# Audio-Visual Separation with Hierarchical Fusion and Representation Alignment

Han Hu*
hxh347@student.bham.ac.uk
Dongheng Lin*
d.lin.2@bham.ac.uk
Qiming Huang
qxh366@student.bham.ac.uk
Yuqi Hou
yxh029@student.bham.ac.uk
Hyung Jin Chang
h.j.change@bham.ac.uk
Jianbo Jiao
j.jiao@bham.ac.uk

The MIx Group, School of Computer Science, University of Birmingham
Birmingham, UK

"Alone we can do so little; together we can do so much."

— *Helen Keller*

**Abstract**

Self-supervised audio-visual source separation leverages natural correlations between audio and vision modalities to separate mixed audio signals. In this work, we first systematically analyze the performance of existing multimodal fusion methods for audio-visual separation task, demonstrating that the performance of different fusion strategies is closely linked to the characteristics of the sound—middle fusion is better suited for handling short, transient sounds, while late fusion is more effective for capturing sustained and harmonically rich sounds. We thus propose a hierarchical fusion strategy that effectively integrates both fusion stages. In addition, training can be made easier by incorporating high-quality external audio representations, rather than relying solely on the audio branch to learn them independently. To explore this, we propose a representation alignment approach that aligns the latent features of the audio encoder with embeddings extracted from pre-trained audio models. Extensive experiments on MUSIC, MUSIC-21 and VGGSound datasets demonstrate that our approach achieves state-of-the-art results, surpassing existing methods under the self-supervised setting. We further analyze the impact of representation alignment on audio features, showing that it reduces modality gap between the audio and visual modalities. The project page is at: https://happy-new-bears.github.io/AudioSep-HFRA/.

## 1 Introduction

In our daily life, sounds come in diverse forms—some are quick and sharp, like a bird chirping or a raindrop hitting the ground, while others are smooth and lingering, like the deep

*Equal contribution.

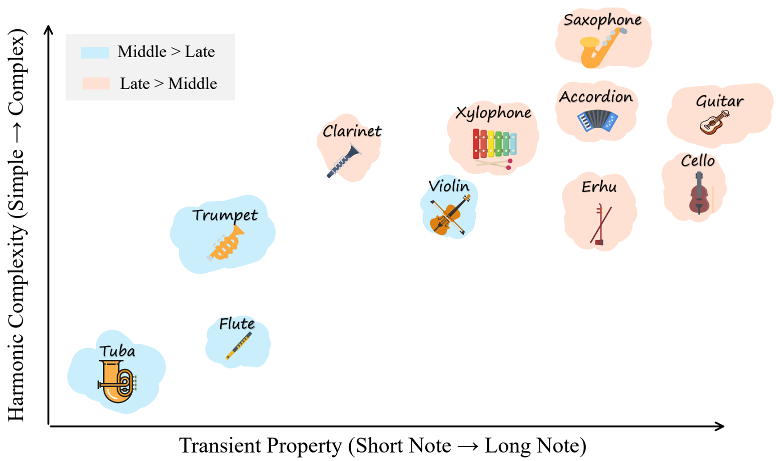


Figure 1: **Relationship Between Acoustic Properties of Musical Instruments and Fusion Strategies.** Instruments with shorter transient properties and simpler harmonic structures are more suited to middle fusion. Conversely, instruments with sustained notes and complex harmonic structures benefit more from late fusion. Details can be found in Appendix S1.

hum of a cello or the steady strumming of a guitar. Transient sounds often carry distinct temporal signatures, while sustained sounds exhibit intricate harmonic structures that evolve over time [1]. These differences in sound are not just something we hear; they also affect how we recognize and separate them.

Traditional audio separation methods rely solely on audio cues, struggling in complex environments where multiple sounds overlap [15, 17]. In recent years, to improve the performance of audio separation, researchers have turned to visual modality as a strong prior [28], known as "audio-visual separation", leveraging the natural correspondences between audio and visual signals. As the visual modality is introduced, a natural question arises: How should the visual information be effectively integrated with the existing audio modality to enhance separation performance? To address this, researchers have explored different fusion strategies to combine the two modalities. Middle fusion integrates visual features at the bottleneck of the audio U-Net [8], while late fusion applies visual features at the final layer of the audio U-Net [7, 28].

