IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5713∼5726
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5713-5726

Fusing audio-visual-movement representation enhanced by language information for sound source separation

Hien-Thanh Duong[1], Huu-Mao Dao[2], The-Thanh-Dat Le[2], Phuong-Bac Nguyen[3], Thi-Lan Le[2]
[1]School of Information and Communications Technology, Hanoi University of Industry, Hanoi, Vietnam
[2]School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam
[3]Department of Information Technology, Hanoi University of Science and Technology, Hanoi, Vietnam

Article Info

Article history:

Received May 18, 2025
Revised Aug 14, 2026
Accepted Sep 20, 2026

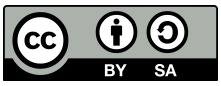
Keywords:

Audio source separation
Audio-visual fusion
Language-guided action representation
Multimodal learning
Pose estimation

ABSTRACT

Audio source separation (ASS) has witnessed significant progress with the integration of multimodal cues, particularly visual and linguistic information. This paper proposes a novel framework named movement-instrument-audio fusion for sound source separation (MIA3S) that leverages movement representations guided by natural language to enhance source separation in musical performances. The proposed approach extends the language knowledge-assisted graph convolutional network (LA-GCN) architecture to encode musician gestures using textual prompts, which inform graph-based movement features aligned with instrumental actions. These features are fused with audio and object-based visual features in a U-Net-based model to estimate separation masks. Evaluated on the MUSIC dataset, the proposed model achieves competitive performance, significantly outperforming baseline and the state-of-the-art methods in terms of signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and F1-score for action recognition. The experimental results confirm the effectiveness of combining language-guided action representations with multimodal fusion for improving the performance of ASS systems.

This is an open access article under the CC BY-SA license.



Corresponding Author:

Thi-Lan Le
School of Electrical and Electronic Engineering, Hanoi University of Science and Technology
No. 1 Dai Co Viet, Hai Ba Trung, Hanoi, Vietnam
Email: lan.lethi1@hust.edu.vn

1. INTRODUCTION

Audio source separation (ASS) is a core research problem in signal processing and artificial intelligence that aims to separate an observed audio mixture into its individual sound sources. Conventional methods, including independent component analysis (ICA) and nonnegative matrix factorization (NMF), typically rely on assumptions such as source independence and linear mixing, which restrict their effectiveness in practical acoustic environments [1]–[3]. More recently, deep learning has substantially advanced source separation by enabling end-to-end optimization of nonlinear mappings through neural architectures such as Wave-U-Net [4] and convolutional time-domain audio separation network (ConvTasNet) [5].

The incorporation of complementary modalities, particularly visual and language information, has recently become an effective strategy for enhancing ASS. In visual-guided ASS, visual information derived from object appearance or motion in video is exploited to facilitate the separation process, particularly in acoustically challenging scenes containing multiple simultaneous sound sources [6], [7]. In music performance

contexts, the visuals of instruments and the gestures of the musicians are particularly informative for distinguishing different instrumental sounds.

In parallel, the use of natural language in audio-visual modeling has gained increasing attention. Language-queried audio source separation (LASS) enables users to interact with the separation system by specifying the desired sound source through natural language descriptions. Recent studies have introduced models that integrate natural language as a control signal in the separation task. For example, Liu *et al.* [8] introduced the task of LASS, demonstrating the potential of natural language queries in guiding ASS. Additionally, Tan *et al.* [9] proposed a self-supervised approach for audio-visual source separation based on natural language queries, highlighting the effectiveness of language guidance in multimodal learning. However, current approaches predominantly treat language as a global semantic condition and do not account for the temporal and structural correspondence between actions and sound. Specifically, the intersection of action representations and linguistic information for guiding source separation remains under-explored.

This paper proposes a novel movement-instrument-audio fusion for sound source separation (MIA3S) framework that unifies language-grounded movement representations and audio-visual signals to enhance ASS. Building on Pham *et al.* [10] previous work on visual-guided sound separation and gesture-informed source modeling [7], this work leverage the language knowledge-assisted graph convolutional network (LA-GCN) architecture to encode natural language prompts related to musician gestures. These prompts generate prior topologies that inform a graph-based action encoder, enabling the system to learn semantically aligned motion features that correspond to instrument-specific sound patterns. This work has three main contributions: i) extending the LA-GCN architecture to model language-guided movement embeddings in the context of instrumental sound separation; ii) proposing an MIA3S framework that integrates audio, visual, movement, and language modalities for enhancing ASS; and iii) demonstrating the new application of LA-GCN for audio applications, extending its original design for action recognition to the domain of ASS. Through empirical evaluation on the MUSIC dataset, this work demonstrate that proposed method significantly outperforms baseline audio-visual separation models, especially in scenarios involving visually similar instruments with distinct playing gestures. By incorporating natural language knowledge into modeling musician gestures/movements, the proposed method bridges the gap between linguistic semantics, human activity, and acoustic modeling, paving the way for more generalizable and interpretable multimodal audio separation systems.

2. RELATED WORKS

In recent years, research on source separation has seen rapid development, especially with the integration of vision and language modalities. To situate the contributions, this section mentions related work in three main approaches. These include: i) audio-only source separation, ii) audio-visual source separation, and iii) query-based source separation and action representation with language priors.

2.1. Audio-only source separation

Audio-only source separation techniques have significantly evolved over the past two decades. Early methods primarily relied on classical signal processing techniques, which exploited statistical independence or sparsity assumptions. One of the foundational approaches was ICA, which separates mixed signals based on the statistical independence of sources in the time or frequency domain. For instance, Makino *et al.* [1] applied ICA with microphone arrays, demonstrating effective separation in structured environments. However, ICA techniques often assume a linear and instantaneous mixing model and require the number of sources not to exceed the number of sensors. This limits their scalability in practical situations, such as single-channel recordings or when the number of audio sources exceeds the number of sensors.

