visual objects and separated sound elements, as well as audio-audio

*Corresponding author. yanliji@uestc.edu.cn

consistency between original sound mixtures and combinations of separated elements. We design a cross-modal sound separation (CMS) solution to perform separation with the aforementioned consistency. This approach involves several key steps: first, separating sound elements and aligning their combination to the original sound mixture; second, matching sound elements with visual objects; and third, performing dense cross-modal attention learning. Consistency learning could add semantic-related information to sound signals, thus assisting in clearly distinguishing sound elements. Even though, due to the absence of ground truth single-source sounds, the separated results are still partially entangled after separation.

To tackle the problem, inspired by the Independent Component Analysis (ICA) (Comon 1994) which attempts to decompose a multivariate signal into independent non-Gaussian signals in an unsupervised term, we delve deeper into the analysis of the relationship among separated sound elements. In most existing approaches (Ma et al. 2021; Zhou et al. 2022), either the supervised information of single-source sound or the guiding information of visual modality, all ignore the inner relationship among sound elements. Therefore, we propose Independency Adversarial Learning (IAL), a method that leverages adversarial separation to scrutinize the inherent relationships of sounds and enhance sound independence, ultimately reducing entanglement during separation.

An effective way to maximize the independence of separated sound elements is to minimize the mutual information shared among separated sound elements. Based on information theory, we formulate the mutual information calculation of separated sound elements by two steps: first, defining a separated element pair (x_1, x_2), the joint probability distribution $\mathbb{P}(x_1, x_2) = P(x_1, x_2)$ and multiplication of marginal probability distributions $\mathbb{Q}(x_1, x_2) = P(x_1)P(x_2)$; second, calculating the Kullback–Leibler divergence (Csiszar 1975) between $\mathbb{P}$ and $\mathbb{Q}$, denoted as $I(x_1, x_2) = D_{KL}(\mathbb{P}(x_1, x_2)||\mathbb{Q}(x_1, x_2))$. However, how to calculate the divergence of distributions when the distribution forms are unknown? Learning from Generative Adversarial Network (GAN), discriminators could estimate distribution divergence between fake images and real images , for instance, GAN (Goodfellow et al. 2014) estimating Jensen–Shannon divergence, f-GAN (Nowozin, Cseke, and Tomioka 2016) estimating F-divergence and WGAN (Arjovsky, Chintala, and Bottou 2017) estimating Wasserstein distance. In IAL, we seek support from adversarial learning and design a discriminator to estimate the divergence between $\mathbb{P}$ and $\mathbb{Q}$ for mutual information estimation. Then we minimize the mutual information shared among separated sound elements to enhance the independence among them.

In this paper, we propose an Independency Adversarial Learning based Cross-Modal Sound Separation (IAL-CMS) approach. The major contributions are summarized as follows:

- We propose the IAL, which builds adversarial learning on the mutual information among separated sound elements to decompose overlap and enhance the independence of sound elements.
- We design the CMS, where audio-visual consistent feature learning and interactive cross-attention learning are presented to emphasize the semantic consistency for accurate separation.
- Expensive experiments are performed on three datasets, MUSIC, VGGSound, and AudioSet, achieving outperform results in both supervised and unsupervised separation, which certifies the effectiveness of our proposed approach.

Related Work

Visual Sound Separation. In recent years, visual information has been employed to guide sound source separation which significantly improves the separation quality. Some methods (Zhao et al. 2018; Zhou et al. 2022) adopted single-source audios as ground truths and performed Mix-and-Separate training paradigm. Besides utilizing ground truth directly, semantic consistency between the separated audio elements and the ground truths was employed to separate mixtures, $e.g.$, Co-Separation, SeCo (Gao and Grauman 2019; Chatterjee et al. 2021b; Ma et al. 2021; Zhou et al. 2023; Saijo and Ogawa 2022). Current methods seek audio-visual consistency between separated sound elements and visual frames (Zhu and Rahtu 2022; Zhou et al. 2023). However, existing solutions still require a portion of single-source videos for training and suffer difficulty in decomposing entanglements in unsupervised separation scenarios. To step further, our work focuses on unsupervised sound separation with solely multi-source videos.

Blind Sources Separation. Unsupervised blind sources separation separate sound mixtures where single-source ground truths are unavailable (Wisdom et al. 2020a,b, 2021; Saijo and Ogawa 2022). MixIT (Wisdom et al. 2020a,b, 2021) introduced a "Mixture of Mixture" paradigm for unsupervised blind separation and improved separation quality by reducing linear correlation and encouraging source sparsity. Adversarial unmix-and-remix (Yedid 2019) learned remix consistency by adversarial learning for unsupervised blind separation of image mixtures. However, it is still difficult to reduce the non-linear overlap, like reverberation, within multi-source audios. Inspired by that, we propose an adversarial learning solution to enhance the independence of separated elements to solve the overlap problem.