Figure 1 illustrates the relationship between **acoustic characteristics** and **fusion stages**. Middle fusion performs better at capturing sharp, short-duration sounds but struggles with harmonically rich sources, while late fusion is more effective for continuous sounds but may overlook transient details. This trade-off naturally leads us to ask: Can we design a fusion strategy that combines the strengths of both approaches to improve separation *across a wide range of sound characteristics?* Our approach builds upon the idea that different sound characteristics require different levels of fusion. We propose a hierarchical fusion strategy that integrates both middle and late fusion. In this way, both short, sharp sounds and long, continuous ones are handled at the most suitable stages.

In audio-visual separation, researchers have successfully leveraged large pre-trained vision models, such as CLIP [19], to extract strong visual representations that significantly improve performance [7, 25]. This naturally leads to the question: *Can we apply large pre-trained audio models to benefit the audio separation task?* At first glance, it seems intuitive to replace the audio U-Net's learned features with embeddings from a large pre-trained audio model. However, audio separation requires fine-grained time-frequency details to disentangle and reconstruct overlapping sounds, which the high-level audio embeddings from large pretrained model often lack. Thus, simply substituting learned features with pre-trained embeddings is not a good choice. Instead of directly replacing the audio latent representations at the bottleneck of U-Net, we propose representation alignment—an approach that aligns the latent space of the separation model with the embeddings from a large pre-trained audio model. In this way, the model can not only preserve fine-grained spectral details but also distill high-level semantic knowledge from the pre-trained embeddings.

To evaluate and understand the representation alignment approach, we investigate two key questions: Whether aligning the U-Net's latent features with a pre-trained model improves audio-visual separation; If so, what underlying factors contribute to this improvement. Our experiment findings show that the proposed representation alignment method not only improves audio separation performance, but also enhances the semantic richness of audio latent features. Interestingly, representation alignment also reduces the modality gap between audio and visual representations, even though no explicit objective was introduced to enforce multimodal alignment.

We highlight the main contributions of this paper below:

- Our study is the first to reveal the correlation between acoustic characteristics of sound sources and performances of different fusion stages in audio-visual separation.
- We propose a self-supervised hierarchical fusion strategy for audio-visual separation.
- We propose a representation alignment loss that bridges the semantic gap between U-Net bottleneck features and pre-trained audio embeddings.
- Experiments demonstrate that our proposed method achieves performance gains on various benchmark datasets.

# 2 Related Works

## 2.1 Audio Visual Separation

Existing audio-visual separation methods can be broadly categorized into self-supervised and weakly-supervised approaches. Self-supervised methods [7, 28, 29] leverage the popular mix-and-separate strategy that creates synthetic audio mixtures by combining audio from different videos, enabling models to learn separation without the need for extra human annotations. Weakly-supervised methods [5, 8] introduce additional semantic information, such as audio category labels, to provide indirect supervision. Compared to self-supervised methods, weakly-supervised methods can offer performance advantages but rely on additional annotations on audio categories, which increase labeling costs and may limit generalization to unseen scenarios. Thus, in this work, we focus on exploring self-supervised approaches.

While our method, like other self-supervised approaches, relies on global visual features, another line of work uses spatially grounded features via object detectors, e.g., Co-Separation [8], CCoL [22], and iQuery [5]. However, these methods have limitations that distinguish them from the self-supervised, global-feature-based approach: their performance is heavily dependent on the detector's accuracy, and they are restricted to categories the detector is trained on. For example, CCoL's masks can be blurry, while iQuery had to use a more general detector to accommodate new instruments in the MUSIC-21 dataset. In contrast, our approach is purely self-supervised, leveraging global video-level features and natural audio-visual correspondences for separation, which makes it more robust to object detection failures and removes the dependency on object-level annotations or detectors.