To overcome these limitations, researchers introduced NMF as a more flexible technique for monaural or underdetermined mixtures. NMF decomposes the magnitude spectrogram of an audio mixture into a set of nonnegative spectral bases and corresponding temporal activations, allowing each source to be represented by distinct spectral patterns. Building upon this formulation, various extensions have incorporated prior knowledge to improve source separation performance. One notable application of NMF is in informed source separation, where additional information about the sources, such as pre-trained spectral templates or side-channel metadata, is used to guide the separation process. Ozerov *et al.* [2] proposed a flexible ASS framework that integrates NMF with spatial models and supports including the prior information such as harmonicity, speaker identity, or temporal continuity, significantly improving separation performance in

real-world scenarios. Besides, Rohlfing *et al.* [11] introduced an informed NMF-based source separation method in which side information is represented by a compact set of parameters derived from reference sources. Zhang *et al.* [12] enhanced multichannel NMF by incorporating a Gaussian mixture prior on spectral basis vectors to improve source discrimination and separation performance. Despite its mathematical elegance, traditional NMF still has limitations in modeling temporal dynamics and often requires carefully tuned constraints (e.g., sparsity and temporal smoothness) for practical performance.

The emergence of deep learning has significantly advanced source separation performance by introducing data-driven models that can learn complex nonlinear mappings from mixtures to source estimates. Among the pioneering neural architectures is Wave-U-Net, introduced by Stoller *et al.* [4], which adapts the U-Net architecture to operate directly on raw waveforms in a multi-scale fashion. Unlike spectrogram-based models, Wave-U-Net avoids explicit phase reconstruction and enables end-to-end learning, achieving competitive performance in music source separation. Further breakthroughs came with ConvTasNet, introduced by Luo and Mesgarani [5], which operates entirely in the time domain using temporal convolutional networks. ConvTasNet surpassed traditional time–frequency masking techniques by learning an optimal latent representation of the signal and employing a powerful separation module with long receptive fields and residual connections. It demonstrated substantial improvements in speech separation benchmarks. Nonetheless, audio-only systems often struggle in complex acoustic scenes where overlapping sources share similar spectral features or where spatial cues are absent. These limitations have motivated multimodal approaches that incorporate visual or semantic information to better disentangle audio sources.

2.2. Audio-visual source separation

Audio-visual source separation (AVSS) aims at enhancing the ability to separate sounds by leveraging visual context from video data. Visual cues such as object appearance, motion, and spatial consistency provide critical information that can disambiguate overlapping sources in challenging auditory environments. With these prospects, numerous studies focusing on the integrated exploitation of audio and visual information have been conducted recently: the method presented in [13] learns pixel-to-sound associations from synchronized audio-visual data, enabling the localization and separation of visually grounded sound sources. Similarly, Ephrat *et al.* [14] developed a speaker-independent AVSS model that uses lip motion to isolate speech in noisy multi-speaker scenarios. More recent advancements, such as Song and Zhang's model [6], integrate audio-visual predictive coding to learn temporal cross-modal correlations, while Zhu and Rahtu [15] proposed a cascaded refinement architecture that incorporates opponent filter networks and visual localization masks to recursively improve separation.

Further, tri-modal architecture based on bidirectional encoder representations from transformers (TriBERT) [16] introduces full-body action features to guide human-centric sound separation, emphasizing the importance of motion beyond simple object detection. Gao and Grauman's work [17] on the co-separation of sounds represents another pivotal direction. Their model learns to associate detected objects in the visual scene with latent audio components. While effective at using object context to separate sources, it does not incorporate explicit motion dynamics. Interestingly, the authors suggest that future work should explore spatio-temporal cues for further improvement. Afouras *et al.* [18] take a different route, using self-supervised learning from unlabelled videos to discover audiovisual correlations without human annotations, opening the door for more scalable AVSS systems. These studies mentioned above, along with several others, have collectively demonstrated AVSS's potential, but limitations remain in handling occlusions, generalizing to open-set environments, and incorporating fine-grained temporal or performer cues.

2.3. Query-based source separation and action representation with language priors

Beyond using visual context, a growing body of work has explored query-based sound separation, where the model responds to user-specified queries in the form of audio, images, or text. Among these, iQuery [19] introduces instrument representations as queries and employs a cross-modal transformer to associate each query with its corresponding sound source in complex mixtures. However, iQuery mainly exploits appearance and motion cues extracted from video frames without explicitly modeling musician gestures or body movements, which could provide more informative cues for sound separation in performance-oriented scenarios. Earlier object-centric models, such as [17], similarly ground audio components in object regions without accounting for dynamic performer behavior.

Other works like CLIPSep [20] utilized pre-trained vision-language models (e.g., CLIP) to match sound components with text queries. Liu *et al.* [8] further advanced this method by introducing LASS [21].

This growing line of research shows promise, especially when integrated with language priors and semantic representations. Still, explicit modeling of human actions, including gestures and motion-based cues tied to sound production, remains underexplored.

In summary, audio-visual and language-guided sound separation has seen promising strides. However, most approaches are modular and dataset-specific. Cross-modal systems often focus on retrieval or localization and treat sound separation as a secondary task. The field lacks a unified architecture that effectively models and aligns the temporal dynamics of actions, visual cues, and semantics from language. This work aims to develop a joint audio-visual-language model for sound source separation. Unlike previous methods, the proposed approach incorporates both pre-trained semantic representations and task-specific alignment objectives, enabling more interpretable, generalizable, and robust performance in audio-visual environments.

3. MOVEMENT-INSTRUMENT-AUDIO FUSION FOR SOUND SOURCE SEPARATION FRAMEWORK

3.1. General workflow

This paper introduce a novel framework MIA3S, to separate different audio sources from video, comprising three key processing components: i) instrument detection and instrument feature extraction, ii) joints estimation and movement feature extraction, and iii) sound source separation as illustrated in Figure 1. These components collaboratively process input data consisting of video and its single-channel audio signal. The object detector identifies potential audio source objects in video frames, followed by the visual feature extractor, which extracts features from these identified objects. An action feature extractor is employed to capture the musician's action features from the video.

To this end, a LA-GCN is used to encode the pose sequence into discriminative movement features. The sound source separation module utilizes U-Net to combine audio features with visual and action features to estimate the spectral masks of the audio sources. This integrated approach enables the AVSS system to effectively isolate and distinguish various audio sources within video content. The following section describes these components in detail.

