

A Visually Guided Sound Source Separation and Localization System for Musical Instrument Videos

Seyedeh Sogand Hashemi

Yazd University

Masoudreza Aghabozorgi Sahaf

aghabozorgi@yazd.ac.ir

Yazd University

Mohammadreza Asadi

Yazd University

Article

Keywords: Audio-visual learning, Persian musical instrument detection, single-channel sound source separation, visual sound source localization, YOLO

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10721104/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Visually Guided Sound Source Separation and Localization System for Musical Instrument Videos

Seyedeh Sogand Hashemi¹, Masoudreza Aghabozorgi Sahaf^{1*}, Mohammadreza Asadi¹

¹Electrical Engineering Department, Yazd University, Yazd, Iran

*Corresponding Author Email: aghabozorgi@yazd.ac.ir

Abstract— Audio-visual sound source separation aims to recover individual sound signals from a mixture by exploiting information from both the audio and visual modalities. This paper presents a modular, visually guided deep learning framework for detecting, localizing, and separating musical instruments in single channel real world videos. The proposed system consists of two sequential components. First, a YOLOv8s model detects the classes and spatial locations of the instruments in video frames. The detections obtained from multiple frames are then aggregated to identify the instruments present in the video and select the corresponding instrument-specific separation networks. Each separation network is a fully connected neural network that independently estimates a binary time-frequency mask for each spectrogram frame. The dataset contains 358 annotated images and 420 audio signals extracted from videos covering 12 Persian and international musical instrument classes. Training and test samples were separated at the original video level to prevent data leakage. On the held-out visual test set, the instrument detection model achieved an mAP@0.5 of 63.9%, a precision of 75.1%, a recall of 59.2%, and an F1-score of 66.2%. Across the 12 instrument classes, the separation models achieved average SDR, SIR, and SAR values of 8.75, 13.42, and 7.24 dB, respectively. In the AudioSet-based comparison setting, an SDR of 10.97 dB was obtained. These results demonstrate the feasibility of using visual instrument detections to guide class-specific sound separation, particularly for underrepresented Persian musical instruments.

Index Terms— Audio-visual learning, Persian musical instrument detection, single-channel sound source separation, visual sound source localization, YOLO.

I. INTRODUCTION

Sound source separation and localization are fundamental components of audio-visual scene understanding, with applications in music production, audio editing, noise reduction, multimedia retrieval, and human-computer interaction. Sound source separation aims to recover individual source signals from an observed mixture, whereas source localization identifies the corresponding sound producing objects within a scene. These tasks are particularly challenging in musical performances, where multiple instruments may produce overlapping harmonic and temporal patterns. The

problem becomes more difficult in single-channel recordings obtained from real-world videos because the audio signal alone may not provide sufficient information to distinguish the contributing sources. Visual information can reduce this ambiguity by identifying the instruments present in a scene and providing information about their appearance, position, and motion. Consequently, visually guided sound source separation has emerged as an effective approach for associating components of an audio mixture with their corresponding visual objects. Sound of Pixels demonstrated that visual object representations could guide the separation of individual sound sources from mixed audio [1]. Co-Separation subsequently exploited the co-occurrence of visually similar objects across videos to improve audio source separation [2]. Visual scene representations and structured relationships among detected objects have also been investigated to provide additional semantic guidance for separation [3].

Subsequent studies incorporated more detailed visual and motion information. Music Gesture modeled the body and hand movements of musicians to improve the separation of both different and identical instrument types [4]. Zhu and Rahtu proposed a two-stage appearance-and-motion framework that learns sound-related motion representations for visual sound separation and localization [5]. SeCo introduced audio-consistency guidance to separate previously unknown musical visual sounds without requiring source-level annotations [6]. In addition, iQuery reformulated musical instruments as learnable queries to improve cross-modal consistency and facilitate generalization to additional instrument categories [7]. Location-guided cross-modal attention [8] have further demonstrated the value of spatial and semantic visual information in multi-source scenes. More recent research has increasingly investigated unified audio-visual architectures. OneAVM jointly addresses sound source localization, separation, and recognition using a shared audio-visual encoder and task-specific decoders, demonstrating that the three tasks can benefit from joint learning [9]. Audio-Visual Predictive Coding integrates audio feature extraction, multimodal feature interaction, and separation mask prediction within a predictive coding

Seyedeh Sogand Hashemi was with Yazd University, Yazd, Iran. She is now a Ph.D. Researcher at KU Leuven, Belgium (e-mail: hashemisogand@stu.yazd.ac.ir). Mohammadreza Asadi Gandomani was with Yazd University, Yazd, Iran. He is now a Research Engineer with Flanders Make, Belgium (e-mail:

mohammadreza.asadi@stu.yazd.ac.ir).

*Masoudreza Aghabozorgi is with the Electrical Engineering Department, Yazd University, Yazd, Iran (aghabozorgi@yazd.ac.ir).

architecture [10]. Continual Audio-Visual Sound Separation further addresses the challenge of learning new sound classes while preserving the model’s performance on previously learned categories [11]. These studies indicate a transition from separately designed audio and visual components toward more integrated and adaptable audio-visual learning frameworks.

Recently reported advances have extended this trend toward joint detection, open-vocabulary separation, and detailed scene understanding. See2Hear jointly performs visual object detection and sound source separation using an end-to-end transformer-based architecture and dynamically filters the detected object queries before audio-visual fusion [12]. OmniSep supports sound separation using text, image, audio, and composed multimodal queries within a unified query-based framework [13]. For localization, an object aware approach was introduced to distinguish active sound producing objects from visually similar but silent objects in complex single- and multi-source scenes [14]. These methods illustrate the growing importance of object level semantics and flexible multimodal guidance in sound source analysis. Several studies have also explored generative and foundation-model-based approaches. A visually guided generative framework based on diffusion and flow-matching models has been proposed to improve sound separation across diverse source categories [15]. MMAudioSep adapts a pretrained video-to-audio generative model to video- and text-queried sound separation, enabling knowledge acquired during audio generation to be transferred to the separation task [16]. The USE framework combines universal sound separation and target sound extraction and interprets different acoustic, semantic, and visual clues within a unified model [17]. These developments represent a shift from conventional discriminative mask estimation toward more flexible generative sound separation systems.

Despite this progress, several challenges remain. Many existing methods are evaluated primarily on established benchmark datasets and commonly represented instrument categories. Their application to domain-specific instruments and real-world videos with variations in illumination, viewpoint, recording quality, background interference, and acoustic conditions remains challenging. Persian musical instruments are particularly underrepresented in commonly used audio-visual datasets. Previous studies related to Persian music have mainly focused on genre classification [18] and Iranian modal music recognition [19]. In contrast, limited research has investigated the combined visual detection and audio separation of Persian musical instruments in multi-instrument videos. To address this gap, this study presents a practical audio-visual deep learning system for the detection, localization, classification, and separation of musical instruments in real-world videos. The proposed system extends our previous studies on Persian musical source separation and audio-visual noise reduction [20], [21]. It consists of two sequential deep learning components. First, a YOLOv8-based visual model detects the instrument classes and their spatial locations within the video frames. Predictions obtained from multiple frames are aggregated to identify the instruments present in the video. The detected classes then guide the

selection of the corresponding instrument-specific SoundNet models. Each SoundNet estimates a binary time-frequency mask from the mixed audio spectrogram and uses it to reconstruct the signal associated with the selected instrument.

Unlike fully unified end-to-end audio-visual architectures, the proposed method follows a modular and visually informed design. The visual component determines the number and types of instruments present and reduces the separation search space by selecting only the relevant instrument-specific models. The audio component subsequently performs frame-wise binary-mask estimation for the selected instruments. This modular structure allows the visual detection and audio separation components to be developed, trained, and evaluated independently while maintaining a functional relationship between the two modalities. The dataset used in this study comprises real-world videos from 12 Persian and international musical-instrument classes recorded under diverse visual and acoustic conditions. The principal contributions of this research are: 1) the development of a modular system that associates detected visual instruments with their corresponding separated audio signals; 2) the inclusion of underrepresented Persian musical instruments alongside international instruments; 3) the use of visual detections to guide the selection of instrument-specific sound separation models; and 4) the evaluation of both visual detection and audio separation performance in realistic single- and multi-instrument scenarios.