Another important design choice in audio-visual separation models is the stage at which audio and visual features are fused. Existing methods primarily follow two fusion strategies: late fusion and middle fusion. In late fusion [7, 28, 29], visual features are applied at the final stages of the U-Net decoder to reweight the audio spectrogram. Middle fusion integrates visual embeddings earlier in the network by tiling and concatenating them with the bottleneck features of the U-Net [20]. Prior studies suggest that late fusion generally outperforms middle fusion in self-supervised settings, while middle fusion has been shown to be beneficial in weakly-supervised scenarios [8, 22], particularly when combined with additional supervision signals, such as classification losses derived from labeled data [2, 3]. Different from previous works, we investigate how different sound characteristics influence the effectiveness of middle and late fusion and propose a hierarchical fusion strategy that combines both fusion mechanisms.

## 2.2 Cross-Modal Representation Learning

The CLIP model, based on contrastive learning, has been widely used as a pretraining framework to build joint embedding spaces for text and image modalities. Its success has also inspired extensions to the audio domain. For instance, [25] proposed a self-supervised approach where an additional audio encoder is trained to align input audio with the pretrained CLIP embedding space, enabling audio representations to inherit the multimodal alignment capabilities of CLIP. Similarly, [7] explore the zero-shot modality transfer capability of CLIP by keeping the pre-trained model frozen while optimizing only the remaining components for the target sound separation task. In our work, we follow the approach of [7] which uses CLIP to extract visual features.

Similarly, the CLAP model [6, 26] extends such contrastive pretraining paradigm to audio-text embedding space CLAP has been applied to audio-visual segmentation and detection tasks [9, 16].However, its potential for audio-visual separation remains underexplored.

# 3 Method

## 3.1 Problem Definition

Audio separation aims to isolate individual sound sources from a mixed audio signal. During training, the model takes as inputs an audio mixture $\mathbf{x} = \sum_{i=1}^{n} s_i$, where $s_1, \ldots, s_n$ are the $n$ audio tracks, along with their corresponding images $\mathbf{y}_1, \ldots, \mathbf{y}_n$ extracted from the videos. We first transform the audio mixture $\mathbf{x}$ into a magnitude spectrogram $\mathbf{X} = |\mathrm{STFT}(\mathbf{x})|$ and pass the spectrogram through an audio U-Net [20] to produce $k$ ($\geq n$) intermediate masks $\tilde{M}_1, \ldots, \tilde{M}_k$. On the other stream, each image $\mathbf{y}_i$ is encoded into an embedding $\mathbf{e}_i = \mathrm{Enc}_{img}(\mathbf{y}_i)$.

Middle fusion and late fusion integrate visual information at different stages of the separation model. The process of middle fusion can be written as follows:

$$\tilde{M}_i = \mathrm{Dec}_a\left(\mathrm{Tile}(\mathbf{e}_i) \oplus \mathrm{Enc}_a(\mathbf{X})\right). \tag{1}$$

The process of late fusion can be written as follows:

$$\tilde{M}_i = \mathrm{Proj}_1(\mathbf{e}_i) \odot \mathrm{Dec}_a(\mathrm{Enc}_a(\mathbf{X})), \tag{2}$$

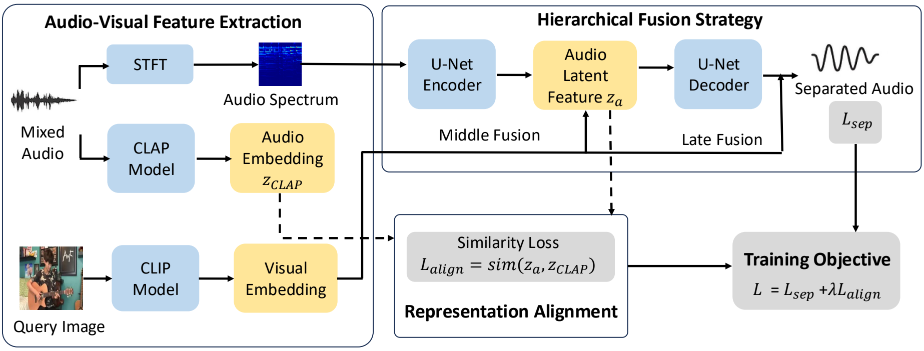


Figure 2: **Pipeline of our proposed method.** The pipeline consists of three key components: audio-visual feature extraction, hierarchical fusion, and representation alignment. It takes an audio mixture and corresponding video frames as input. The Audio-Visual Feature Extraction module processes the input through dedicated encoders, extracting audio features from spectrograms and CLAP and extracting visual features from CLIP. Hierarchical fusion includes middle fusion and late fusion, which happens at the bottleneck of the audio U-Net and the final layer of the audio decoder separately.

