Visually Guided Sound Source Separation and Localization using Self-Supervised Motion Representations –Supplementary Material

Lingyu Zhu
Tampere University, Finland
lingyu.zhu@tuni.fi

Esa Rahtu
Tampere University, Finland
esa.rahtu@tuni.fi

The supplementary material is organized as follows: Section A provides additional visualization of the source separation and localization; Section B contains additional details of the network architectures; and Section C presents the optimization and evaluation configurations.

A. Additional Qualitative Results

This section provides additional qualitative results of the visual sound source separation and source localization results. The experimental setups are as explained in the main paper.

Sound Source Localization Figures B and C provide additional qualitative results of the sound source localization with the proposed Audio-Motion Embedding (AME) framework, Cascaded Opponent Filter (COF), and Multisensory using MUSIC-21 and AVE datasets, respectively.

Visually Guided Sound Source Separation Figures D and E present additional qualitative results of separating mixtures of two sound sources using AMnet from the MUSIC-21 and AVE datasets, respectively. Figure F shows results of separating mixtures of three sound sources from MUSIC-21. Figure G contains results of separating sources of the same type from the MUSIC-21 dataset.

Quantitative experiments in fully natural scenarios are not possible due to the lack of ground truth for the source components. However, a qualitative example is shown in Figure A (click to play).

B. Network Architectures

This section provides additional details of the network structures and implementation.

B.1. Audio-Motion Embedding Framework

Motion Network The Motion Network E_M utilizes a 3D version of Res18 on the input video sequence of size

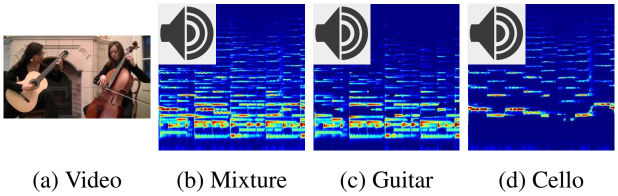
(a) Video (b) Mixture (c) Guitar (d) Cello

Figure A: Visually guided sound source separation in natural scenario. Use Adobe Acrobat Reader to play.

$3 \times T \times H \times W$, where $T = 48$ and $H = W = 224$. With the *stride*=16 on spatial dimension and *stride=4* on the temporal dimension, we yield the motion representation f_{M1} of size $C_M \times T^{'} \times H^{'} \times W^{'}$, where $C_M = 512$, $T^{'}$=12 and $H^{'} = W^{'} = 14$. With an additional 3D convolution, we obtain a single channel feature map f_{M2} of size $1 \times T^{'} \times H^{'} \times W^{'}$. Furthermore, we add a spatial average pooling over the $H^{'}$ and $W^{'}$ dimensions to achieve the final motion embedding vector f_{M3} of size $1 \times T^{'}$.

Sound Network The Sound Network E_S uses Res18-1D architecture to map the input audio waveform into a common vector space with the Motion Network. The Sound Network is composed of a series of 1D convolutions. A fractional poling and a 1D convolution layers are applied on top to obtain the final one channel embedding vector of size $1 \times T^{'}$.

B.2. Audio-Appearance Sound Source Separation

Appearance Network We adopt frame augmentation of random scaling, random horizontal flipping, and random cropping (224 × 224) during training for all datasets. We apply a dilated Res18-2D with *dilation*=2 to obtain the appearance representations. For an input RGB image of size $3 \times H \times W$, we truncate the Res18-2D after *stride*=16 and achieve the appearance feature of size $C_A \times H^{'} \times W^{'}$, where $H^{'} = W^{'} = 14$, C_A equals to 21 and 28 for MUSIC-21 and AVE datasets respectively. C_A represents the category numbers of dataset. By performing a spatial average pooling operation on the top, the Appearance Network produces the representation f_A of size *1* $\times$ C_A.