The remainder of this paper is organized as follows. Section 2 describes the signal processing principles and neural network architectures used in the proposed method. Section 3 presents the dataset, preprocessing, and implementation details. Section 4 reports and discusses the visual detection and sound separation results. Finally, Section 5 summarizes the conclusions and potential directions for future research.

II. METHODOLOGY

In audio-visual data, the term “audio source localization” refers to the process of recognizing the specific regions and pixels responsible for generating an audio signal. While many studies typically focus on a single active sound source, often a speaker or a musical instrument, real-world scenarios can involve multiple sound sources of diverse origins, including human and nonhuman sources such as speech, music, vehicles, and animal sounds [22].

The problem of single-channel audio source separation involves taking a single-channel audio signal, which is essentially a mixture of different sound sources, and then breaking it down into individual sound sources. In this context, we work with a finite and discrete mixed signal in the time domain, denoted as $s_{mix} \in \mathbb{R}^{N_t}$, which is a two-channel signal represented as an array with $N_t \times 2$ samples [20]. This mixed signal is essentially a combination of various single-source audio signals, i.e.:

$$s_{mix} = \sum_{i=1}^{N_s} s_i, \quad (1)$$

where $s_1, s_2, \dots, s_{N_s} \in \mathbb{R}^{N_t}$, each containing N_t samples per

channel, corresponds to individual sources contributing to the mixed audio signal. N_s is the total number of single sound sources, and τ denotes the sample period. Each of these audio signals has a duration of 10 s and is sampled at a rate of 44.1 kHz.

The sound separation algorithm should possess the capability to accurately estimate the individual sources $s_1, s_2, \dots, s_{N_s}$ solely from an analysis of the combined signal (s_{mix}). For instance, as illustrated in Fig. 1, the algorithm should be able to identify and separate three distinct musical instruments as sound sources from the mixed signal. It is noteworthy that the complexity of this challenge scales with the number of sound sources. Importantly, we work under the assumption that all individual and combined signals are free of external noise and are in their pristine, undistorted form.

To separate a mixed sound signal into its two single sources, we follow the general block diagram shown in Fig. 2 as follows:

- First, we separate the frames and audio signal of the test video.
- In the second step, we start with the mixed audio signal in the time domain (s_{mix}) and transform it to the time–frequency domain (T-F) using the STFT converter (S_{mix}).
- In the third step, we predict the locations and labels of the active audio sources in video frames using an object detection model.
- In the next step, by considering the former predicted results, we separate the estimation of the spectrograms of individual sound signals in the T–F domain by special deep neural network models.

Finally, we reconstruct the separated single signals from the T-F domain to the time domain using inverted STFT (ISTFT). Subsequently, we combine each of the separated sound signals with video frames, in which the name and location of that source are specified and produce new videos for each sound.

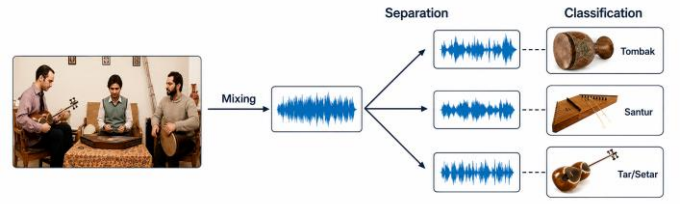


Fig. 1. Separation of three distinct sound sources from a mixed audio signal

A. Time-Frequency Analysis

An audio space containing different audio sources can be represented by a set of audio attributes related to the audio sources; therefore, various processes can be performed on the basis of these attributes. Although audio signals are originally in the time domain, processing in the frequency domain leads to better feature analysis and better results. Moreover, audio signal processing in the time–frequency domain provides more comprehensive information about both time and frequency attributes. As a result, in this paper, we implement short-time Fourier transform (STFT) as a time–frequency (TF) transformer because this method is efficiently executed in audio signal separation and masking applications.

STFT is a powerful processing tool for converting a signal into its constituent frequency components in specific quantized time bins. Thus, sound changes can be observed and analyzed at different time–frequency intervals. Furthermore, the performance of STFT in the signal separation process is almost similar to human beings’ ear function and close to reality. Considering the effect of frequency changes on TF values in STFT of a signal, we mainly use STFT to identify the dominant frequency ranges of each time interval [23].

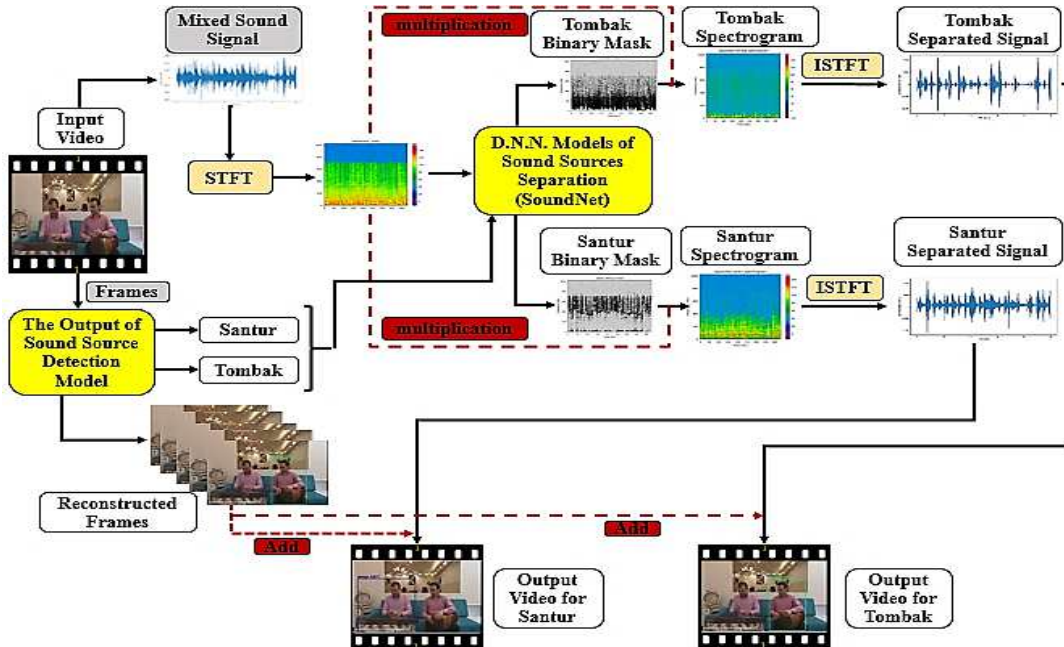


Fig. 2. Block diagram of the proposed method

According to the DFT relation, we define the discrete STFT and Invers STFT (ISTFT) relation as:

$$X_{km} = \sum_{n=-\infty}^{\infty} [x_n w_{n-mR}] e^{-j\omega_k n}, \quad (2)$$

and

$$x_n = \frac{1}{N} \sum_{m=-\infty}^{\infty} \sum_{k=0}^{N-1} X_{km} e^{j\omega_k n}, \quad (3)$$

where x_n represents the n^{th} component of an input sound vector x and w_{n-mR} defines the $n - mR^{th}$ component of the window vector w . X_{km} represents the $(k, m)^{th}$ component of the complex mixed matrix X in time-frequency domain, M is window length of samples, N is DFT size of samples, R is value of hop in samples, $\omega_k = 2\pi k/N$, and $k = 0, 1, 2, \dots, N-1$. Considering the conjugate symmetry in DFT for real input signals, only $N_f = \frac{N}{2} + 1$ first rows of the final STFT output are stored [23].