where the function $\mathrm{Enc}_a(\cdot)$ denotes the audio encoder, which extracts bottleneck audio features, while $\mathrm{Dec}_a(\cdot)$ represents the audio decoder. In middle fusion, $\mathrm{Tile}(\mathbf{e}_i)$ expands the visual embedding spatially to match the dimensions of $\mathrm{Enc}_a(\mathbf{X})$, and $\oplus$ denotes the channel-wise concatenation of visual and audio features. Late fusion applies a projection function $\mathrm{Proj}_1(\mathbf{e}_i)$ mapping the visual embedding to a compatible space before performing element-wise multiplication ($\odot$) with the decoded audio representation.

The separated spectrogram $\hat{\mathbf{X}}_i$ for each source $i$ is then obtained by multiplying the estimated mask $\tilde{M}_i$ to the input mixed spectrogram $\mathbf{X}$ through element-wise multiplication:

$$\hat{\mathbf{X}}_i = \tilde{M}_i \odot \mathbf{X}. \tag{3}$$

The separation loss $\mathcal{L}_{\text{sep}}$ is then defined using the $L_1$ distance between the predicted spectrogram $\hat{\mathbf{X}}_i$ and the ground-truth spectrogram $\mathbf{X}_i$:

$$\mathcal{L}_{\text{sep}} = \sum_{i=1}^{n} \|\hat{\mathbf{X}}_i - \mathbf{X}_i\|_1. \tag{4}$$

## 3.2 Hierarchical fusion for Audio-Visual Separation

We can observe from Figure 1 that middle and late fusion strategies exhibit complementary advantages depending on the characteristics of the target sound. In Specifically, middle fusion is more effective at separating transient sound instruments (*e.g.*, trumpet, flute) and stable low-frequency instruments (*e.g.*, tuba), whereas late fusion performs better for sustained instruments (*e.g.*, saxophone, cello) and harmonically rich instruments (*e.g.*, acoustic guitar, xylophone). Appendix S3 further provide a quantitative analysis of the relationship between separation performance and fusion stage. Motivated by such observation, we propose a hierarchical fusion strategy that integrates both middle and late fusion to leverage their complementary strengths. The hierarchical fusion mechanism is formulated as:

$$\tilde{M}_i = \mathrm{Proj}_1(\mathbf{e}_i) \odot \mathrm{Dec}_a\left(\mathrm{Tile}(\mathrm{Proj}_2(\mathbf{e}_i)) \oplus \mathrm{Enc}_a(\mathbf{X})\right), \tag{5}$$

where $\text{Proj}_1$ and $\text{Proj}_2$ are both single fully connected layers that project the visual embedding $\mathbf{e}_i$ into appropriate latent spaces for different fusion mechanisms.

### 3.3 Audio Representation Alignment

Previous studies have shown that using CLIP-extracted visual features greatly improves audio-visual separation [7] because CLIP learns strong multimodal representations that align well with semantic categories. Inspired by this, we hypothesize that learning high-quality audio representations can similarly enhance source generation performance in audio-visual separation. In this way, we propose audio representation alignment, which encourages the U-Net encoder's latent representations to align with self-supervised audio embeddings extracted by the pretrained large audio model.

The audio representation alignment loss function is defined as follows:

$$\mathcal{L}_{\text{audio\_align}} = 1 - \text{sim}(\mathbf{z}_a, \mathbf{z}_{\text{CLAP}}), \tag{6}$$

where the latent representations are defined as $\mathbf{z}_a = \text{Enc}_a(\mathbf{X})$ and $\mathbf{z}_{\text{CLAP}} = \text{Enc}_{\text{CLAP}}(\mathbf{x})$.