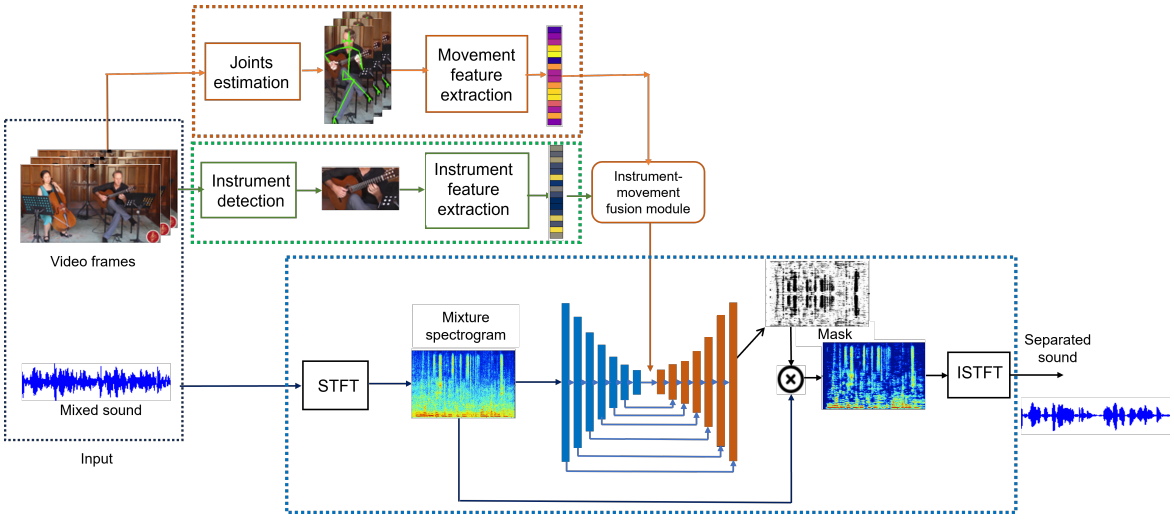


Figure 1. The general workflow of the proposed MIA3S system

3.2. Instrument detection and instrument feature extraction

Previous studies [17] have demonstrated that objects, particularly instruments, serve as important cues in guiding the sound source separation process. Building on this insight, thee proposed study incorporates instrument-related visual features into the separation network. It is worth noting that any object detection model can be applied. In proposed implementation adopt the you only look once version 5 (YOLOv5) model to detect instruments within video frames. Subsequently, a residual network with 18 layers (ResNet-18) network is used to extract visual features from the detected instrument regions as introduced [10].

3.3. Joints estimation and movement feature extraction

Besides instrument appearance, performers' body movements during instrument performance provide important cues for guiding sound source separation. To integrate these action-based features into the separation framework, various visual modalities such as red, green, and blue (RGB) images, skeleton data, or features from video representation models like video masked autoencoder (VideoMAE) [22] can be utilized. In this study choose to use skeleton information due to its lightweight nature and robustness to appearance variations such as clothing changes.

For movement representation, this work employ AlphaPose [23] to estimate human joints from video frames. AlphaPose combines object detection with deep learning techniques to produce accurate and reliable 2D joint estimations. It has been extensively trained to detect and localize key joints (including fingers, wrists, elbows, shoulders, knees, and ankles) even under challenging conditions such as occlusions and inconsistent lighting. Given a 10-second video clip estimates a sequence of 68-joint skeletons, consisting of 26 body joints and 21 joints for each hand. Each joint is described by its 2D coordinates (x, y) and an associated confidence score.

To capture the temporal dynamics and spatial structure of human joint movements, this work apply a LA-GCN [24], which effectively encodes the pose sequence into a discriminative movement feature for guiding the source separation process. LA-GCN is a novel approach that integrates graph neural networks (GNN) with language model knowledge to enhance activity recognition. The model comprises two key components: the global prior relationship (GPR) graph and the category prior relationship (CPR) graph. The GPR graph leverages bidirectional encoder representations from transformers (BERT) to extract joint features and establish relationships between distant nodes while maintaining essential "bones" for simplified topology modeling. The CPR graph encodes prior knowledge from large language models for various action classes, utilizing a PC-AC module to create category template topologies that provide additional supervision for the GCN. This approach is particularly effective for complex action classifications. LA-GCN also introduces a multi-hop attention graph convolution (MHA-GC) block to improve inter-node message passing efficiency. By integrating natural language knowledge into graph data feature learning, LA-GCN enhances the robustness and efficiency of activity recognition tasks.

To configure the LA-GCN model for the task, this work first create a list of textual prompts with (1).

What will [J] act like when [C] (1)

Where *[J]* represents the joints and *[C]* represents the action name, which are then input into the BERT model for encoding. The defined list of textual prompts in this study is shown in Figure 2.

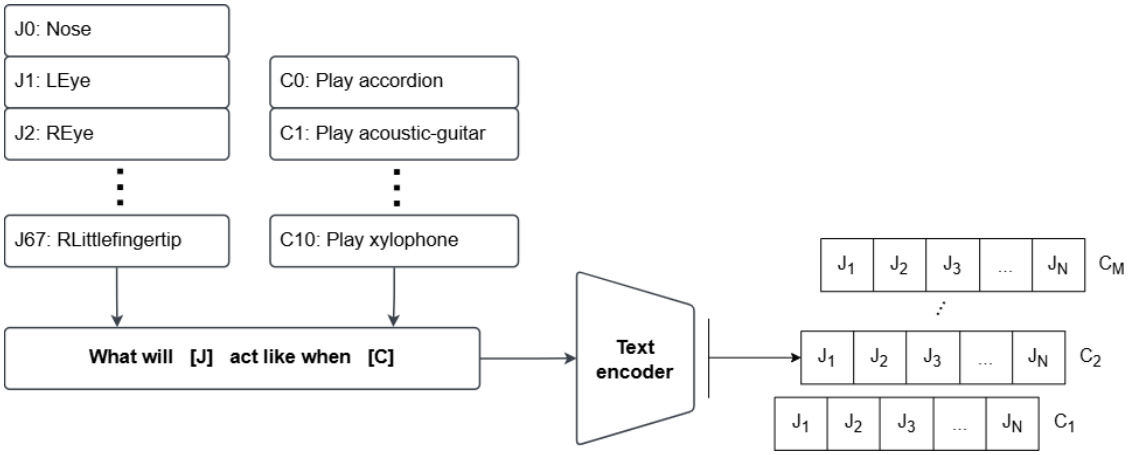


Figure 2. The list of textual prompts

BERT encodes the input sentences by adding [CLS] tokens at the beginning and [SEP] tokens at the end. After passing through N transformer layers, the final output of the model includes the features of each token and the [CLS] feature after the Pooler. The Pooler feature contains the semantic information of the entire sentence, which aids in tasks like text classification. Hence, this feature is used to represent the joints. With M action categories and N number of joints, the feature C of all joints in an action is calculated.