Using STFT, we can determine which frequencies are present at any given time. To avoid frequency perturbation or spectral leakage, we use the signal windowing technique and Hanning window. Each part of the signal is multiplied by the Hanning window function $w(t)$ before the Fourier transform. Moreover, time resolution and frequency resolution must be in balance; for instance, a larger FFT frame size leads to more frequency resolution but less time resolution. To achieve better time resolution, we apply the consecutive overlapping windows solution, providing better time resolution for short instant events [20]. In the following, we use STFT to obtain and display an image of the sound signal energy. For this purpose, we compute the magnitude of the matrix $V = |X|^2$ in the T-F domain as the spectrogram of signal; next, we display the result using the logarithmic scaling method as follows:

$$V_{(dB)} = 10 \log_{10} V, \quad (4)$$

where $V_{(dB)}$ denotes the logarithmic scale of the spectrogram. This sound analysis technique can comprehend sound signals almost similar to human ears [20].

B. Mask Filtering

The windowed-disjoint orthogonal (WDO) hypothesis has led to the emergence of sound source separation methods based on binary masks, one of the common methods for separating a desired time-frequency interval from the spectrum of a mixed

signal [20]. Another mask, called a soft mask, is equal to the ratio of a single sound signal spectrogram to a mixed sound signal spectrogram; thereafter, a binary mask is obtained by thresholding the soft mask. Equation (5) and (6) show the soft mask and binary mask formulas, respectively:

$$SM_{s_j} = \frac{V_{s_j(dB)}}{\sum_{i=1}^{N_s} V_{s_i(dB)} + \varepsilon}, \quad (5)$$

and

$$BM_{s_j}(k, m) = \begin{cases} 1 & SM_{s_j}(k, m) \geq 0.5 \\ 0 & \text{O.w.} \end{cases}, \quad (6)$$

where SM_{s_j} is a soft mask of the source s_j with size of $N_f \times N_t$ and values range $[0, 1]$, $V_{s_j(dB)}$ represents the spectrogram of single source s_j and N_s is the number of sources present in the mixed signal. As a guarantee for the non-zero denominator in (5), we add a very small positive value $\varepsilon = 2^{-52}$ to the dominator. In addition, we assume 0.5 as the threshold value for calculating binary masks. In (6) the expression $SM_{s_j}(k, m)$ is equal to the $(k, m)^{th}$ time-frequency interval in the soft mask of the sound source s_j , and $BM_{s_j}(k, m)$ shows the corresponding interval in the binary mask. Fig. 3 exhibits soft and binary masks corresponding to the Guitar and Violin.

Although real audio signals do not have the complete orthogonality specification of the WDO hypothesis, the assumption that the signals are WDO is still applicable to the implementation of the separation algorithm and has good results [20]. As a result, we use appropriate masks based on our application to separate the audio signals from different sources. In this method, multiplying the associated mask by the mixed signal results in the spectrogram of each single sound source, i.e.:

$$\hat{V}_{s_j(dB)} = V_{s_{mix}(dB)} BM_{s_j}, \quad (7)$$

where $\hat{V}_{s_j(dB)}$ is the estimated spectrogram for the s_j source, and $V_{s_{mix}(dB)}$ and BM_{s_j} are the spectrogram of the mixed signal and the estimated mask for the s_j source, respectively.

C. Deep Neural Networks

Deep Neural Networks (DNNs) are a type of Artificial Neural Network (ANN) with multiple layers between the input

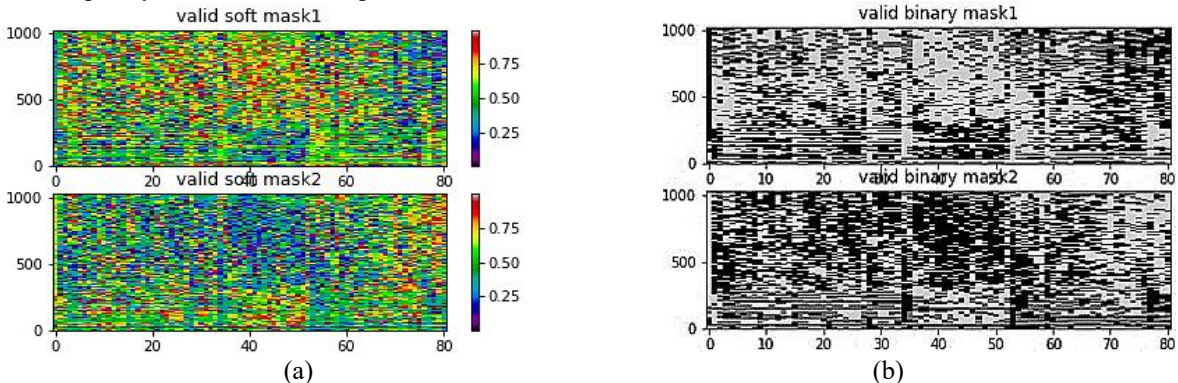


Fig. 3. (a) Soft masks and (b) binary masks of Guitar (up) and Violin (down)

and output. A Convolutional Neural Network (CNN) is a type of multilayer neural network designed for two-dimensional data such as images and computer vision applications. In general, a convolutional neural network is a hierarchical neural network consisting of three main layers: Convolution, Pooling, and Fully Connected Layer (FC). Layers such as Batch_Normalization or Dropout can also be added to the model to improve the performance of the network.

D. Object Detection

Object detection is defined as the localization and recognition of an object with a specific class within an image. In general, object detection methods are divided into traditional and modern categories. Traditional methods, which are called two-stage methods, use the segmentation technique and object classification (e.g., Faster R-CNN and Mask R-CNN methods). In modern methods, which are called one-stage methods, the machine sees the input image only once and extracts the necessary features of the image using a combination of deep convolutional and fully connected networks (e.g., YOLO and SSD methods). Modern methods have higher accuracy and speed than traditional methods.

In this study, we use YOLOv8 [24] model to detect instruments within a video. This can help our sound separation system to localize, recognize, and separate the sound of each audio source from the mixed signal more accurately.

E. YOLOv8 Architecture

The You Only Look Once version 8 (YOLOv8) model stands as a seminal advancement in the field of computer vision, particularly in object detection tasks. At its foundation, YOLOv8 employs a deep CNN structure with multiple layers, including convolutional, pooling, and residual blocks. It revolutionizes object detection applying a unified neural network to predict bounding boxes and class probabilities simultaneously. This unique approach allows YOLOv8 to operate with remarkable speed while preserving a high level of precision in object detection [24].

Moreover, YOLOv8 demonstrates a refined understanding of feature representation through the integration of advanced techniques such as feature pyramid networks (FPNs) and focal loss. FPNs facilitate multi-scale feature extraction, enabling the model to capture objects of various sizes more effectively by aggregating features from different network layers. The focal loss function aids in prioritizing challenging samples during training, effectively addressing the issue of imbalanced datasets, and enhancing the model's ability to focus on hard-to-detect objects. YOLOv8 adopts an anchor-free approach, predicting the object's center directly instead of estimating the offset from predefined anchor boxes. These processes contribute significantly to YOLOv8's ability to generalize well across diverse datasets and exhibit robust performance in real-world scenarios [24].

YOLOv8 comes in multiple variants (YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x), offering a range of trade-offs between accuracy and speed. YOLOv8x is the most accurate but slower, while YOLOv8 Nano is the fastest and smallest among them.

Fig. 4 is a summarized diagram of the YOLOv8 architecture [24], where “f” and “e” illustrates the total features number and the growth rate, respectively. CBS block involves a Convolution, a BatchNorm, and a SiLU layer.

F. Sound source separation

Our sound source separation problem is defined as follows:

“We have a mixed signal consisting of two or more audio signals from different musical instruments. First, we want to determine the signals of which audio sources are present in the mixed signal; second, we want to separate the audio of different sound sources.”

Regarding this situation, two challenges must be considered:

- 1) Identification of the names of the present and active musical instruments;
- 2) Sound separation of identified musical instruments from the mixed signal.

After passing the first challenge and considering its results, we face a sound source separation problem as the second challenge. The proposed solution is to design a deep neural network model to separate the audio signals of the sources in the video. This network is the second deep neural network used in the proposed algorithm. We name this network "SoundNet" which is shown in Fig. 5.