Proposed IAL-CMS Approach

The IAL-CMS Approach performs cross-modal sound separation through Independency Adversarial Learning (IAL). The pipeline of the proposed IAL-CMS is in Figure 2.

Preliminary

Given a pair of videos, V_1 and V_2, each contains a sound mixture of multiple sound sources, corresponding to A_1 and A_2. We mix A_1 and A_2 to be one sound mixture, $A_{mix} = A_1 + A_2$. The mixture is input into our proposed separation model for obtaining single-source sounds. To perform cross-modal separation, we detect visual objects in V_1 and V_2, and the set of detected objects is noted as $\{O_n, n = 1, ..., N\}$. Each object O_n corresponds to one separated sound element

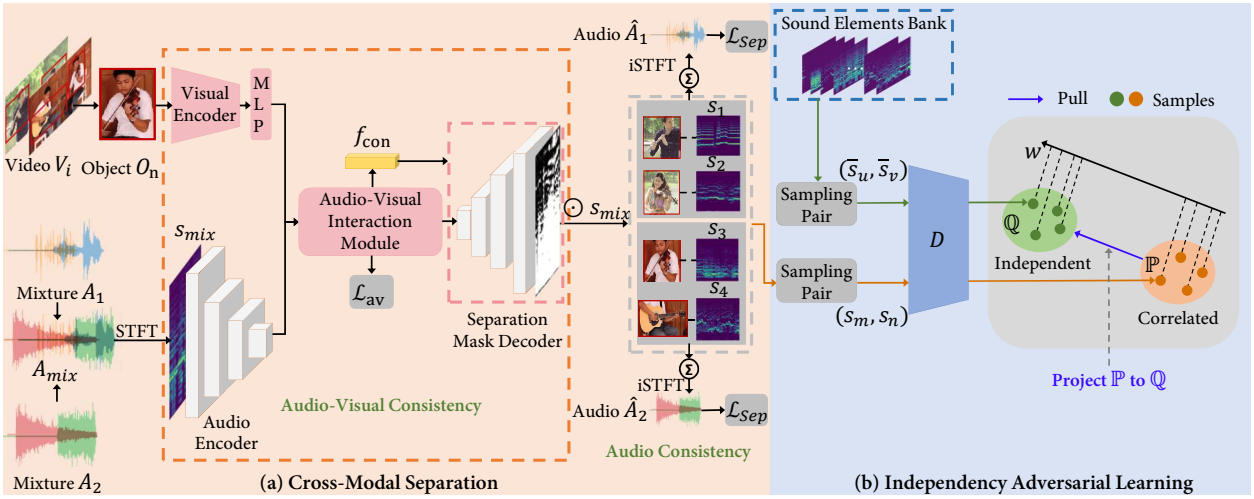


Figure 2: The pipeline of our proposed IAL-CMS approach. It is majorly composed of two parts, (a) the Cross-Modal Separation (CMS), and (b) Independency Adversarial Learning (IAL). The CMS realizes sound mixture separation, where an Audio-Visual Interaction Module employs contrastive learning and attention to enhance semantic consistency for two modalities and fuse semantic-enhanced features. We employ Unet as a decoder to generate separation masks for sound separation. The IAL maximizes the independence of separated sound elements through adversarial training for accurate separation.

a_n, whose spectrum is denoted by s_n. The task of cross-modal sound separation is to extract clean single-source sounds a_n from A_{mix} without the guidance of ground truths.

Independency Adversarial Learning (IAL)

Due to heavy overlapping, separated sound elements remain parts of entanglements. To reduce these entanglements, we propose the IAL to enhance the independence of sound elements which improves the quality of sound separation.

Independence Formulation We measure the independence between separated sound elements by mutual information. It is noted that the less mutual information, the higher the independence. To calculate the mutual information I, we first define a separated sound element set $\{s_n, n = 1, ..., N\}$, and we use function $P()$ to define the joint probability distribution calculation of an element pair (s_m, s_n). So the joint distribution $\mathbb{P}$ of (s_m, s_n) is obtained by Eqn. (1).

$$\mathbb{P}(s_m, s_n) = P(s_m, s_n). \tag{1}$$

where $m, n \in \{1, ..., N\}, m \neq n$, and s_m, s_n are separated from one video.

Since the sound elements s_m, s_n are separated from one source mixture, it is difficult to ensure their independence. There often remain parts of entanglements, thus the (s_m, s_n) cannot be directly used to calculate the marginal probability distributions. To solve the problem, we build a sound element bank that saves separated elements that are obtained from other mixtures. The bank is defined as $\{\bar{s}_u, u = 1, ..., U\}$, and a randomly selected sample pair from the bank is defined as $(\bar{s}_u, \bar{s}_v), u, v \in \{1, ..., U\}, u \neq v$. The elements of the bank have nearly no correlation with currently separated elements (s_m, s_n).

Therefore, the randomly selected sample pairs $(\bar{s}_u, \bar{s}_v)$ should approximate independence as they are independent and typically originate from different mixtures. Then we can use the randomly selected elements $(\bar{s}_u, \bar{s}_v)$ for the marginal probability distribution calculation.