Once this work have the textual prompt features, two feature graphs in the LA-GCN model that are GPR graph and CPR graph are constructed based on 2D coordinates of joints and confidence scores from AlphaPose. The LA-GCN model is then trained on the MUSIC dataset to recognize the instrumental actions. By removing the final fully connected layer, the model serves as an effective movement feature extractor. This framework extracts relevant features specific to each instrument-playing activity, improving the model's ability to separate sound sources.

3.4. Audio source separation

For audio processing, this work employ the short-time Fourier transform (STFT) [25] to transform the input waveform into a time-frequency spectrogram, enabling the analysis of its spectral characteristics over time. The STFT employs a 1,022-point analysis window and a hop size of 256 samples, allowing the preservation of fine-grained temporal and spectral details in the audio signal. This setup helps to capture changes in frequency and intensity within the signal, thereby supporting detailed audio analysis and processing.

A convolutional U-Net architecture is adopted as the backbone for source separation because of its strong capability in audio reconstruction tasks. The network first extracts latent audio representations from the time-frequency representation of the mixed audio through a seven-stage convolutional encoder with progressive down-sampling. Each stage employs a 4×4 convolutional filter, followed by batch normalization and a leaky rectified linear unit (LeakyReLU) activation with a negative slope of 0.2, while down-sampling the feature maps with a stride of 2. As the encoder proceeds, the spatial resolution is progressively reduced, whereas the feature dimension is expanded from 2 to 512 channels, enabling the network to learn rich spectral patterns and complex acoustic structures.

The instrument features and movement features are extracted by the instrument detector and movement detector, respectively. Then, these components are integrated through an instrument-movement fusion module. The representations from the two modalities are merged through feature concatenation to make the instrument-movement features and then passed through a dual-layer convolutional network for further processing. At the bottleneck stage, these extracted instrument-movement features are integrated with the audio features. Specifically, the instrument-movement features are replicated to match the spatial dimensions of the audio features (1, 512) and then concatenated along the channel dimension as illustrated in Figure 3, resulting in a unified representation with 1,024 channels.

The fused representation is subsequently fed into a seven-layer transposed convolutional decoder. Each decoding layer employs a 4×4 transposed convolution with a stride of 2, followed by batch normalization and a rectified linear unit (ReLU) activation, while the final layer applies a Sigmoid function. Throughout the decoding process, the feature maps are progressively upsampled as the channel dimensionality is gradually reduced to 2. Skip connections from the encoder are utilized at each decoding step, concatenating the corresponding encoded features to preserve fine-grained frequency details, thereby enabling accurate estimation of the spectrogram masks for each individual audio source.

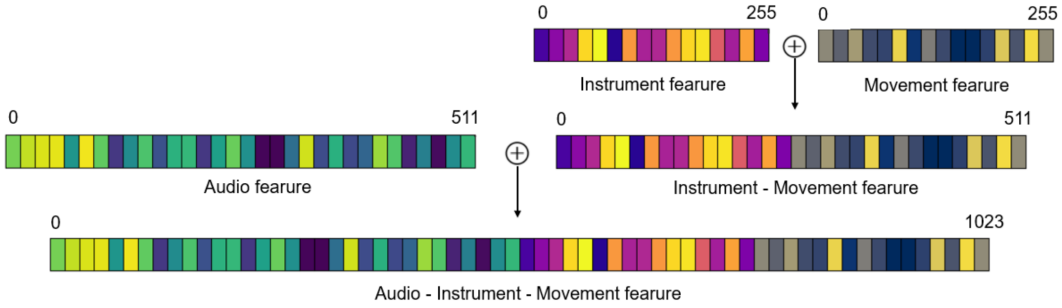


Figure 3. Audio-instrument-movement fusion diagram

To estimate the separated sound, the masks generated by the U-Net are multiplied by the time-frequency representation of the input audio mixture. The resulting spectrograms are then transformed back into the time domain using the inverse short-time Fourier transform (ISTFT), producing individual audio signals for each source. This approach allows the extraction of distinct audio components corresponding to different instrument-movement features.

3.5. Loss function formulation

As described in the previous section, the objective of proposed framework is to recover each constituent sound source from an input audio mixture associated with a given video clip. Consistent with the previous studies [7], [10], this work adopt the widely used mix-and-separate training strategy. Let $(V_i, s_i(t))_{i\in[1,N]}$ denote a collection of N video clips and their corresponding audio signals. Based on this set, a synthetic audio mixture is constructed as (2).

$$s(t) = \sum_{i=1}^{N} s_i(t) \tag{2}$$

The model is designed to recover the target audio signal $s_i(t)$ from the mixture $s(t)$ for the sound source associated with video V_i.

3.5.1. Loss functions for audio component

Starting from the mixed audio signal $s(t)$, the waveform is transformed into a time-frequency representation $S \in \mathbb{R}^{F\times T}$ using the STFT [25], where F and T denote the spectral and temporal dimensions, respectively. The resulting spectrogram is subsequently processed by the U-Net encoder to learn a compact latent representation. The decoder then progressively reconstructs the feature maps through successive upsampling operations, producing audio embeddings that are finally used to estimate the separated spectrogram mask $\hat{\mathcal{M}}_i$.

The network is trained by learning spectrogram masks. Specifically, the reference ratio mask corresponding to the i-th sound source, denoted by $\mathcal{M}_i$, is defined as (3).

$$\mathcal{M}_i(t,f) = \frac{S_i(t,f)}{S(t,f)} \tag{3}$$

Where (t, f) represents the corresponding time-frequency bin. The network parameters are optimized using the L_1 loss [26], and the resulting audio separation loss L_{A} is presented as (4).

$$L_{\text{A}} = \sum_{i=1}^{N} \left\| \hat{\mathcal{M}}_i - \mathcal{M}_i \right\|_1 \tag{4}$$

Where $\hat{\mathcal{M}}_i$, $\mathcal{M}_i$ are the predicted spectrogram mask and the ground truth spectrogram mask of the i-th sound source corresponding to the i-th video, respectively.

3.5.2. Loss functions for instrument-movement component

In addition to the audio, information about objects (i.e., instruments) and musicians' movements while playing different instruments provides valuable cues for sound source isolation. In addition, each instrument has a corresponding playing action. Thus, instruments and playing actions will correspond to each other in the same layer.