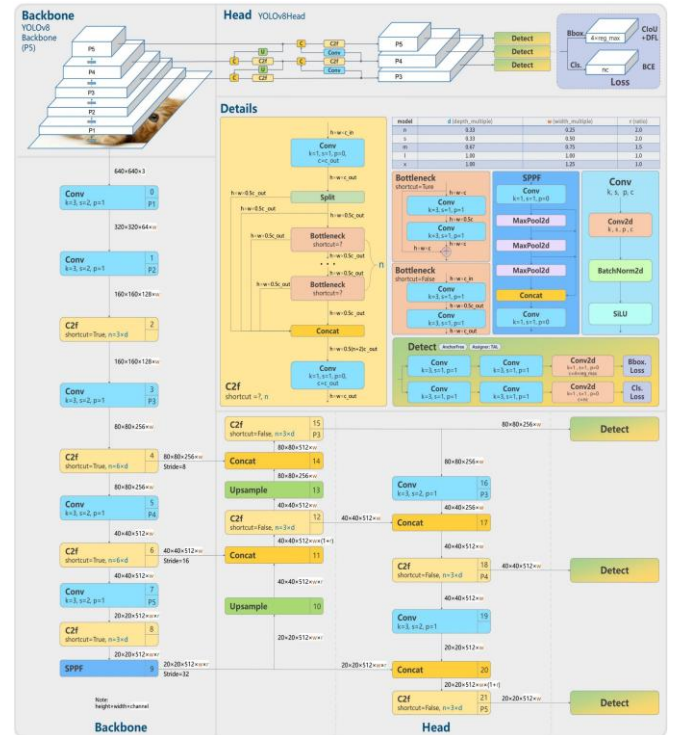


Fig. 4. Architecture of YOLOv8 [24]

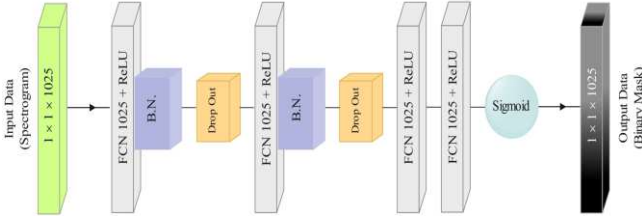


Fig. 5. The SoundNet network

The results of the SoundNet model are separated audio signals of each detected sound source. SoundNet consists of dense layers with ReLU activation functions, dropout layers, and batch_normalization layers. The batch_normalization layer helps the network achieve faster convergence, less training time, and prevention of model saturation. In addition, dropout layers added to the first two layers of the model prevent the overfitting problem and compensate for the lack of training data. The activation function of the last layer is the sigmoid function.

Data dimensions in all layers is $1 \times 1 \times 1025$. To train the SoundNet model, we use the Adam optimizer function with a learning rate of 0.0001. To evaluate the model, the binary_crossentropy loss function and accuracy criterion are considered. Model trains on the Google Colab environment and takes about 59 min.

III. ALGORITHM IMPLEMENTATION

In this section, we describe the practical issues of the proposed sound source separation and localization method in detail. All algorithms are written in Python using PyTorch and TensorFlow libraries. As mentioned before, the algorithm implementation, including the design and training of deep neural networks, is executed on the Google Colab environment (NVIDIA T4 GPU).

In this paper, the main structure consists of two individual deep neural networks; the network 1 is trained for the sound source detection in video frames, and the network 2 is responsible for the sound signal separation task.

A. Dataset

The dataset for this study comprises audio-visual (video) data categorized into 12 classes of audio sources. These sources include Persian musical instruments as well as international ones, specifically Daf, Kamancheh, Santur, Setar, Tombak, Accordion, Guitar, Clarinet, Piano, Flute, Violin, and Tuba. The data was collected from publicly available videos and images on the Internet [25]. For the international instruments, we also utilized some videos of the “Audioset” [26] dataset, which includes 10-second real-world video clips arranged in a wide range of sound sources, making it one of the most comprehensive video datasets globally.

The dataset used for training the visual model consists of 358 labeled images (split 80/20 for training and test), manually annotated to ensure high-quality ground truth labels. Each image is annotated with bounding boxes around the musical

instruments using the YOLO format. For the audio model, sound signals were extracted from 420 real-world videos (split 85/15 for training and testing) sourced from both the Internet and the Audioset dataset. No samples originating from the same source video were shared between the two subsets.

B. Data Preprocessing

First, we clip 10s parts of each video in which audio sources are present and played. Next, we reduce the frame rate of clipped videos to 10 fps because a lower frame rate leads to easier data processing for the processor. Next, we extract the audio signal and frames of each clipped video; as a result, we obtain a 10-s audio signal and approximately 100 images from each video. The sampling frequency of audio signals is 44.1 kHz, so an audio signal is a 2-channel 441263×2 signal.

- 1) **STFT:** To transfer an audio signal to the T-F domain, we normalize it and then apply STFT on its first channel. The STFT configuration is as follows:

- Window length: 2048;
- Overlap value: 1024;
- Nfft: 2048;
- Sampling frequency: 44.1 kHz;
- Window type: Hann window.

Furthermore, because the STFT of real-valued signal results in a conjugate symmetric matrix, we only use the first $\frac{nfft}{2} + 1$ or 1025 rows of the STFT matrix.

- 2) **Spectrogram:** After the STFT calculation, we computed the signal spectrogram from the 2D STFT matrix. The spectrogram represents the power distribution of different frequencies in each time bin, as shown by the spectrograms of some instruments in Fig. 6. It is obvious that the red area indicates the dominant frequencies in each time bin, which vary from one instrument to another. Thus, we use this fact as the feature extraction rule for training SoundNet.

- 3) **Binary Mask:** We calculate binary masks as a means to play the role of training labels and SoundNet outputs. To reduce information loss, high-energy time-frequency bins (red areas) of a spectrogram are assumed to be the main content of an audio signal, so they are identified and separated by masks. In fact, a binary mask plays the role of a filter, resulting in a separated audio signal. Considering our experiments in this study, the binary mask has better results and less interference than the soft mask.

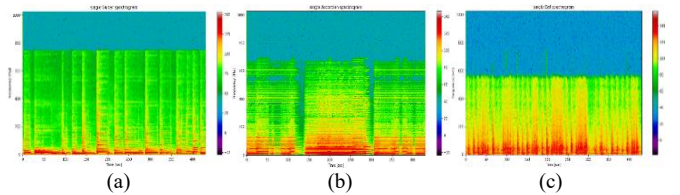


Fig. 6. The Spectrograms: (a) Guitar, (b) Accordion, (c) Daf

- 4) **Dimension reformation:** To match the SoundNet input size, each audio spectrogram and its corresponding binary mask, originally of size 1025×431 , were divided column-wise into 431 samples of size 1025×1 . Each sample was then reshaped into a $1 \times 1 \times 1025$ tensor. This process generated approximately 10,000 frame-level samples per instrument class (Fig. 7). SoundNet processes each time frame independently and therefore does not explicitly model temporal dependencies between consecutive frames. The training and test sets were separated at the original file level before this process to prevent data leakage.

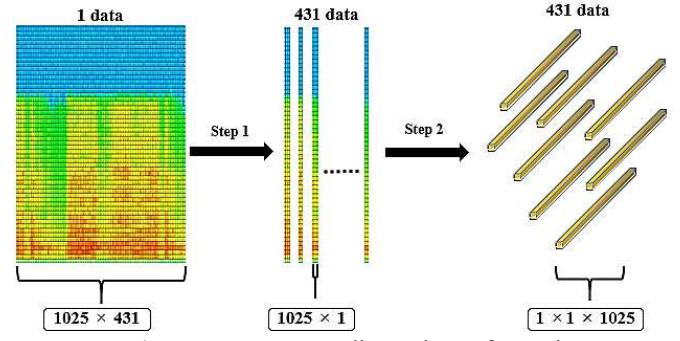


Fig. 7. Spectrogram dimension reformation

In the visual part, after the normalization of RGB values of image, the image data dimension is changed to $640 \times 640 \times 3$ before applying the instrument detection model (YOLOv8).