Here, $\text{Enc}_a(\cdot)$ denotes the audio encoder of the U-Net, which extracts latent features from the input spectrogram $\mathbf{X}$, and $\text{Enc}_{\text{CLAP}}(\cdot)$ represents the pretrained self-supervised audio encoder from CLAP, which generates a semantic embedding from the waveform of the same mixed audio input $\mathbf{x}$. The function $\text{sim}(\cdot,\cdot)$ is thecosine similarity function:

$$\text{sim}(\mathbf{z}_a, \mathbf{z}_{\text{CLAP}}) = \frac{\mathbf{z}_a \cdot \mathbf{z}_{\text{CLAP}}}{\|\mathbf{z}_a\| \|\mathbf{z}_{\text{CLAP}}\|}. \tag{7}$$

In practice, we add this term to the original audio separation objectives described in Section 3.1. Therefore, the overall training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{sep}} + \lambda \mathcal{L}_{\text{audio\_align}}, \tag{8}$$

where $\lambda > 0$ is a hyperparameter that controls the tradeoff between audio separation performance and representation alignment.

## 4 Experiments

In this section, we conducted experiments not only to demonstrate that our proposed method can improve the performance of audio-visual separation but also to gain deeper insights into **why** these improvements occur.

### 4.1 Experimental Settings

**Datasets.** We conduct experiments on three widely-used datasets: MUSIC [28], MUSIC-21 [29] and VGGSound[4]. The MUSIC dataset contains 601 untrimmed videos of musical solos and duets across 11 categories of musical instruments due to some videos becoming online unavailable over time. Following the data splits of [28], we use 483 videos for training and 118 videos for testing, with the test set exclusively consisting of solo performances. The MUSIC-21 dataset [29] consists of solo videos across 21 instrument categories. We utilize 1,039 online available solo videos and adopt the training/testing split of [29], with 831 videos for training and 208 for testing. VGGSound [4] is a large-scale audio-visual dataset consisting of 10-second video clips "in the wild", covering 309 sound event categories. Our training set contains 132,760 videos, and our test set includes 11,147 videos.

**Baselines.** We compare our method with several recent self-supervised approaches across all three datasets. On the MUSIC dataset, we include: NMF-MFCC [24], a non-learnable audio-only baseline (results from [8]); Sound-of-Pixels [28]; CLIPSep [7], retrained under the same settings; and Semantic Grouping Network (SGN) [14]. For MUSIC-21, we compare against NMF-MFCC, Sound-of-Pixels, and CLIPSep following the same protocol. On VGGSound, we evaluate our method against CLIPSep as the main baseline, given its strong performance in audio-visual separation.

**Evaluation Metrics.** We assess the performance of sound separation using three standard metrics: Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifacts Ratio (SAR). SIR evaluates how well interfering sources are suppressed, while SAR reflects the level of artifacts introduced during separation [23]. Among these, SDR is generally considered the most important metric, as it provides an overall measure of separation quality [18] by accounting for both interference and artifacts. In summary, higher values for all three metrics indicate better performance while we primarily focus on the SDR.

**Implementation Details.** All audio is resampled to 11,625 Hz. For MUSIC and MUSIC-21, we extract 5-second segments from each video (random crop for training, center crop for testing). VGGSound clips are 10 seconds long and used without cropping to retain real-world diversity. Audio is processed using STFT with a Hann window of size 1024 and hop length of 256, producing complex spectrograms. We then compute the magnitude and apply a logarithmic mapping along the frequency axis to reflect human perception. The final input is a $256 \times 256$ log-scaled magnitude spectrogram, where $T = 256$ time frames and $F = 256$ frequency bins. Visual inputs are sampled at 8 frames per second. A single frame per clip is extracted (randomly for training, center for testing) and encoded using a frozen CLIP ViT-B/32 to obtain a 512-dimension embedding. We train our model using a batch size of 32. The learning rate is initially set to $10^{-4}$ and reduced by a factor of 0.1 at the 60th epoch.