The instrument-movement-consistency loss, denoted by $L_{ins-mov}$, is designed to ensure that the separated spectrograms align with both the corresponding musical instrument and the musician's movements. This objective is optimized using the cross-entropy loss, as in (5).

$$L_{\text{ins-mov}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} -y_{i,c}, \log(p_{i,c}) \tag{5}$$

Here, N denotes the count of musical instruments appearing in the videos, whereas C corresponds to the overall set of instrument categories. The binary indicator $y_{i,c}$ specifies whether the c-th category matches the target class of the estimated spectrogram, while $p_{i,c}$ denotes the estimated probability assigned to category c.

The overall loss function is intended to optimize the spectral features of the separated sound while also ensuring consistency between the separated sound spectrograms, instrument categories, and movement categories. It is obtained as (6).

$$L_{\text{MIA3S}} = L_{\text{A}} + \lambda L_{\text{ins-mov}} \tag{6}$$

Where λ denotes the weighting coefficient for the instrument-movement-consistency loss L_{ViMo}. This coefficient controls the trade-off between audio separation performance and consistency with both instrument identity and performer movement.

4. EXPERIMENTAL RESULTS

4.1. Dataset

This study utilizes a subset of the MUSIC dataset [13], which was originally compiled from YouTube videos using keyword searches. This dataset encompasses 685 video clips of musical performances, including both solos and duets. The proposed research focuses on the 536 solo videos. This collection spans 11 distinct instrument categories: acoustic guitar, accordion, clarinet, cello, flute, erhu, saxophone, tuba, trumpet, xylophone, and violin. This study follows the same preprocessing and evaluation protocol as in [7]. The original videos are divided into 10-second segments. These segments were created with varying overlap ratios of 50%. This preprocessing resulted in a total of 9,760 10-second clips. Then, the dataset is divided into three sets: a training set of 9,260 clips, an evaluation set of 250 clips, and a testing set of 250 clips.

4.2. Evaluation metrics

Since the proposed framework exploits performer movement as an important cue for sound source separation, this work perform a series of experiments to assess the robustness of the movement feature extraction module as well as the overall separation performance. Therefore, two sets of evaluation metrics are employed:

i) Evaluation metrics of the movement recognition: for movement recognition, this work utilize precision, recall, and F1-score, which are widely used metrics to measure the accuracy of the LA-GCN model used in movement feature extraction.

ii) Evaluation metrics of sound source separation: to evaluate the quality of sound source separation, this work adopt three standard metrics from the source separation community [27] that are signal-to-distortion ratio (SDR), signal-to-artifact ratio (SAR), and signal-to-interference ratio (SIR). Expressed in decibels (dB), these metrics provide a comprehensive evaluation of separation quality, where larger values correspond to better performance.

4.3. Implementation details

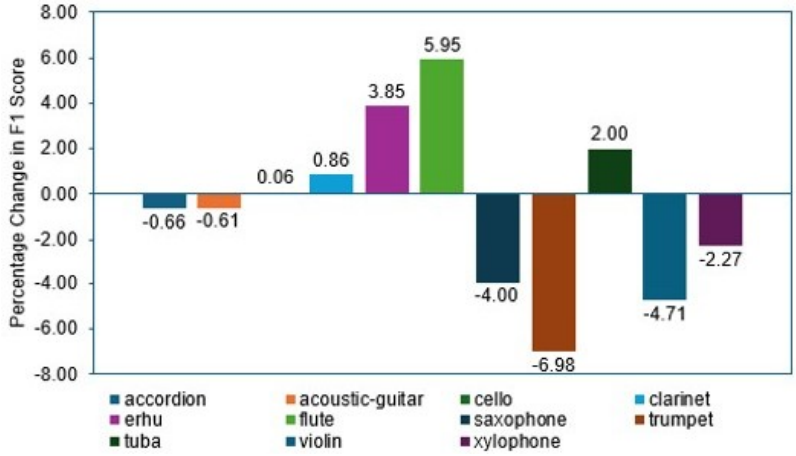
All experiments were performed on a workstation running Ubuntu 20.04, featuring two NVIDIA graphics processing units (GPUs): a GeForce GTX 1080 Ti with 10 GB memory and another with 8 GB memory. The proposed framework was implemented using PyTorch version 1.10.2. The model is trained with mini-batches of 8 samples using the adaptive moment estimation (Adam) optimizer. An initial learning rate of 1×10^{-4} is adopted for the entire network, whereas the visual representation module is optimized with a lower learning rate of 1×10^{-5}. This learning rate scheduling strategy enables more effective fine-tuning of the visual encoder. In the proposed loss function L_{MIA3S}, the weighting coefficient is set to $\lambda = 0.05$ for the instrument-movement consistency term.

4.4. Ablation study

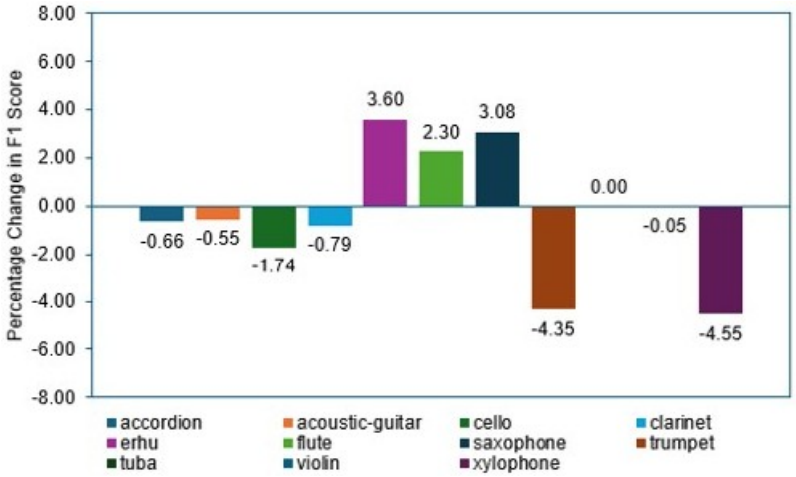
4.4.1. Evaluation of joints used for action representation

From a given video frame, 68 body joints can be estimated using the AlphaPose model. However, not all joints contribute equally to action representation. Therefore, in this study performs an evaluation to assess the relative importance of each joint in representing and recognizing actions. Three scenarios of the MIA3S are defined as follows: i) using only the upper body joints from the hips upwards, ii) using only the joints from the knees upwards, and iii) using all 68 joints.