C. Deep Neural Network Implementation

The proposed method integrates two deep neural networks within a multimodal cross-modal learning framework. This structure establishes a connection between the results of the networks, enabling the networks to utilize both audio and visual information effectively for a specific task. Our algorithm is the first method designed specifically for the separation and localization of Persian musical instruments in videos. The deep neural networks work sequentially as follows:

- 1) **Instrument Detection:** For detecting musical instruments in images and videos, the YOLOv8s model is chosen due to its balance of efficiency and performance. Fig. 8 shows that the training dataset contains images captured in real environments with different conditions, qualities, and colors. Prior to training, all images are resized to a standard dimension of $640 \times 640 \times 3$ pixels to maintain consistency and enhance training efficiency. Data augmentation techniques, including random horizontal flips, cropping, blurring, scaling, and adjustments to image attributes such as saturation, brightness, and exposure, are applied to increase the model's robustness and prevent overfitting. The YOLOv8s variant (small model) is specifically selected for the detection model implementation. It is initialized with pre-trained weights from the COCO dataset, which includes a diverse array of object classes, providing a solid foundation for transfer learning. The backbone network utilizes a modified CSP (Cross Stage Partial) network to enhance gradient flow and reduce computational load. Neck network incorporates PANet (Path Aggregation Network) for effective multi-scale feature fusion and head network employs a decoupled head for object classification and bounding box regression. This comprehensive architecture ensures efficient and accurate detection of instruments in varied visual contexts. Model was trained using the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. The training process was carried out for 25 epochs with a batch size of 16.

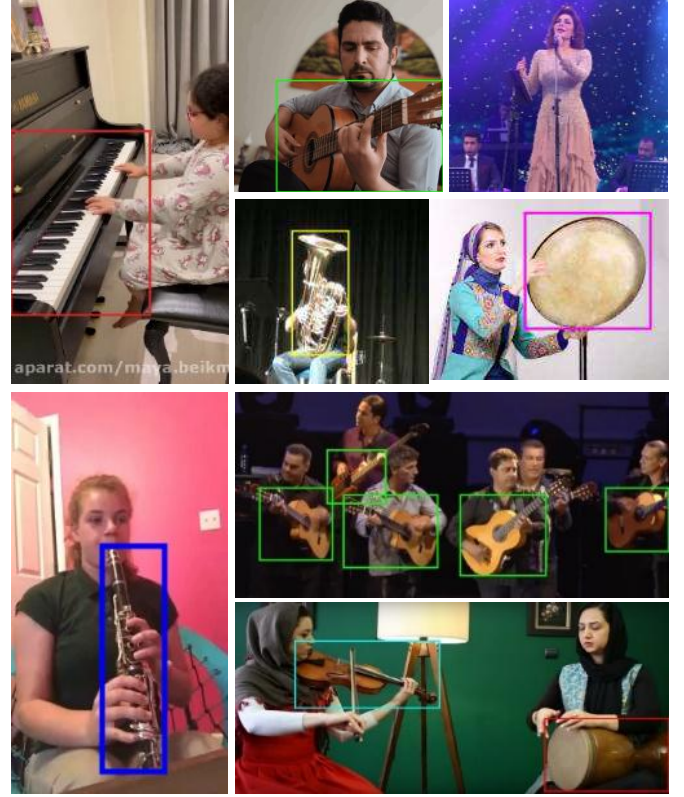


Fig. 8. Samples of the training dataset for the detection model

A cosine annealing learning rate scheduler was employed to gradually reduce the learning rate, thereby allowing the model to converge more effectively.

The loss function used for training consisted of three components: distribution focal loss, classification loss, and bounding box regression loss. Distribution focal loss is a loss function used to improve the precision of bounding box localization. It is designed to refine the predicted bounding box coordinates by focusing on the distribution of predicted probabilities for each bounding box. It helps the model achieve more accurate bounding box predictions by considering the distribution of bounding box coordinates. Classification loss stands for classification loss, which measures the error in the predicted class labels of the objects. It is usually computed using a cross-entropy loss function. It ensures that the model correctly identifies the classes of the detected objects. Bounding box

regression loss measures the error between the predicted bounding box coordinates and the ground truth bounding box coordinates. It often combines multiple components, such as the IoU (Intersection over Union) loss, to ensure accurate localization. It helps the model learn to predict the precise location and size of the bounding boxes that encompass the detected objects. These components were weighted to balance their contributions to the overall loss. Fig. 9 illustrates the training and validation loss curves for the instrument detection model, encompassing (a) bounding box regression loss, (b) classification loss, and (c) distribution focal loss. These diagrams indicate that validation loss trends closely follow those of training loss, demonstrating that model learning process is consistent and generalizes well to validation data. Fig. 10 showcases original video frames alongside the detected musical instruments, highlighted with bounding boxes generated by the detection model. In the final step of our system's visual analysis, the detection model's results are applied to the original frames. This process reconstructs individual frames for each instrument, as examples in Fig. 11.

- 2) **Sound Separation:** As mentioned before, we are facing two challenges in the sound separation part: source recognition and separation. The visual part results act as an appropriate solution for the first challenge. In fact, this is where the importance of visual feature used in the sound separation domain could be highlighted. Labels predicted by the instrument detection model (YOLOv8s) in visual processing are considered as the names of the present instruments in the video. In addition, the accuracy of the results is enhanced by considering the labels repeated at least in 4 frames among all 10 frames of a video as the final labels for the present sound sources in the video. Afterward, the sound source separation process is executed on the basis of the final predicted labels. We can figure out

that visual features can help us to identify the number and type of sound sources more precisely, leading to have better separation quality. moreover, it reduces the problem complexity and system computational cost.

In the proposed algorithm, each instrument as a single category has its own trained SoundNet model. For instance, the SoundNet model associated with the Guitar is trained only on the Guitar training dataset, so it can separate only the Guitar sound signals. In this way, we have 12 individual SoundNet models with similar architectures but trained on 12 different datasets related to the musical instruments in the research. The input data of a SoundNet consist of normalized reformed pieces of mixed audio signal spectrograms. The reformed pieces of relevant binary masks of single sounds are also used as training labels. For example, Fig. 12 shows a two-source mixed signal spectrogram in (a), which is a mixture of the source 1 and 2 spectrograms shown in (b).