Suppose $P_m(s_u)$ refers to a marginal distribution of s_u, the marginal probability distributions of the selected element pair $(\bar{s}_u, \bar{s}_v)$ is calculated in Eqn. (2).

$$\mathbb{Q}(\bar{s}_u, \bar{s}_v) = P_m(\bar{s}_u)P_m(\bar{s}_v). \tag{2}$$

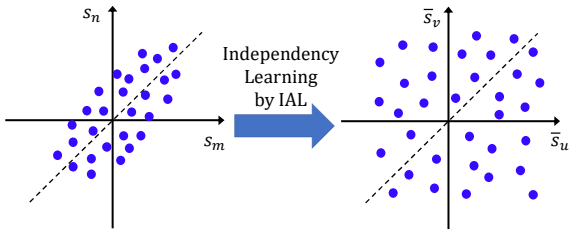
The mutual information $I(s_m, s_n)$ between s_m and s_n can be calculated by Kullback-Leibler (KL) divergence between the joint distribution $\mathbb{P}$ and the marginal probability distributions $\mathbb{Q}$, as shown in Eqn. (3).

$$I(s_m, s_n) = KL(\mathbb{P}(s_m, s_n)||\mathbb{Q}(\bar{s}_u, \bar{s}_v)). \tag{3}$$

We illustrate the point distribution of $\mathbb{P}$ and $\mathbb{Q}$ in Figure 3. The joint distribution of $\mathbb{P}(s_m, s_n)$ are non-ideal distribution, where samples are heavily correlated because they are still partially entangled. In contrast, the distribution of $\mathbb{Q}(\bar{s}_u, \bar{s}_v)$ are sparse, where samples are independent. The IAL aims to minimize the mutual information between (s_m, s_n) by projecting the correlated distribution $\mathbb{P}$ to a sparse distribution, similar to the distribution of $\mathbb{Q}$, as shown in Figure 2 (b).

Because it is difficult to calculate the probability value of $\mathbb{P}$ and $\mathbb{Q}$, we can't directly calculate $KL(\mathbb{P}||\mathbb{Q})$. Inspired by the GAN learning which adopts a discriminator to estimate the divergence between distributions of real and fake images, we learn from the McGAN (Mroueh, Sercu, and Goel 2017) to utilize the integral probability metric for estimating the divergence between distribution $\mathbb{P}$ and distribution $\mathbb{Q}$:

$$I(s_m, s_n) \approx \max_{\|w\|_2 \leq 1} \big\langle w, \mathbb{E}_{(\bar{s}_u, \bar{s}_v)\sim\mathbb{Q}}[f(\bar{s}_u, \bar{s}_v)] - \mathbb{E}_{(s_m, s_n)\sim\mathbb{P}}[f(s_m, s_n)] \big\rangle \tag{4}$$



(a) Distribution $\mathbb{P}$ of correlated elements

(b) Distribution $\mathbb{Q}$ of independent samples

Figure 3: Illustration of distribution $\mathbb{P}$ and $\mathbb{Q}$. (a) presents the joint distribution $\mathbb{P}$ of pair (s_m, s_n), where the separated sound elements s_m and s_n are entangled; (b) presents the distribution $\mathbb{Q}(\bar{s}_u, \bar{s}_v)$, where $\bar{s}_u$ and $\bar{s}_v$ are unrelated and nearly independent. We learn clear separation by projecting distribution $\mathbb{P}$ to $\mathbb{Q}$.

where f is a bounded map that projects spectrum pairs to feature vectors; $\mathbb{E}$ denotes expectation calculation for calculating mean features; $\langle \cdot, \cdot \rangle$ denotes the inner product of vectors; w is a discrimination vector that projects mean features to discrimination values; $\|w\|_2 \leq 1$ denotes the bounded condition of w. Eqn. (4) is adopted to measure the distance between two distributions.

Following that, we design a discriminator D which consists of an encoder E and a binary classifier, where the encoder E is used to model the function of f for feature mapping. The binary classifier determines the encoded feature belonging to the joint distribution or independent distribution. So we reformulate the mutual information estimation in Eqn. (4) as a discrimination network training problem, as illustrated in Eqn. (5) and Eqn. (6).

$$D(s_m, s_n) = \langle w, E(s_m, s_n) \rangle \tag{5}$$

$$\begin{aligned} I(s_m, s_n) \approx \max_D \{ & \mathbb{E}_{(\bar{s}_u, \bar{s}_v) \sim \mathbb{Q}}[D(\bar{s}_u, \bar{s}_v)] \\ & - \mathbb{E}_{(s_m, s_n) \sim \mathbb{P}}[D(s_m, s_n)] \} \end{aligned} \tag{6}$$

In conclusion, we use mutual information $I(s_m, s_n)$ in Eqn. (3) to measure independence and convert the estimation of $I(s_m, s_n)$ into discriminator training by Eqn. (6).