The results of these scenarios are illustrated in Figure 4. As shown, using all 68 joints generally yields the best performance across most instrument-playing activities. An exception is observed in Figure 4(a) in the flute playing, where using only upper body joints from the hips upwards get better activity recognition results than using all 68 joints of the body. The reason is that lower-body move very little during this activity, so they do not contribute significantly to recognition accuracy. Interestingly, for playing the erhu and saxophone, using only the joints from the knees upward leads or using upper body joints both to slightly better results then all body joints, with improvements ranging from 2.08% to 3.6% when using only the joints from the knees upward leads, as shown in Figure 4(b), and from 2.00% to 3.85% when using upper body joints. This is likely because these instruments are typically played in a seated position, rendering lower-leg joints less informative for distinguishing the activity. Although there are exceptions, the overall experimental results point out that incorporating all joints enables the model to exploit richer spatial and temporal features, thereby enhancing the recognition of complex activities. As a result, all 68 joints are utilized in our subsequent experiments.



(a)



(b)

Figure 4. Comparison of F1-score ratios between scenarios using only joints from the: (a) upper body versus using all 68 joints and (b) knees up versus using all 68 joints

4.4.2. Evaluation of model used for movement feature extraction

Although many deep learning models can be used to extract the movement feature. In this study proposes to apply LA-GCN as this method is particularly powerful for activity recognition, where both visual and conceptual cues are needed for accurate classification. Table 1 compares the action recognition results of LA-GCN and the model spatio-temporal graph convolutional network (ST-GCN) used in [7] for 11 playing instrument activities: playing the acoustic guitar, accordion, clarinet, cello, flute, erhu, saxophone, tuba, trumpet, xylophone, and violin. The experimental results demonstrate that LA-GCN performs exceptionally well in classifying musical instrument-related activities. Most activity classes attain F1-scores between 0.94 and 1.00. Notably, the activity of playing the tuba has the lowest F1-score of 0.94, while activities like playing the guitar, accordion, and xylophone achieve an F1-score of 1.00. This demonstrates the model's ability to accurately recognize activities based on the human skeleton.

Compared to the ST-GCN model [7], the LA-GCN model demonstrates noticeable improvements in recognition performance across most activities, as detailed in Table 1. Specifically, the LA-GCN achieves higher F1-score in 10 out of 11 activities. For example, in the case of playing clarinet and erhu, the F1-scores increase by 0.09 and 0.08, respectively, indicating a significant enhancement in recognizing complex hand and upper body movements. Similarly, activities such as playing the cello, saxophone, and tuba also see improvements ranging from 0.03 to 0.06. The only exception is the activity of playing the flute, where the F1-score slightly decreases from 0.99 in ST-GCN to 0.97 in LA-GCN. This result can be explained by the

limited movement involved in playing the flute, particularly in the hand movements, where the additional joints utilized by LA-GCN provide minimal contribution.

Table 1. Comparison of action recognition performance between the proposed MIA3S framework and the ST-GCN model [7]

ID	Activity	LA-GCN model			ST-GCN model		
		Precision	Recall	F1-score	Precision	Recall	F1-score
0	Playing accordion	1.00	1.00	1.00	0.99	0.99	0.99
1	Playing acoustic-guitar	1.00	1.00	1.00	0.98	1.00	0.99
2	Playing cello	0.98	1.00	0.99	0.95	0.96	0.96
3	Playing clarinet	0.99	0.87	0.98	0.93	0.85	0.89
4	Playing erhu	1.00	0.94	0.97	0.82	0.98	0.89
5	Playing flute	0.98	0.96	0.97	1.00	0.98	0.99
6	Playing saxophone	0.98	0.92	0.95	0.95	0.79	0.86
7	Playing trumpet	0.92	1.00	0.96	0.95	0.95	0.95
8	Playing tuba	0.88	1.00	0.94	0.94	0.82	0.88
9	Playing violin	1.00	1.00	1.00	0.98	0.97	0.98
10	Playing xylophone	1.00	1.00	1.00	0.96	0.96	0.96

4.4.3. Evaluation of the role of components in the proposed method

This study conduct ablation experiments to evaluate the contribution of each component in the proposed method. Three scenarios are considered: i) audio-instrument, ii) audio-movement, and iii) audio-instrument-movement. In the first scenario, only audio and instrument features are used. In the second, the instrument features are removed, retaining only audio and movement information. The third scenario incorporates all three types of features for sound source separation.

The sound source separation results of three scenarios are presented in Table 2. From there, this study observe that the audio-instrument configuration outperforms the audio-movement setup in terms of both SDR and SAR, highlighting the effectiveness of using visual features of the instrument alongside audio cues to guide the U-Net model. This suggests that visual information related to the instrument provides more discriminative cues than movement features alone.

Table 2. Ablation study on the components of the MIA3S method

Test scenario	SDR [↑]	SIR [↑]	SAR [↑]
Audio-instrument	7.40	9.80	14.52
Audio-movement	6.93	10.73	11.90
Audio-instrument-movement	8.77	12.72	12.81

The audio-instrument-movement configuration yields the best performance, achieving an SDR of 8.77 dB and an SIR of 12.72 dB. Specifically, the SDR improves by 1.37 dB compared to the audio-instrument scenario and by 1.84 dB compared to the audio-movement scenario. This is also consistent with the qualitative evaluation given in Figure 5. The leftmost column presents the musical instruments identified in the corresponding video frames.

Columns 2 to 5 illustrate the spectrogram masks estimated by the proposed MIA3S model under three scenarios: i) audio-instrument, ii) audio-movement, and iii) audio-instrument-movement, together with the reference mask. Columns 6 to 9 display the corresponding reconstructed spectrograms for the same three settings alongside the reference spectrogram. This study observes that the predicted spectrograms of the MIA3S model have the highest similarity to the ground-truth spectrogram. These results confirm that combining audio, instrument appearance, and movement features yields the most effective results for sound source separation.

4.4.4. Comparison with the state-of-the-art methods

Table 3 presents the experimental results of our approach alongside those of representative state-of-the-art methods for sound source separation. This study observe that the MIA3S model achieves strong performance, with its SDR value exceeding that of most other methods-only slightly lower than the iQuery model based on a transformer architecture [19]. Compared to the method in [7], which also integrates audio, object, and movement information, the proposed approach improves the SDR by 3.35 dB. Likewise, the

SIR increases by 5.65 dB, rising from 7.07 to 12.72. However, the SAR experiences a decrease of 3.97 dB, dropping from 16.78 to 12.81.