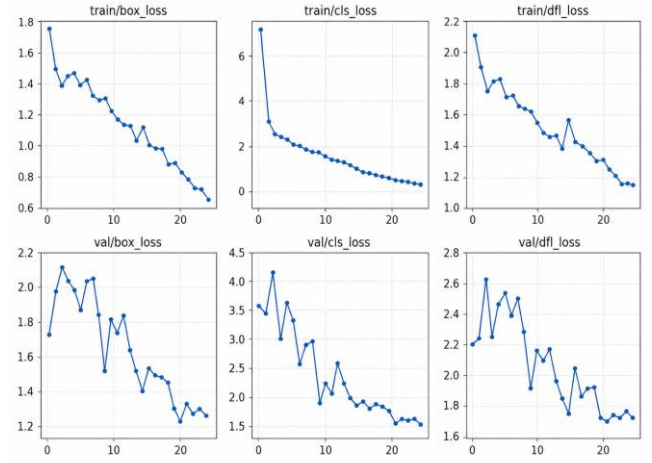


Fig. 9. Loss/epoch diagrams for training the detection model

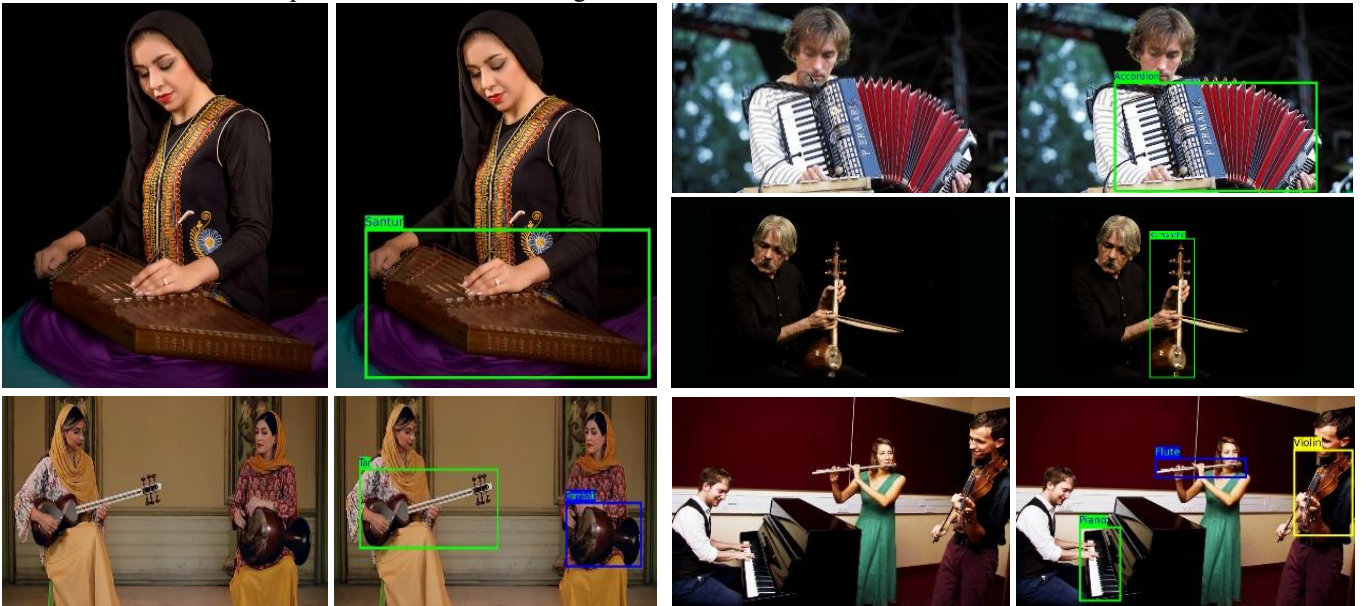


Fig. 10. The results of the instrument detection model: Main frame (left) and Detected instruments (right)

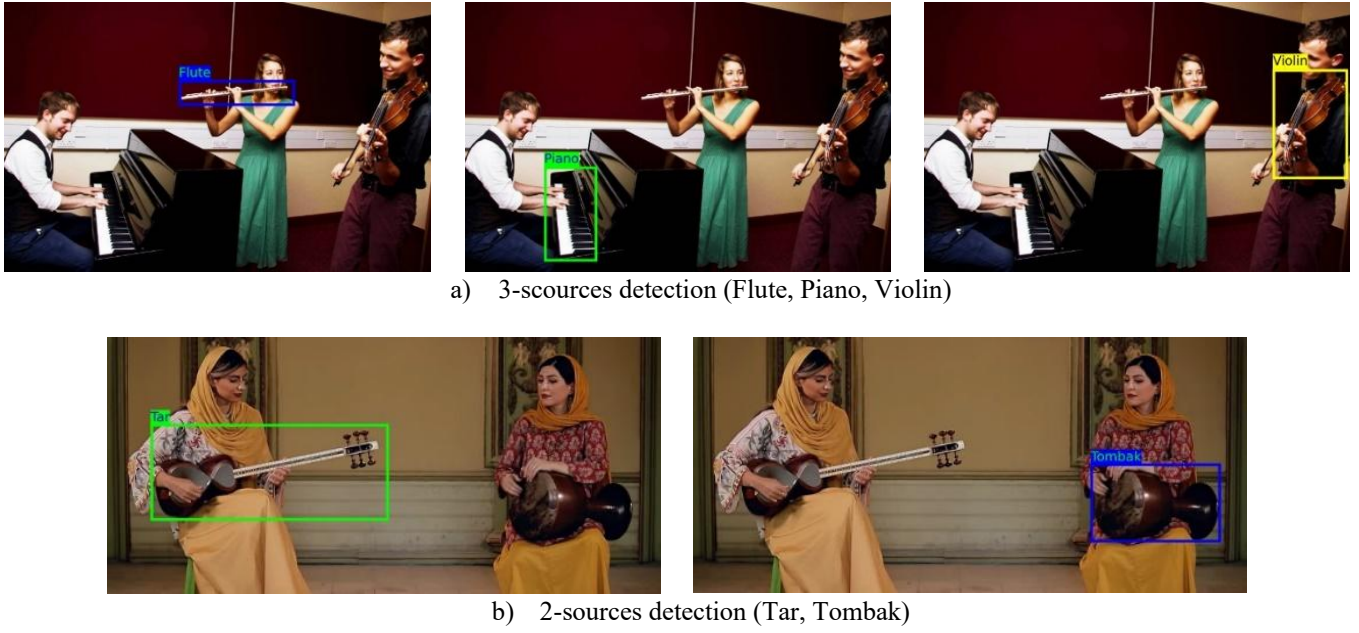


Fig. 11. Separately reconstructed frames for each instrument

SoundNet models for both sources are trained on spectrograms of mixed signals. Binary masks in Fig. 12 - (c), which are made from related single spectrograms of sources 1 and 2 in (b), are implemented as training labels for SoundNet models for sources 1 and 2, separately. Each binary mask acts as a sound separator for one source within a mixed signal of several sources. In addition, binary masks are converted to $1 \times 1 \times 1025$ data in the preprocessing stage and before entering the network. The output of each SoundNet model involves 431 components with a dimension of $1 \times 1 \times 1025$. Then, they are converted to a 1025×431 binary mask, belonging to the corresponding audio source making up the mixed audio signal, by some operations like concatenating.

In other words, the deep neural network in the sound separation section (SoundNet) can determine whether in each T-F bin of the mixed spectrogram, the desired audio

is multiplied by the spectrogram of the mixed signal to obtain the predicted spectrogram of a single sound signal. For example, Fig. 13 shows how to perform this step. On the left, there is a spectrogram of a mixed signal composed of two different audio sources. Furthermore, two binary masks have been estimated by related SoundNet models. The separated single spectrograms on the right-hand side of the figure are acquired by multiplying each predicted mask by the mixed spectrogram. As is obvious in each separated spectrogram, the energy of the T-F bins corresponding to the T-F bins of the binary mask with the value of 1 are separated, and the other T-F bins have a value of 0.

The separated spectrograms for sound sources are in the T-F domain. Therefore, we normalize them and reconstruct the separated sound signals in the time domain by ISTFT calculation for each sound source (instrument). The separated sound signals are the final outputs of the sound separation section of the proposed system.

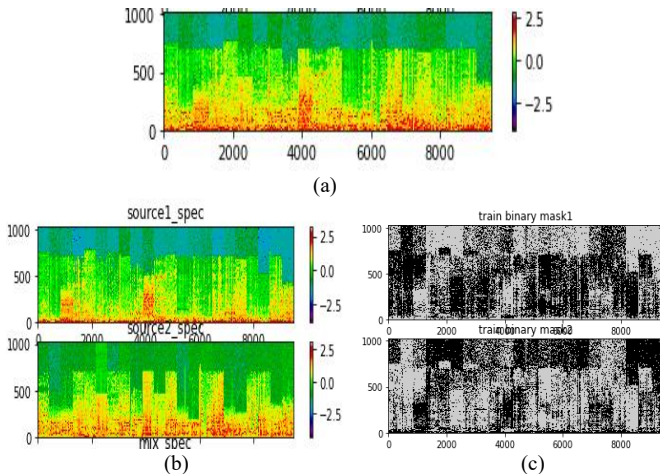


Fig. 12. (a) Spectrogram of the mixed sound, (b) the single spectrograms, and (c) the single binary masks source is active or not. Next, each predicted binary mask

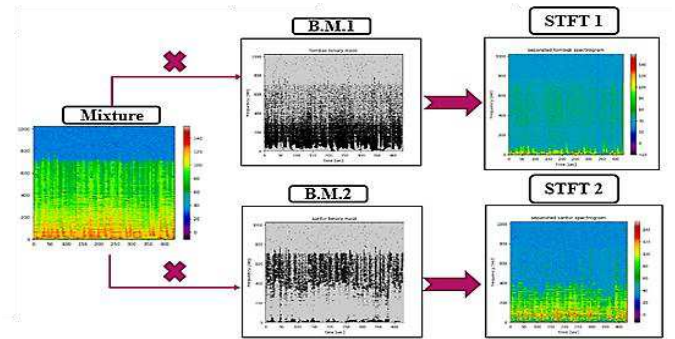


Fig. 13. Sound source separation using the predicted binary masks in the T-F domain

IV. RESULTS

To evaluate the results of the proposed system implementation, we address two general aspects: audio and visual. In the visual section, we analyze the results of the instrument detection model in video frames. In the audio section, we assess and compare sound separation performance using appropriate standard criteria. All results were obtained on the held-out test videos and their corresponding frames and sounds, none of which were used during model training.