Adversarial Learning For clear sound separation, we train our separator to maximize the independence between separated sound elements, which equals minimizing the mutual information $I(s_m, s_n)$. Based on Eqn. (6), the min-max training is formulated as:

$$\begin{aligned} \min_G \max_D \{ & \mathbb{E}_{(\bar{s}_u, \bar{s}_v) \sim \mathbb{Q}}[D(\bar{s}_u, \bar{s}_v)] \\ & - \mathbb{E}_{(s_m, s_n) \sim \mathbb{P}}[D(s_m, s_n)] \} \end{aligned} \tag{7}$$

where the separated sound elements are generated by the separator G.

In the min-max training, the discriminator maximizes the distribution divergence of sound element pairs, while the separator G minimizes the distance between joint distribution $\mathbb{P}$ and independent distribution $\mathbb{Q}$ by separating more independent sound elements.

Due to the well-known difficulty of min-max adversarial training, we try several adversarial algorithms. Finally, we implement spectral normalization on discriminator D to avoid gradient vanishing problems (Miyato et al. 2018) and satisfy the bounded constraint in Eqn. (4). We design losses $\mathcal{L}_D, \mathcal{L}_G$ with L2 regularization for adversarial training.

$$\begin{aligned} \mathcal{L}_D = & \max(0, 1 - D(\bar{s}_u, \bar{s}_v)) \\ & + \max(0, 1 + D(s_m, s_n)) \end{aligned} \tag{8}$$

$$\mathcal{L}_G = D(\bar{s}_u, \bar{s}_v) - D(s_m, s_n) \tag{9}$$

Instead of directly optimizing Eqn. (6), $\mathcal{L}_D$ is used to train the optimal discriminator that maximizes the distance between $\mathbb{P}$ and $\mathbb{Q}$ (Cortes and Vapnik 1995) for estimating the mutual information $I(s_m, s_n)$ in Eqn. (6); $\mathcal{L}_G$ train the separator to directly minimize the estimated mutual information $I(s_m, s_n)$.

We alternately train the discriminator by $\mathcal{L}_D$ and the separator by $\mathcal{L}_G$ in an adversarial manner. We could learn independent sound elements that share less mutual information to overcome the overlap problem in sound separation.

Cross-Modal Sound Separation

We perform Cross-Modal sound mixture Separation (CMS) relying on audio consistency and audio-visual consistency. As shown in Figure 2 (a), we employ a cross-modal guided sound separation solution to learn separation masks, which are used to extract separated sound elements.

Audio-Visual Features We first prepare audio-visual features for the separation model. We detect visual objects O_n in video V_i, and extract visual feature f_{O_n} by the visual encoder for each detected object. For audios, we obtain mel-spectrum s_{mix} of sound mixture A_{mix} via Short-Time Fourier Transform (STFT) and Mel-spectrum transformation, and fed s_{mix} into the audio encoder to extract feature f_a^{mix}.

Audio-Visual Interaction Module To perform visual-guided sound separation, we design an audio-visual interaction module that performs interactive feature learning to emphasize audio-visual consistency. As illustrated in Figure 4, the module includes two processing strategies, **audio-visual consistent feature learning** and **interactive cross-attention learning**. The module produces attention-emphasized fusion feature f_{con}, which serves as the initial input to the decoder for mask generation. Since feature f_{con} well represents object sounds in the guidance of visual images, we further utilize it to control the global category semantic of separated sound elements.

Audio-Visual Consistent Feature Learning Different from Co-Separation (Gao et al. 2019) which directly fuses audio-visual features and uses them for separation, we design the cross-modal consistent feature learning to perform cross-modal alignment and enhance semantic consistency. We employ contrast learning (Chen et al. 2021) to enhance semantic consistency between the visual feature $f_v = f_{O_n}$ and the audio mixture feature f_a^{mix}. As f_a^{mix} involves multiple-source sounds, we first fuse visual and sound mixture features to emphasize visual-related semantic information and use the fusion feature in contrastive learning.

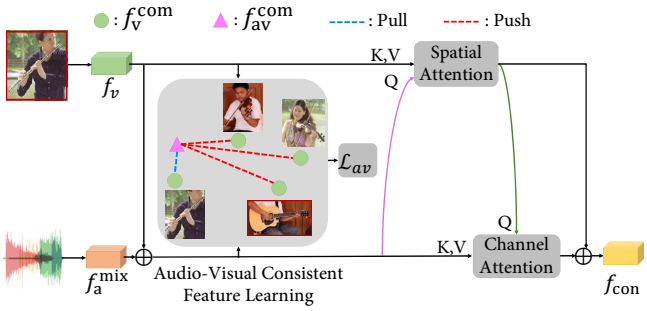


Figure 4: The structure of Audio-Visual Interaction Module. The audio-visual consistent feature learning employs contrastive learning to match corresponding visual objects and sound mixtures, drawing near the feature distributions of f_{av}^{com} and f_v^{com}. Following that, cross-modal interactive attention learning is performed to further enhance the semantic consistency of two-modal features. The attention-enhanced features are finally fused for separation mask generation.