These results indicate that the use of LA-GCN in our proposed method significantly enhances sound source separation, particularly by improving source clarity and reducing interference. Nonetheless, this comes with a modest trade-off in artifact suppression. Overall, the experimental results highlight the critical role of motion feature extraction in ASS, offering further gains in the perceptual quality of the separated audio.

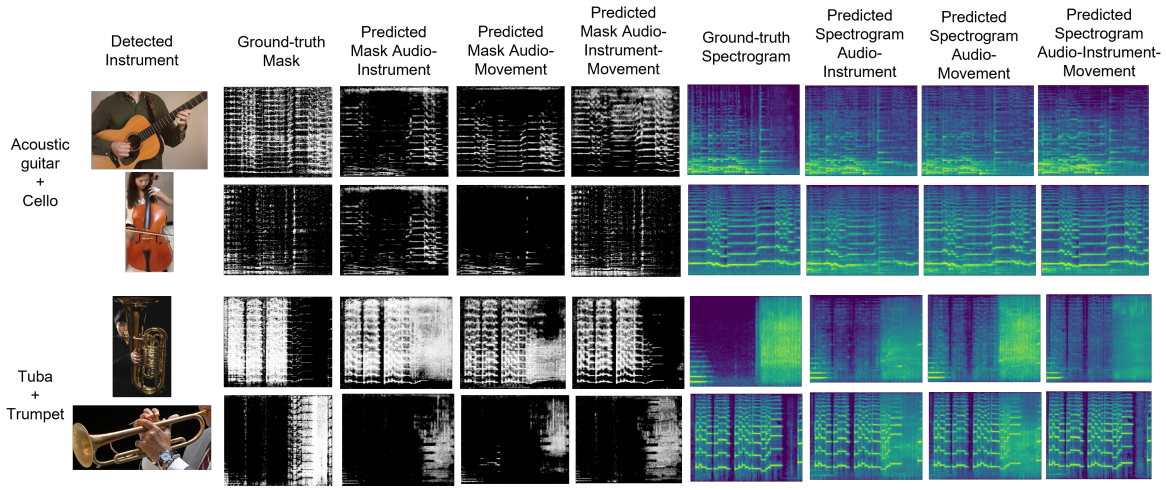


Figure 5. Qualitative comparison on two examples from the MUSIC evaluation set

Table 3. Experimental results of ASS on the MUSIC dataset

Model	Feature	SDR [↑]	SIR [↑]	SAR [↑]
NMF-MFCC [28]	Audio	0.92	5.68	6.84
Sound-of-pixels [13]	Audio and object	4.23	9.39	9.85
Co-separation [17]		6.54	11.37	9.46
Cyclic co-learning [29]		7.74	13.22	11.54
iQuery [19]		11.17	15.84	14.27
Model in [7]	Audio, object, and movement	5.42	7.07	16.78
Proposed MIA3S model		8.77	12.72	12.81

5. CONCLUSION

In this work introduces a novel approach to ASS MIA3S that fuses audio, instrument appearance, and language-guided movement representations. By employing the LA-GCN model, our approach explicitly models the semantic relationship between musician movements and sound production using natural language prompts. The proposed fusion mechanism at the bottleneck of the U-Net architecture, combining instrument-movement features with encoded audio representations, demonstrates strong synergy among modalities. Empirical evaluations on the MUSIC dataset demonstrated that our proposed model outperforms existing approaches across multiple evaluation metrics, including SDR, SIR, and F1-score. The ablation studies further confirmed the importance of integrating both instrument and movement cues for fine-grained source separation. Despite these promising results, several directions for future work remain. First, while this study utilized a fixed set of textual prompts, adapting these prompts dynamically based on context or user interaction could further align system behavior with human intent, as well as further enhance performance. Second, incorporating 3D pose estimation or depth-aware motion features might improve robustness in complex scenes involving occlusions or multiple overlapping performers. Lastly, extending the framework to support open-set scenarios and real-time processing would make it more practical for deployment in live music or multimedia applications.

ACKNOWLEDGMENTS

The authors would like to thank the Artificial Intelligence Research and Applications Center (AIC), School of Information and Communications Technology, Hanoi University of Industry, for their support during this research.

FUNDING INFORMATION

This research is funded by the Vietnam Ministry of Education and Training under grant number B2023-MDA-02.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hien-Thanh Duong	✓	✓			✓			✓		✓	✓		✓	✓
Huu-Mao Dao		✓	✓			✓		✓	✓	✓	✓			
The-Thanh-Dat Le	✓		✓	✓			✓			✓	✓			
Phuong-Bac Nguyen				✓		✓		✓		✓				
Thi-Lan Le	✓	✓			✓	✓			✓	✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

This section is not applicable, as this study did not involve human participants or personally identifiable information.

ETHICAL APPROVAL

This section is not applicable, as this study used only non-sensitive computational or public data.

DATA AVAILABILITY

The data that support the findings of this study are openly available in GitHub at https://github.com/roudimit/MUSIC_dataset.

REFERENCES

[1] S. Makino, S. Araki, R. Mukai, and H. Sawada, "Audio source separation based on independent component analysis," in *2004 IEEE International Symposium on Circuits and Systems*, pp. V-668-V–671, doi: 10.1109/ISCAS.2004.1329896.

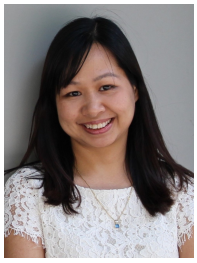
[2] A. Ozerov, C. Févotte, and E. Vincent, "An introduction to multichannel NMF for audio source separation," in *Audio Source Separation*, Cham: Springer, 2018, pp. 73–94, doi: 10.1007/978-3-319-73031-8_4.

[3] C. Févotte, E. Vincent, and A. Ozerov, "Single-channel audio source separation with NMF: divergences, constraints and algorithms," in *Audio Source Separation*, Cham, Switzerland: Springer, 2018, pp. 1–24, doi: 10.1007/978-3-319-73031-8_1.