A. Visual Evaluation

In object detection tasks, evaluating the performance of a model requires comprehensive metrics beyond accuracy. Precision, Recall, F1-Score, and Mean Average Precision (mAP) are fundamental metrics used to measure the effectiveness of detection models. Here they are defined as:

- **Precision:** Precision measures the accuracy of correctly recognized instruments made by the instrument detection model. A higher precision means fewer false recognition, indicating the ability of model to avoid incorrect detections.

$$Precision = \frac{TP}{TP+FP}, \quad (8)$$

- **Recall:** In our problem, recall, also known as sensitivity, signifies the capability of model to correctly detect all actual instruments within the frames. A higher recall value implies the model can capture more relevant objects.

$$Recall = \frac{TP}{TP+FN}, \quad (9)$$

- **F1-Score:** The F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both precision and recall, demonstrating the overall performance of the detection model.

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \quad (10)$$

where TP, FP, FN, are True Positive, False Positive, and False Negative, respectively.

- **Mean Average Precision (mAP):** AP measures the area of precision-recall curve for a particular object class. It evaluates how well a model detects objects by considering both precision and recall across various confidence thresholds. In scenarios with multiple object classes, mAP calculates the average AP across all classes. Higher mAP values indicate better overall object detection performance across multiple classes, considering both precision and recall for each category. The mAP@0.5 means the value of mAP at an Intersection over Union (IoU) threshold of 0.5.

$$AP = \int_0^1 precision(r) dr, \quad (11)$$

In case of object detection, YOLOv8 operates significantly better than a great number of object detection models because of its structure, which is capable of analyzing real images with different scenes and scales, low quality, and unclear objects more precisely and conceptually.

To evaluate the proposed instrument detection model, we use our test dataset, which includes 172 test images. According to

the results, the detection process is performed well with mAP@0.5 of 63.9%, Precision of 75.1%, Recall of 59.2%, and F1-score of 66.2%. For instruments with more different and simpler shapes (e.g., Daf) and playing pose, the detection model obtains better results. Compared to the Sound of Pixels method [1] with mAP@50 of 32.43%, our instrument detection model gains almost 2 times higher detection performance, confirming the efficiency of the transfer learning method on a small dataset.

Fig. 14 and 15 shows some output examples of the proposed instrument detection model on some real-world single-/multi-source test data, respectively. Each image is a video frame in which only one of the sound sources and its predicted probability is determined.

B. Audio Evaluation

To evaluate the algorithm and compare it with previous research quantitatively, we use the AudioSet Single Source test dataset [26], in which there is only one video for every single audio source. Thus, we use a set of standard criteria called BSS-EVAL [27] to measure the quality of audio signal separation. This set includes the power ratio of source signal to interference signal (SIR), the power ratio of source signal to artifact signal (SAR), and the power ratio of source signal to distortion signal (SDR). The SDR criterion expresses the overall average value of the separation quality, which includes the effects of both interference and artifact signals.

All of these statistical criteria are calculated by comparing the estimated single source signal with the original recorded single signal present in the mixed signal. In other words, SDR, SAR, and SIR are obtained from decomposing the estimated source signal $\hat{S}$ into

$$\hat{S} = S_{target} + e_{interf} + e_{noise} + e_{artif}, \quad (12)$$

where S_{target} is a good approximation of the power of the original source signal, e_{interf} is the power of the interference error signal caused by other sources, e_{noise} is the power of the environment noise signal, and e_{artif} is the power of the artifact error signal produced by other factors. Consequently, the evaluation criteria are calculated as follows [26]:

$$SAR = 10 \log_{10} \frac{\|S_{target} + e_{interf} + e_{noise}\|^2}{\|e_{artif}\|^2}, \quad (13)$$

$$SIR = 10 \log_{10} \frac{\|S_{target}\|^2}{\|e_{interf}\|^2}, \quad (14)$$

$$SDR = 10 \log_{10} \frac{\|S_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2}, \quad (15)$$

In this study, no artificial noise is added to the sound separation system or the mixture of sound signals; therefore, only the audio signals of the videos have been combined to produce mixed signals, without any noise changes in the preprocessing stage.

Table. I presents the obtained results of sound source separation for the test dataset using the proposed method. According to the SDR values, the highest quality of signal separation belongs to the Clarinet, Guitar, Violin, and Tombak.



Fig. 14. Example of single-source instrument detection

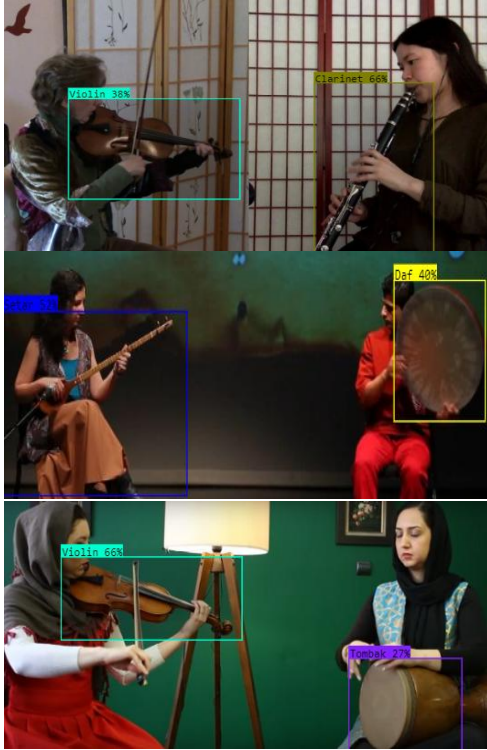


Fig. 15. Example of multi-source instrument detection

In general, the model is more efficient for sound separation in instances with louder voices, lower harmonics, and lower frequency range instruments. In contrast, in instruments with more vibration, a wider frequency range is used, so the frequency interference will be increased. As a result, the sound separation process will have lower accuracy.

The diagram in Fig. 16 illustrates the results reported in [2] along with our outcome, in the case of applying state-of-the-art sound source separation methods on the AudioSet dataset in the multi-source mode. For this purpose, a combination of single source signals is used to generate mixed signals.

TABLE I
RESULTS OF THE SOUND SOURCE SEPARATION MODEL

Instrument \ Criterion	SDR	SIR	SAR
Santur	1.36	25.70	1.71
Tombak	8.92	16.62	10.55
Flute	3.71	9.67	4.33
Setar	5.81	12.28	4.06
Daf	7.31	17.5	7.95
Kamancheh	4.38	16.31	5.02
Clarinet	22.29	19.12	4.14
Guitar	17.61	14.74	13.56
Violin	17.39	6.75	16.63
Piano	4.31	5.74	7.93
Accordion	3.73	9.88	5.57
Tuba	7.76	6.71	5.39
Average	8.75	13.42	7.24

Considering the significant differences in the SDR and SIR values between the proposed method and prior ones, we can claim that this technique yields the best separation quality (SDR) and interference reduction (SIR) results; thus, it can better perform in real world and noisy multi-source videos. As the second rank, the co-separation method [2] gains good results, especially on heavily overlapped sounds, because it considers synchronous movements of the player's hands with the played music as visual features.

In Fig. 17, the separated signals (blue signals) of the four sound sources and their corresponding original single sounds (red signals) are plotted in the time domain. As shown for each audio source, the original and separated signals have a significant overlapping portion, indicating good quality of sound separation.