For contrastive learning, we first adopt an encoder to project f_{av} and f_v into a common feature space, as shown in Eqn.10.

$$f_{av}^c = \Phi_{av}(AVP(f_{av})), f_v^c = \Phi_v(f_v) \tag{10}$$

where AVP denotes global average pooling in both the frequency and temporal dimensions; Φ_{av} and Φ_v are projectors involving MLP layers and L2 normalization.

We compose the fusion feature $f_{av,n}^c$ and the visual feature $f_{v,n}^c$ corresponding to the same sounding object as a positive pair, thus other visual features $f_{v,m}^c$ corresponding to other objects are used as negative pairs, $m \in \mathcal{N}$. We employ the infoNCE (van den Oord, Li, and Vinyals 2018) for optimization. The loss is set in Eqn. (11).

$$\mathcal{L}_{av} = -\sum_{n=1}^{N} \log \frac{\exp\left(< f_{av,n}^c, f_{v,n}^c > /\tau\right)}{\sum_{m\in\mathcal{N}} \exp\left(< f_{av,n}^c, f_{v,m}^c > /\tau\right)} \tag{11}$$

where τ denotes the temperature hyperparameter, and $\mathcal{N}$ refers to negative sample space.

Interactive Cross-attention Learning With the fusion features f_{av} and visual feature f_v, we perform interactive cross-modal attention learning to further enhance their cross-modal consistency through spatial and channel attention learning (Xu et al. 2020). The fusion feature f_{av} and visual feature f_v are iteratively used as queries in the attention calculation, extracting attention-emphasized representations with consistent semantics.

Sound Mixture Separation We employ a Unet decoder to generate separation masks, denoted as Mask Generator. The semantic-enhanced feature f_{con} is input to the generator for separation mask generation. The preceding audio-visual consistency learning effectively encodes category semantics in fusion feature f_{con}. This enables us to exert global control over the separated spectrums using f_{con}, ensuring that every patch in separated spectrums would share the same category semantics. To achieve this, in the Mask Generator, we replace BatchNorm with AdaIN (Huang and Belongie 2017; Karras, Laine, and Aila 2019) in the Unet decoder (Ronneberger, Fischer, and Brox 2015), where channel-wise standard deviation and mean of features are controlled according to f_{con}.

Method	2-mix			3-mix		
	SDR↑	SIR↑	SAR↑	SDR↑	SIR↑	SAR↑
SOP (Zhao et al. 2018)	7.30	11.9	11.9	3.65	8.77	8.48
MP-Net (Xu et al. 2019)	7.00	10.39	15.31	5.75	5.37	**13.68**
Co-Separation (Gao et al. 2019)	7.38	13.7	10.8	3.94	8.93	-
FCSN (Ma et al. 2021)	11.74	16.23	-	6.30	**11.47**	-
AVSGS (Chatterjee et al. 2021b)	11.4	17.3	13.5	-	-	-
SepFusion (Zhou et al. 2022)	8.76	16.65	11.81	-	-	-
AMNet (Zhu and Rahtu 2022)	11.08	**18.00**	13.22	-	-	-
iQuery (Chen et al. 2023)	11.17	15.84	14.27	-	-	-
IAL-CMS (Ours)†	10.22	17.08	12.82	6.29	10.73	10.37
IAL-CMS (Ours)	**12.26**	17.93	**15.40**	**6.54**	11.31	10.43

Table 1: Sound mixture separation results on MUSIC. Here, † refers to unsupervised separation.

We obtain the separated mel-spectrums s_j by multiplying the decoded mask M_j with the sound mixture spectrum, $s_j = s_{mix} \odot M_j$. Then the spectrums s_j are transformed back to temporal signals a_j. If the separation does not lose audio information, the audio mixtures before and after the separation should be the same. We reconstruct the sound mixture $\hat{A}_i$ by adding up N_a separated sound elements in video V_i via Eqn. 12. Then we define $\mathcal{L}_{sep}$ (Eqn.13) using the signal-to-noise ratio (SNR) to train our model for keeping audio consistency between the sound mixture A_i and the reconstructed sound mixture $\hat{A}_i$.

$$\hat{A}_i = \sum_{j=1}^{N_a} a_j \tag{12}$$

$$\mathcal{L}_{sep} = -10 \lg \frac{\|A_i\|^2}{\|A_i - \hat{A}_i\|^2 + \tau\|A_i\|^2} \tag{13}$$

Finally, we alternately train the proposed model by the total loss $\mathcal{L}$ defined in Eqn.14 where λ_1, λ_2 refer to weights.

$$\mathcal{L} = \mathcal{L}_{sep} + \lambda_1(\mathcal{L}_G + \mathcal{L}_D) + \lambda_2\mathcal{L}_{av} \tag{14}$$

Method	SDR↑	SIR↑
SOP* (Zhao et al. 2018)	3.67	6.51
MP-Net* (Xu et al. 2019)	2.76	3.43
Co-Separation* (Gao et al. 2019)	3.32	7.15
FCSN (Ma et al. 2021)	6.39	10.14
IAL-CMS (Ours)	**9.96**	**14.74**

Table 2: Result comparison of 2-mix separation on VGGSound-15Instrument. The results noted by * are reproduced by (Ma et al. 2021).