[4] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: a multi-scale neural network for end-to-end audio source separation," in *19th International Society for Music Information Retrieval Conference*, 2018, pp. 334–340.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019, doi: 10.1109/TASLP.2019.2915167.
[6] Z. Song and Z. Zhang, "Visually guided sound source separation with audio-visual predictive coding," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 11, pp. 15528–15542, Nov. 2024, doi: 10.1109/TNNLS.2023.3288022.
[7] T. T.-H. Duong, T.-H. Nguyen, T. T.-D. Le, T.-L. Nghiem, D.-H. Pham, and T.-L. Le, "Investigating the role of human action detector in visual-guide audio source separation system," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Oct. 2023, pp. 1296–1303, doi: 10.1109/APSIPAASC58517.2023.10317504.
[8] X. Liu *et al.*, "Separate what you describe: language-queried audio source separation," in *Interspeech 2022*, Sep. 2022, pp. 1801–1805, doi: 10.21437/Interspeech.2022-10894.
[9] R. Tan *et al.*, "Language-guided audio-visual source separation via trimodal consistency," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 10575–10584, doi: 10.1109/CVPR52729.2023.01019.
[10] D.-H. Pham, Q.-A. Do, T. T.-H. Duong, T.-L. Le, and P.-L. Nguyen, "End-to-end visual-guided audio source separation with enhanced losses," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Nov. 2022, pp. 2022–2028, doi: 10.23919/APSIPAASC55919.2022.9980162.
[11] C. Rohlfing, J. M. Becker, and M. Wien, "NMF-based informed source separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 474–478, doi: 10.1109/ICASSP.2016.7471720.
[12] C. Zhang, F. Yang, and J. Yang, "Basis clustering regularized NMF for multichannel blind source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 153–168, 2026, doi: 10.1109/TASLPRO.2025.3642556.
[13] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Computer Vision – ECCV 2018*, Cham, Switzerland: Springer, 2018, pp. 587–604, doi: 10.1007/978-3-030-01246-5_35.
[14] A. Ephrat *et al.*, "Looking to listen at the cocktail party," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, Aug. 2018, doi: 10.1145/3197517.3201357.
[15] L. Zhu and E. Rahtu, "Visually guided sound source separation using cascaded opponent filter network," in *Computer Vision – ACCV 2020*, Cham: Springer, 2021, pp. 409–426, doi: 10.1007/978-3-030-69544-6_25.
[16] T. Rahman, M. Yang, and L. Sigal, "TriBERT: full-body human-centric audio-visual representation learning for visual sound separation," in *35th International Conference on Neural Information Processing Systems*, 2021, pp. 9774–9787.
[17] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 3878–3887, doi: 10.1109/ICCV.2019.00398.
[18] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *Computer Vision – ECCV 2020*, Cham: Springer, 2020, pp. 208–224, doi: 10.1007/978-3-030-58523-5_13.
[19] J. Chen, R. Zhang, D. Lian, J. Yang, Z. Zeng, and J. Shi, "iQuery: instruments as queries for audio-visual sound separation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 14675–14686, doi: 10.1109/CVPR52729.2023.01410.
[20] H.-W. Dong, N. Takahashi, Y. Mitsufuji, J. McAuley, and T. B.-Kirkpatrick, "CLIPSep: learning text-queried sound separation with noisy unlabeled videos," in *International Conference on Learning Representations (ICLR) 2023*, 2023.
[21] X. Liu *et al.*, "Separate anything you describe," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 458–471, 2025, doi: 10.1109/TASLP.2024.3520017.
[22] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training," in *36th International Conference on Neural Information Processing Systems*, 2022, pp. 10078–10093.
[23] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: regional multi-person pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2353–2362, doi: 10.1109/ICCV.2017.256.
[24] H. Xu, Y. Gao, Z. Hui, J. Li, and X. Gao, "Language knowledge-assisted representation learning for skeleton-based action recognition," *IEEE Transactions on Multimedia*, vol. 27, pp. 5784–5799, 2025, doi: 10.1109/TMM.2025.3543034.
[25] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984, doi: 10.1109/TASSP.1984.1164317.
[26] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, Mar. 2017, doi: 10.1109/TCI.2016.2644865.
[27] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006, doi: 10.1109/TSA.2005.858005.
[28] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proceeding of International Conference on Digital Audio Effects DAFx '09*, 2009.
[29] Y. Tian, D. Hu, and C. Xu, "Cyclic co-learning of sounding object visual grounding and sound separation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2021, pp. 2744–2753, doi: 10.1109/CVPR46437.2021.00277.

BIOGRAPHIES OF AUTHORS



Hien-Thanh Duong holds a doctor of Computer Science from the Hanoi University of Science and Technology, Vietnam, in 2019. She also received her bachelor's degree in Mathematics and Information Technology from Hanoi National University of Education, Vietnam, in 2000, and her master's degree in Computer Science from Hanoi University of Science and Technology in 2009. She is currently a senior lecturer and researcher with the Hanoi University of Industry, Vietnam. Her research interests include signal processing and AI/ML applied to audio, image, and video analysis. She can be contacted at email: thanh.duong@haui.edu.vn.



Huu-Mao Dao is an undergraduate student majoring in Electronics and Telecommunications Engineering at the Hanoi University of Science and Technology (HUST), Vietnam. His research interests include multimodal machine learning, audio source separation, and signal processing. He can be contacted at email: mao.dh233865@sis.hust.edu.vn.



The-Thanh-Dat Le received the B.E. degree in Electronics and Telecommunications from the Hanoi University of Science and Technology (HUST), Vietnam. He is currently working at ITOKI Corporation in Japan and has been part of the company's DX Promotion Headquarters since April 2025. His research interests include multimodal machine learning, sound source separation, and signal processing. He can be contacted at email: dat.ltt2311@gmail.com.



Phuong-Bac Nguyen received the B.Eng. degree in Information Technology from Hanoi University of Science and Technology in 2005 and the M.Sc. degree in Information Technology from the same university in 2019. She is currently a lecturer in the Department of Computer Science at Hanoi University of Mining and Geology. Her research interests include signal processing, data mining, machine learning, and the internet of things (IoT). She can be contacted at email: nguyenthiphuongbac@humg.edu.vn.



Thi-Lan Le received the degree in Information Technology and the M.S. degree in Signal Processing and Communication from Hanoi University of Science and Technology (HUST), Vietnam, and the Ph.D. degree in video retrieval from INRIA, Sophia Antipolis, France, in 2009. She is currently an associate professor with the School of Electrical and Electronic Engineering (SEEE), Hanoi University of Science and Technology. Her research interests include image processing, computer vision, content-based indexing and retrieval, video understanding, and human–robot interaction. She can be contacted at email: lan.lethi1@hust.edu.vn.