Table 1: **Audio-visual separation performance comparison on the MUSIC dataset.** Best results in **bold**, second-best <u>underlined</u>.

| Method | SDR ↑ | SIR ↑ | SAR ↑ |
|---|---|---|---|
| RPCA [10] | −0.62 | 2.32 | 2.41 |
| Wave-U-Net [21] | 3.80 | 6.75 | 6.62 |
| ResUNetDecouple+ [12] | 3.98 | 7.17 | 6.91 |
| MP-Net [27] | 4.82 | 10.19 | 10.56 |
| SGN [14] | 5.20 | 10.81 | 10.67 |
| NMF-MFCC [24] | 0.92 | 5.68 | 6.84 |
| Sound-of-Pixels [28] | 3.84 | 9.66 | 9.32 |
| CLIPSep (Middle Fusion) [7] | 5.57 | **12.99** | 8.60 |
| CLIPSep (Late Fusion) [7] | <u>5.86</u> | 11.65 | <u>9.72</u> |
| **Ours** (Hierarchical + Align) | **6.72** | <u>12.60</u> | **10.21** |

Table 2: **Audio-visual separation results on MUSIC21 dataset.** Best results in **bold**, second-best underlined. †Results from Table I in [5]. ‡Results from Table 3 in [11].

| Methods | SDR ↑ | SIR ↑ | SAR ↑ |
|---|---|---|---|
| NMF-MFCC [24]† | 2.78 | 6.70 | 9.21 |
| AV-MMix-and-Separate [8]‡ | 3.23 | 7.01 | 9.14 |
| Sound-of-Pixels [28] | 6.51 | 12.84 | 10.58 |
| CLIPSep (Middle Fusion) [7] | 7.36 | **14.28** | 10.22 |
| CLIPSep (Late Fusion) [7] | 7.27 | 13.10 | <u>11.14</u> |
| **Ours** (Hierarchical) | <u>7.72</u> | 13.63 | 10.94 |
| **Ours** (Hierarchical + Align) | **8.03** | <u>13.92</u> | **11.36** |

Table 3: **Audio-visual separation results on VGGSound.** Best results in **bold**, second-best underlined.

| Method | SDR ↑ | SIR ↑ | SAR ↑ |
|---|---|---|---|
| CLIPSep (Late Fusion) | 0.90 | 7.47 | <u>8.26</u> |
| CLIPSep (Middle Fusion) | <u>1.16</u> | **9.41** | 6.95 |
| **Ours** (Hierarchical + Align) | **1.97** | <u>8.96</u> | **9.62** |

## 4.2 Audio-Visual Sound Source Separation Results

**Quantitative Evaluation.** We evaluate our method on MUSIC, MUSIC-21, and VGGSound, with results summarized in Table 1, Table 2, and Table 3, respectively. Overall, our approach consistently outperforms existing baselines, especially in terms of SDR, which is widely regarded as the most important metric for source separation.

These results validate the effectiveness of our design in handling a wide range of audio-visual scenarios.

While our method achieves state-of-the-art SDR scores, we observe a slightly lower SIR compared to some baselines, such as CLIPSep with Middle Fusion on the MUSIC-21 and VGGSound datasets. We attribute this to the trade-off between suppressing interfering sources (SIR) and avoiding the introduction of artifacts (SAR). Our approach prioritizes reducing artifacts, as evidenced by our consistently high SAR scores across all datasets, particularly on MUSIC and VGGSound. Since SDR provides a comprehensive measure of separation quality by accounting for both interference and artifacts, our superior SDR performance indicates that the gain from reducing artifacts outweighs the minor decrease in interference suppression, leading to a higher overall separation quality.

**Qualitative Evaluation.** Figure 3 provides qualitative examples of sound separation on the MUSIC dataset. Our method produces cleaner and more distinct separated sounds, reducing interference. The visualization results further support that our approach can separating mixed musical components with high precision.

## 4.3 Audio Representation and Modality Gap Evaluation

To evaluate the effect of representation alignment, we perform linear probing on the MUSIC test set. As shown in Table 4, for U-Net latent feature, classification accuracy significantly