Method	SDR↑	SIR↑	SAR↑
SOP (Zhao et al. 2018)	1.66	3.58	11.5
MIML (Gao et al. 2018)	1.83	-	-
MP-Net* (Xu et al. 2019)	1.79	2.25	-
Co-Separation (Gao et al. 2019)	4.26	7.07	13.0
FCSN (Ma et al. 2021)	4.45	**7.92**	-
IAL-CMS (Ours)	**4.71**	6.70	**14.51**

Table 3: Result comparison of 2-mix separation on AudioSet-15Instrument. The results noted by * are reproduced by (Ma et al. 2021).

Experiments

Datasets and Evaluation Metrics

Datasets We evaluate our proposed approach on three datasets, MUSIC, VGGSound, and AudioSet. **MUSIC** (Zhao et al. 2018) has 714 untrimmed videos of musical solos and duets, spanning 11 instrument categories. MUSIC-Solo is a subset of MUSIC that contains videos involving solo sounds, used for supervised separation evaluation. Synthetic-Duet is an artificially synthesized set by mixing two solos in MUSIC-Solo, used for unsupervised separation evaluation. **VGGSound** (Chen et al. 2020) is a large-scale audio-visual dataset containing over 200k 10s video clips for 310 categories of general classes. **AudioSet** (Gemmeke et al. 2017) contains 10s video clips and audio category annotations. For VGGSound and Audioset, following (Gao and Grauman 2019; Ma et al. 2021), we use the subset involving 15 instruments for evaluation. Thus we use the "15Instrument" suffix in the following.

Evaluation Metrics For evaluation metrics, we use the open-source $mir\_eval\ library$ (Raffel et al. 2014) following (Zhao et al. 2018; Gao and Grauman 2019; Ma et al. 2021) to calculate SDR, SIR, and SAR for separation quality assessing.

Comparison with SOTA Approaches

MUSIC We perform both the supervised and unsupervised (denoted by †) separation, and list results in Table 1. For supervised separation, our approach outperforms all existing methods in 2-mix evaluation under a supervised setting. Our approach improves the SDR by 4.88 compared with the Co-Separation (Gao et al. 2019). Compared with the FSCN (Ma et al. 2021), which performs additional refining operations for fine-grained separation, our approach surpasses SDR by 0.48 and SIR by 1.70. In the 3-mix evaluation, we surpass the FSCN (Ma et al. 2021) in SDR by 0.24, but a little weak in the SIR. It is worth noting that existing methods all use the ground truth of single-source sounds for model training, but our approach has no such requirement. We use sound mixtures for model training. Even using sound mixtures, noted as the unsupervised setting, our approach still obtains encouraging separation results compared with supervised methods, $e.g.$ SepFusion (Zhou et al. 2022) and Co-Separation (Gao et al. 2019). It demonstrates that IAL and CMS are essential and valuable solutions for separation without much supervision.

$\mathcal{L}_{sep}$	$\mathcal{L}_G/\mathcal{L}_D$	$\mathcal{L}_{av}$	SDR↑	SIR↑	SAR↑
✓	×	×	6.93	13.6	9.91
✓	✓	×	8.31	14.18	11.55
✓	×	✓	8.08	13.10	12.14
✓	✓	✓	**10.22**	**17.08**	**12.28**

Table 4: Evaluation of loss functions on MUSIC. The results indicate the valid contribution from $\mathcal{L}_G/\mathcal{L}_D$ (IAL) and the cross-modal consistent learning with $\mathcal{L}_{av}$.

VGGSound-15Instrument Table 2 exhibits the results on VGGSound-15Instrument. Compared with the SOP (Zhao et al. 2018) and MP-Net (Xu, Dai, and Lin 2019), which are trained using clean single-source sounds, fail in the in-the-wild scenario. Compared with the FSCN (Ma et al. 2021), our method outperforms it by 3.57 on SDR and 4.60 on SIR without extra refining operation. The results certify the effectiveness of cross-modal consistency and independence adversarial learning in our approach, which makes better use of multi-source audios and effectively separates sound elements with reduced entanglement in in-the-wild scenarios.

AudioSet-15Instrument Table 3 shows the results on AudioSet-15Instrument which contains severe noise of off-screen sounds. Compared with FCSN (Ma et al. 2021) and Co-Separation (Gao et al. 2019), our approach achieves higher results in SDR and SAR, a little weak in SIR than the FCSN. With the interference of noise, our method still achieves competitive performance without targeted single-source ground truth sounds. It indicates that our approach has better generalization in solving noise-involved sound separation problems.