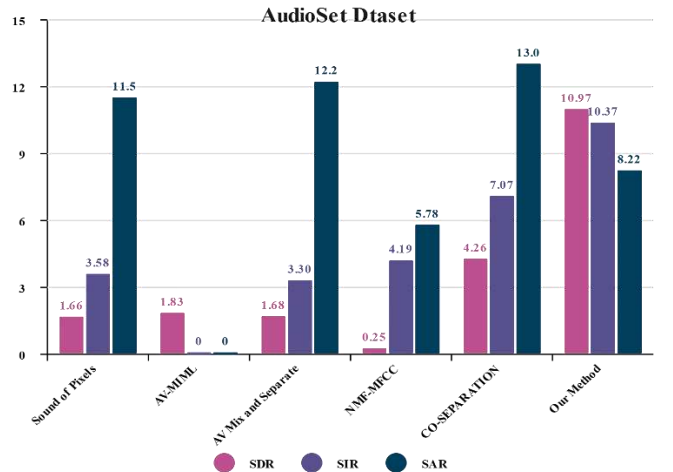


Fig. 16. Comparison of different sound separation methods (Sound of Pixels, AV-MIML, Av Mix and Separate, NMF MFCC, Co-Separation, Our Method) [2]

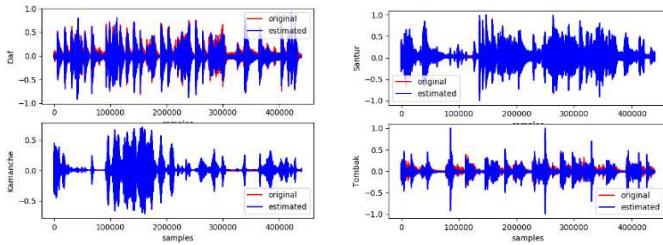


Fig. 17. Comparison of original and separated sound signals in four different instruments (Daf, Santur, Kamancheh, Tombak)

A key advantage of the proposed method is the effective association between the detected visual sources and their corresponding separated audio signals. The visual model identifies the number and type of active instruments, thereby guiding the selection of the relevant instrument-specific SoundNet models and reducing the separation search space. SoundNet then estimates a binary mask independently for each spectrogram time frame using a fully connected network. Although this frame-wise architecture does not explicitly model temporal dependencies between consecutive frames, it provides a simpler solution for spectral mask estimation. Moreover, the modular decomposition of the system into visual instrument detection, localization, and audio separation facilitates its implementation and extension to different instrument classes. Overall, the proposed method demonstrates effective performance on real-world videos recorded under diverse visual and acoustic conditions.

V. CONCLUSION

This study presented a modular audio-visual deep learning framework for detecting, visually localizing, and separating Persian and international musical instruments in single-channel real-world videos. The system uses a YOLOv8s model to identify the instruments present in the scene and select the corresponding instrument-specific separation networks, which estimate binary time-frequency masks from the mixed audio spectrogram. The results demonstrate the feasibility of using visual detections to guide class-specific sound separation, particularly for underrepresented Persian musical instruments. The main limitations are the use of a separate separation model for each instrument class, the lack of explicit temporal modeling, and the restriction to previously seen instrument categories.

REFERENCES

- [1] Zhao, H., et al., The sound of pixels. *In Proceedings of the European conference on computer vision (ECCV)*, 570-586 (2018).
- [2] Gao, R. and Grauman, K. Co-separating sounds of visual objects. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3879-3888 (2019).
- [3] Chatterjee, M., Roux, J.L., Ahuja, N. and Cherian, A. Visual scene graphs for audio source separation. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1204-1213 (2021).
- [4] Gan, C., Huang, D., Zhao, H., Tenenbaum, J.B., and Torralba, A. Music gesture for visual sound separation. *In Proceedings of the IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition*, 10478-10487 (2020).
- [5] Zhu, L. and Rahtu, E. 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*, 1289-1299 (2022).
- [6] Zhou, X., Zhou, D., Ouyang, W., Zhou, H. and Hu, D. Separating unknown musical visual sounds with consistency guidance. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5168-5177 (2023).
- [7] Chen, J., et al. iQuery: Instruments as queries for audio-visual sound separation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14675-14686 (2023).
- [8] Ye, Y., Yang, W. and Tian, Y. Lavss: location-guided audio-visual spatial audio separation. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5508-5519 (2024).
- [9] Mo, S. and Morgado, P. A unified audio-visual learning framework for localization, separation, and recognition. *in Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 25006-25017 (2023).
- [10] Song, Z. and Zhang, Z. Visually guided sound source separation with audio-visual predictive coding. *IEEE Trans. Neural Netw. Learn. Syst.*, **35**(11) 15528-15542 (2024).
- [11] Pian, W., et al., Continual audio-visual sound separation. *Advances in Neural Information Processing Systems*, **37**, (2024).
- [12] Kim, S., et al., Joint object detection and sound source separation. *Proc. 26th Int. Society for Music Information Retrieval Conf. (ISMIR)*, 813-820 (2025).
- [13] Cheng, X., et al., OmniSep: unified omni-modality sound separation with query-mixup. *Proc. Int. Conf. Learn. Representations (ICLR)*, (2025).
- [14] Um, S.J., Kim, D., Lee, S. and Kim, J.U. Object-aware sound source localization via audio-visual scene understanding. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 8342-8351 (2025).
- [15] Huang, C., Liang, S., Tian, Y., Kumar, A. and Xu, C. High-quality sound separation across diverse categories via visually-guided generative modeling. *Int. J. Comput. Vis.*, **134**(104) (2026).
- [16] Takahashi, A.S. and Mitsufuji, Y. "MMAudioSep: taming video-to-audio generative model towards video/text-queried sound separation. *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, (2026).
- [17] H. Wang, C. Li, X. Zhou, S. Wang, and Y. Qian, "USE: A unified model for universal sound separation and extraction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 39, pp. 33476-33484, 2026.
- [18] Farajzadeh, N., Sadeghzadeh, N. and Hashemzadeh, M. PMG-Net: persian music genre classification using deep neural networks. *Entertainment Computing*, **44**, 100518 (2023).
- [19] Ebrat, D., Didehvar, F. and Dadgar, M. Iranian modal music (dastgah) detection using deep neural networks. *arXiv preprint*, arXiv:2203.15335, (2022).
- [20] Hashemi, S.S., Aghabozorgi, M.R. and Sadeghi, M.T. Persian music source separation in audio-visual data using deep learning. *In 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, 1-5 (2020).
- [21] Hashemi, S.S., Asadi, M. and Aghabozorgi, M.R. An Audio-Visual System for Sound Noise Reduction Based on Deep Neural Networks. *In 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, 1-6 (2021).
- [22] Shaikh, M.B.D., Shamsul Islam, S. M. and Akhtar, N. Multimodal fusion for audio-image and video action recognition. *Neural Computing and Applications*, **36**(10), 5499-5513 (2024).
- [23] Allen, J.B. and Rabiner, L.R. A unified approach to short-time Fourier analysis and synthesis. *In Proceedings of the IEEE*, **65**(11), 1558-1564 (1977).
- [24] Jocher, G., Chaurasia, A. and Qiu, J. Ultralytics YOLOv8 version 8.0.0. <https://github.com/ultralytics/ultralytics>, (2024).
- [25] Persian music instruments. <https://www.aparat.com/Coocmusic.ac>
- [26] Gemmeke, J.F. et al. Audio set: an ontology and human-labeled dataset for audio events. *In 2017 IEEE international conference on acoustics, speech and signal processing*, 776-780 <https://research.google.com/audioset/>, (2017).
- [27] Vincent, E., Gribonval, R., and Févotte, C. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, **14**(4), 1462-1469 (2006).

DECLARATIONS

Funding information: The authors received no specific funding for this work.

Competing interests: The authors declare no conflicts of interests.

Data availability: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Figures statement: The images in figures 1, 2, 8 ,10, 11, 14 and 15 are sourced from publicly available dataset (Google search).

Datasets statement: The dataset for this study comprises audio-visual (video) data categorized into 12 classes of audio sources. These sources include Persian musical instruments as well as international ones. The data was collected from publicly available videos and images on the Internet [25]. For the international instruments, we also utilized some videos of the “Audioset” [26] dataset.