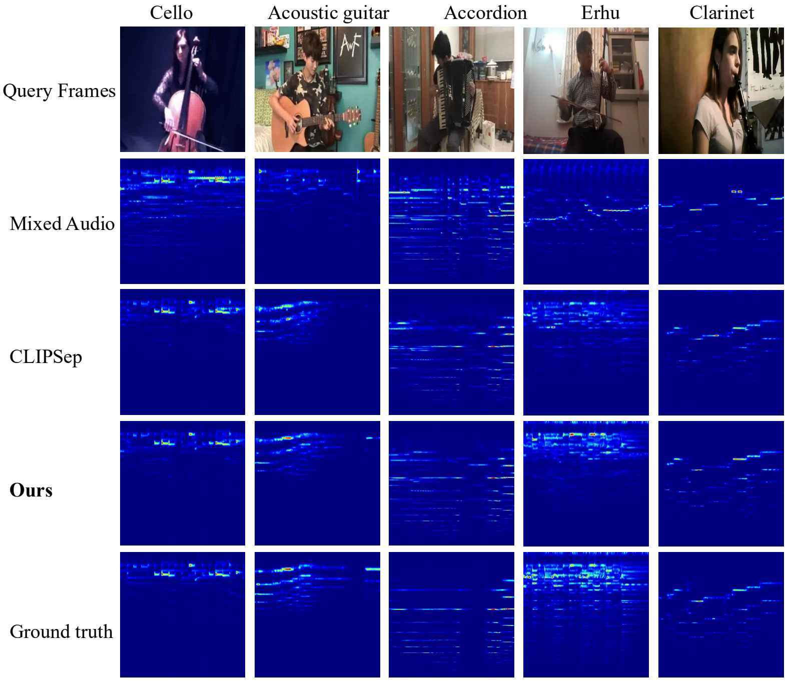


Figure 3: **Qualitative Performance on MUSIC dataset.** We compared our method (the fourth row) with Dong et al. [7] (the third row).

improves after using representation alignment, indicating that aligned U-Net features capture more semantic information. Although classification accuracy remains slightly lower than that of frozen CLAP embeddings, this is reasonable. CLAP is trained explicitly for semantic discrimination, while the U-Net encoder is optimized for source separation, focusing more on preserving fine-grained spectral details than on maximizing classification accuracy.

Table 4: **Linear probing accuracy on different audio representations.**

| Method | Representation Source | Accuracy (%) |
|---|---|---|
| W/O Alignment | U-Net Bottleneck | 49.58 |
| W/ Alignment | U-Net Bottleneck | 58.82 |
| Frozen CLAP | CLAP Embedding | 59.66 |

We further calculate the modality gap between audio and visual representations on the MUSIC test dataset following [13]. Given audio embeddings $\mathbf{X} = \{x_i\}_{i=1}^{N}$ and visual embeddings $\mathbf{Y} = \{y_i\}_{i=1}^{N}$, the modality gap is defined as follows:

$$\Delta_{\text{gap}} = \frac{1}{N}\sum_{i=1}^{N} x_i - \frac{1}{N}\sum_{i=1}^{N} y_i \tag{9}$$

Results are shown in Table 5. We can observe that representation alignment leads to an obvious reduction in modality gap. Interestingly, this reduction occurs despite the absence of an explicit objective to minimize the gap between audio and visual modalities. Although the alignment process is not directly designed for this purpose, it implicitly promotes better interaction between audio and visual features, ultimately benefiting the source separation task.

Table 5: **Comparison of modality gap with and without representation alignment.**

| Method | Modality Gap |
|---|---|
| W/O Alignment | 1.171 |
| W/ Alignment | 0.976 |

## 4.4 Ablation Study

Table 6 presents an ablation study evaluating the impact of visual backbone, fusion strategy, and representation alignment on audio-visual separation performance. Results show that hierarchical fusion consistently yields higher SDR and SAR scores, while maintaining a competitive SIR score. Additionally, with the combination of hierarchical fusion and alignment achieving the best overall results.

Table 6: **Ablations on fusion strategies and with or without representation alignment.**

| Visual Backbone | Fusion | Alignment | SDR ↑ | SIR ↑ | SAR ↑ |
|---|---|---|---|---|---|
| ResNet-18 | Late | ✗ | 3.84 | 9.66 | 9.32 |
| CLIP ViT-B/32 | Middle | ✗ | 5.57 | **12.99** | 8.60 |
| CLIP ViT-B/32 | Late | ✗ | 5.86 | 11.64 | 9.72 |
| CLIP ViT-B/32 | Late | ✓ | 6.07 | 12.10 | 9.73 |
| CLIP ViT-B/32 | Hierarchical | ✗ | 6.65 | 12.39 | 10.11 |
| CLIP ViT-B/32 | Hierarchical | ✓ | **6.72** | 12.60 | **10.21** |