Ablation Study

Evaluation on Loss Functions We evaluate the contribution from four loss functions for sound separation on MUSIC, and the results are shown in Table 4. Because the losses $\mathcal{L}_G, \mathcal{L}_D$ are always simultaneously adopted for model training, we list them together. The loss $\mathcal{L}_{sep}$ is used to keep audio consistency in the CMS. As we do not use single-source sound as supervision for model training, it is required to use loss $\mathcal{L}_{sep}$ to keep the consistency of mixtures.

Removing $\mathcal{L}_G/\mathcal{L}_D$ equals removing the IAL module in our approach. The results show that the $\mathcal{L}_G/\mathcal{L}_D$ (IAL) con-

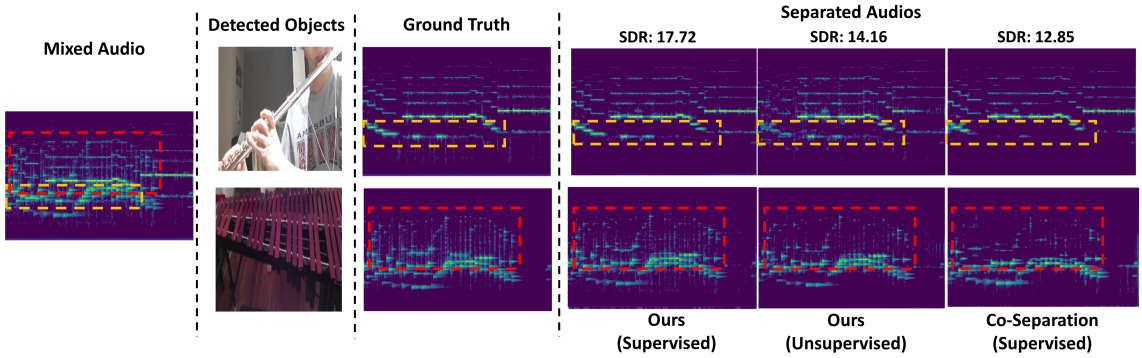


Figure 5: Visualization of separation results on MUSIC Dataset. It exhibits the separation results of our approach under supervised and unsupervised settings, and the results of Co-Separation (Gao et al. 2019). The red and yellow boxes denote the sound disturbance area. As shown, Co-separation fails to separate sound components in the box, while our model successfully separates them more clearly.

Spatial	Channel	AdaIN	SDR↑	SIR↑	SAR↑
×	×	×	11.01	16.16	14.60
✓	×	×	11.33	16.93	14.80
✓	✓	×	11.84	17.03	15.34
✓	✓	✓	**12.26**	**17.93**	**15.40**

Table 5: Evaluation on attention learning, including spatial attention and channel attention, in the audio-visual interaction module and AdaIN in the decoder on MUSIC.

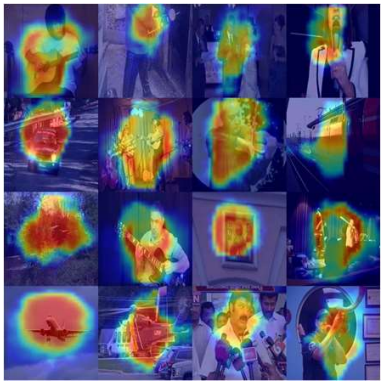


Figure 6: Visualization of learned spatial attention results. Interactive Attention Learning correctly locates objects.

tributes to improving separation performance by 2.14 on SDR, 3.98 on SIR, certifying the effects of IAL. The loss $\mathcal{L}_{av}$ corresponds to cross-modal consistent learning. Without $\mathcal{L}_{av}$, the separation performance drops 1.91 on SDR and 2.9 on SIR. Thus, cross-modal consistency emerges as a crucial factor for cross-modal separation. Enhancing audio-visual consistency indeed contributes to better separation performance.

Evaluation on Interactive Attention Learning Interactive Attention Learning is designed to enhance the semantic representation of features. Here, we evaluate the contribution of three components, spatial attention, channel attention in the audio-visual interaction module, and AdaIN in the decoder. As illustrated in Table 5, without the spatial and channel attention in the audio-visual interaction module, the separation results fall 0.83 on SDR and 0.87 on SIR, respectively. AdaIN in the decoder also assists in improving 0.42 on SDR and 0.9 on SIR. The evaluation certifies that every component contributes to improving separation quality. Figure 6 shows learned attention where crucial object areas are correctly located.

Visualization of Separation Result Comparison

We exhibit the 2-mix separation results of our approach under supervised and unsupervised settings in Figure 5, where we compare our results with the Co-separation (Gao et al. 2019) in the MUSIC dataset. Comparing the separated spectrograms obtained by our approach under the supervised setting and unsupervised setting, we may find that both of them achieve a clear separation with minimal disparity compared to the ground truths. While the results of unsupervised separation are a little worse than our supervised results, they still surpass the results of Co-separation (Gao et al. 2019). The comparison demonstrates that our model, with or without supervised training, can obtain reliable separation performance.