# 5 Conclusion

In this work, we analyze the complementary strengths of middle and late fusion in audio-visual separation and propose a hierarchical fusion strategy that leverages both to improve performance across diverse scenarios. To further enhance semantic understanding, we introduce a representation alignment mechanism that aligns U-Net audio features with CLAP embeddings, enriching semantic information while preserving spectral details. Experiments confirm that both components contribute to separation performance. While our model uses global CLIP features without explicit sound source localization, future work can explore self-supervised localization to guide visual attention toward sounding objects and further improve separation performance.

# Acknowledgements

This project is partially supported by the Royal Society grants (SIF\R1\231009, IES\R3\223050) and an Amazon Research Award. The computations in this research were performed using the Baskerville Tier 2 HPC service. Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP\T022221\1) and the Digital Research Infrastructure programme (EP\W032244\1) and is operated by Advanced Research Computing at the University of Birmingham.

# References


[1] Pauli Brattico, Elvira Brattico, and Peter Vuust. Global sensory qualities and aesthetic experience in music. *Frontiers in Neuroscience*, 11, 2017. URL https://api.semanticscholar.org/CorpusID:12298812.

[2] 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*, pages 1204–1213, 2021.

[3] Moitreya Chatterjee, Narendra Ahuja, and Anoop Cherian. Learning audio-visual dynamics using scene graphs for audio source separation. *Advances in Neural Information Processing Systems*, 35:16975–16988, 2022.

[4] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. Vgg-sound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[5] Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. iquery: Instruments as queries for audio-visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14675–14686, 2023.

[6] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2022.

[7] Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian McAuley, and Taylor Berg-Kirkpatrick. Clipsep: Learning text-queried sound separation with noisy unlabeled videos. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2023.

[8] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3879–3888, 2019.

[9] Ruohao Guo, Liao Qu, Dantong Niu, Yanyu Qi, Wenzhen Yue, Ji Shi, Bowei Xing, and Xianghua Ying. Open-vocabulary audio-visual semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7533–7541, 2024.

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

[11] Md. Amirul Islam, Seyed Shahabeddin Nabavi, Irina Kezele, Yang Wang, Yuanhao Yu, and Jin Tang. Visually guided audio source separation with meta consistency learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3014–3023, 2024. URL https://openaccess.thecvf.com/content/WACV2024/papers/Islam_Visually_Guided_Audio_Source_Separation_With_Meta_Consistency_Learning_WACV_2024_paper.pdf.

[12] Qiuqiang Kong, Yin Cao, Haohe Liu, Keunwoo Choi, and Yuxuan Wang. Decoupling magnitude and phase estimation with deep resunet for music source separation. *arXiv preprint arXiv:2109.05418*, 2021.

[13] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022. URL https://openreview.net/forum?id=S7Evzt9uit3.

[14] Shentong Mo and Yapeng Tian. Semantic grouping network for audio source separation, 2024. URL https://arxiv.org/abs/2407.03736.

[15] Md Khademul Islam Molla and Keikichi Hirose. Single-mixture audio source separation by subspace decomposition of hilbert spectrum. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):893–900, 2007.

[16] Shota Nakada, Taichi Nishimura, Hokuto Munakata, Masayoshi Kondo, and Tatsuya Komatsu. Deteclap: Enhancing audio-visual representation learning with object information. *arXiv preprint arXiv:2409.11729*, 2024.

[17] Kakali Nath and Kandarpa Kumar Sarma. Separation of overlapping audio signals: a review on current trends and evolving approaches. *Signal Processing*, page 109487, 2024.

[18] Darius Petermann, Pritish Chandna, Helena Cuesta, Jordi Bonada, and Emilia Gomez. Deep learning based source separation applied to choir ensembles, 2020. URL https://arxiv.org/abs/2008.07645.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[21] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018.

[22] 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)*, pages 2745–2754, 2021.

[23] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14 (4):1462–1469, 2006. doi: 10.1109/TSA.2005.858005.

[24] 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.

[25] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[26] Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[27] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019.

[28] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[29] 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*, pages 1735–1744, 2019.