Conclusion

A major challenge in sound mixture separation is the heavy overlap and disturbance. To solve the problem, we propose an Independency Adversarial Learning based Cross-Modal Sound Separation (IAL-CMS) approach. The proposed approach is evaluated for supervised and unsupervised separation on MUSIC, VGGSound, and AudioSet datasets. Extensive experiments certify its effective performance in both supervised and unsupervised sound separation.

Acknowledgments

This work was supported by the Science and Technology Innovation Committee of Shenzhen Municipality Foundation (No.JCYJ20210324132203007).

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning (ICML)*, 214–223.

Chatterjee, M.; Roux, J. L.; Ahuja, N.; and Cherian, A. 2021a. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1204–1213.

Chatterjee, M.; Roux, J. L.; Ahuja, N.; and Cherian, A. 2021b. Visual Scene Graphs for Audio Source Separation. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 1184–1193.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. VGGSound: A Large-scale Audio-Visual Dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Chen, J.; Zhang, R.; Lian, D.; Yang, J.; Zeng, Z.; and Shi, J. 2023. iQuery: Instruments as Queries for Audio-Visual Sound Separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14675–14686.

Chen, Y.; Xian, Y.; Koepke, A. S.; Shan, Y.; and Akata, Z. 2021. Distilling Audio-Visual Knowledge by Compositional Contrastive Learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 7016–7025.

Comon, P. 1994. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314.

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20: 273–297.

Csiszar, I. 1975. I-Divergence Geometry of Probability Distributions and Minimization Problems. *Ann. Probab.*, 3(1): 146–158.

Gao, R.; and Grauman, K. 2019. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 3878–3887.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. *Advances in Neural Information Processing Systems (NeurIPS)*, 27.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1501–1510.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 4401–4410.

Ma, S.; Ji, Y.; Xu, X.; and Zhu, X. 2021. Vision-guided music source separation via a fine-grained cycle-separation network. In *Proceedings of ACM International Conference on Multimedia (ACM MM)*, 4202–4210.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. *International Conference on Learning Representations (ICLR)*.

Mroueh, Y.; Sercu, T.; and Goel, V. 2017. McGan: Mean and covariance feature matching GAN. In *International conference on machine learning (ICML)*, 2527–2535.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-GAN: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems (NeurIPS)*, 29.

Raffel, C.; McFee, B.; Humphrey, E. J.; Salamon, J.; Nieto, O.; Liang, D.; and Ellis, D. P. W. 2014. MIR-EVAL: A transparent implementation of common MIR metrics. In *The International Society for Music Information Retrieval (ISMIR)*.

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)*, 234–241.

Saijo, K.; and Ogawa, T. 2022. Remix-Cycle-Consistent Learning on Adversarially Learned Separator for Accurate and Stable Unsupervised Speech Separation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4373–4377.

Tian, Y.; Hu, D.; and Xu, C. 2021. 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)*, 2745–2754.

Tzinis, E.; Wisdom, S.; Remez, T.; and Hershey, J. R. 2022. Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation. In *European Conference on Computer Vision (ECCV)*, 368–385.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Wisdom, S.; Jansen, A.; Weiss, R. J.; Erdogan, H.; and Hershey, J. R. 2021. Sparse, efficient, and semantic mixture invariant training: Taming in-the-wild unsupervised sound separation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 51–55.

Wisdom, S.; Tzinis, E.; Erdogan, H.; Weiss, R.; Wilson, K.; and Hershey, J. 2020a. Unsupervised sound separation using mixture invariant training. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 3846–3857.

Wisdom, S.; Tzinis, E.; Erdogan, H.; Weiss, R. J.; Wilson, K.; and Hershey, J. R. 2020b. Unsupervised speech separation using mixtures of mixtures. In *ICML Workshop on Self-supervision in Audio and Speech*.

Xu, H.; Zeng, R.; Wu, Q.; Tan, M.; and Gan, C. 2020. Cross-modal relation-aware networks for audio-visual event localization. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 3893–3901.

Xu, X.; Dai, B.; and Lin, D. 2019. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 882–891.

Yedid, H. 2019. Towards unsupervised single-channel blind source separation using adversarial pair unmix-and-remix. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3272–3276.

Zhao, H.; Gan, C.; Rouditchenko, A.; Vondrick, C.; McDermott, J.; and Torralba, A. 2018. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 11205, 587–604.

Zhou, D.; Zhou, X.; Hu, D.; Zhou, H.; Bai, L.; Liu, Z.; and Ouyang, W. 2022. Sepfusion: Finding optimal fusion structures for visual sound separation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 3544–3552.

Zhou, X.; Zhou, D.; Ouyang, W.; Zhou, H.; and Hu, D. 2023. SeCo: Separating Unknown Musical Visual Sounds with Consistency Guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5168–5177.

Zhu, L.; and Rahtu, E. 2022. Visually guided sound source separation and localization using self-supervised motion representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1289–1299.