Sound Spectrogram Network We firstly convert the input audio waveform into a spectrogram presentation X_{mix} using Short-time Fourier Transform (STFT), and then forward the mixture spectrogram as the input of the Sound Spectrogram Network. The Sound Spectrogram Network is implemented using MobileNetV2 (MV2) architecture. The network converts the input spectrogram of size $1 \times H_S \times W_S$ to a feature map f_{mix} of size $C_S \times H_S \times W_S$, where $H_S = W_S = 256$, C_S equals to 21 and 28 for MUSIC-21 and AVE datasets respectively. Note that the number of produced feature maps C_S is equal to the appearance feature vector dimension C_A in the previous section.

Sound Source Separation The sound source separation module combines the appearance representations $f_{A,n}$ of n-th source with the sound spectrogram network output f_{mix} using a linear combination to produce the spectrum features $f_{S,n}^{appearance}$ (the superscript *appearance* refers to the Audio-Appearance stage) of size *1* $\times$ *256* $\times$ *256*. With the *sigmoid* and *thresholding* (th = *0.5*) operations, the spectrum features are converted to binary masks $\hat{B}_n^{appearance}$. The output spectrogram is formed by an element-wise multiplication between the binary mask and the original mixture spectrogram. We forward the output spectrograms $\hat{X}_S^{appeearance}$ of all the sources from the Audio-Appearance stage to the upcoming Audio-Motion stage as inputs.

B.3. Audio-Motion Sound Source Separation

Motion Network The Motion Network in the Audio-Motion stage is pre-trained by the Audio-Motion Embedding (AME) framework in Section B.1. We apply a spatial average pooling operation over the $H^{'}$ and $W^{'}$ dimension of the motion features f_{M1} to obtain the motion representation of size $C_M \times T^{'}$, where C_M = *512* and $T^{'}$ = *12*.

Sound Spectrogram Refinement Network The Sound Spectrogram Refinement (SSR) network takes the output spectrograms from the Audio-Appearance stage as inputs. The SSR has an encoder-decoder architecture. The encoder *SSR*$_E$ processes the input spectrogram into sound features $f_S^{motion,encoder}$ (the superscript *motion* refers to the Audio-Motion stage) of size *512* $\times$ *16* $\times$ *16*. The encoder is followed by the Audio-Motion Transformer (AMT) module to fuse the motion and spectrogram features. We employ 8 parallel heads attention layers in the AMT module. The following decoder *SSR*$_D$ produces residual spectrum features $f_{S,n->m}^{motion,decoder}$ of size *1*$\times$*256*$\times$*256*. We relocate the identified residual spectrum components from Audio-Appearance outputs to our final corresponding spectrum feature f_S by using a Residual Fusion module (Eq. 4). With the *sigmoid* and *thresholding* (th = *0.5*) operations, the spectrum features are converted to binary masks $\hat{B}^{motion}$. The output spectrogram $\hat{X}_S^{motion}$ is formulated by an element-wise multiplication between the resulted binary mask $\hat{B}^{motion}$ and the original mixture spectrogram X_{mix}. With an inverse STFT, we obtain the final separated audio waveforms.

C. Implementation Details

Optimization The proposed model was implemented in Pytorch using stochastic gradient descent (SGD) with momentum 0.9, weight decay 1e-4, and batch size 10 for training. Except the Appearance Network that was pre-trained on ImageNet uses a learning rate of 1e-4, all the other modules are trained from scratch using a learning rate of 1e-3.

Evaluation We assess the AME based motions cues in three different motion related tasks: i) sound source localization; ii) action recognition; and iii) audio-visual sound source separation. For all the evaluation metrics, higher value indicates better performance.

In order to give a quantitative evaluation of the AME motions, in addition to the qualitative visualizations, on the task of sound source localization, we measure the consensus Intersection over Union (cIoU) and Area Under Curve (AUC) metrics. Though with the fact that there is no direct dataset which has the ground truth of motion localization, we use the detected bounding boxes of mask r-cnn [1] to indicate the coarse localization of sounding objects.

For the action recognition task, we simply add a fully connected layer on top of the motion features for classifying the actions. We measure the performance by reporting the classification accuracy (Acc) on UCF-101 [2] dataset.

The sound separation performance is measured in terms of: Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), and Signal to Artifact Ratio (SAR). SDR and SIR scores measure the separation accuracy. SAR captures only the absence of artifacts, hence can be high even if separation is poor.

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[2] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012.

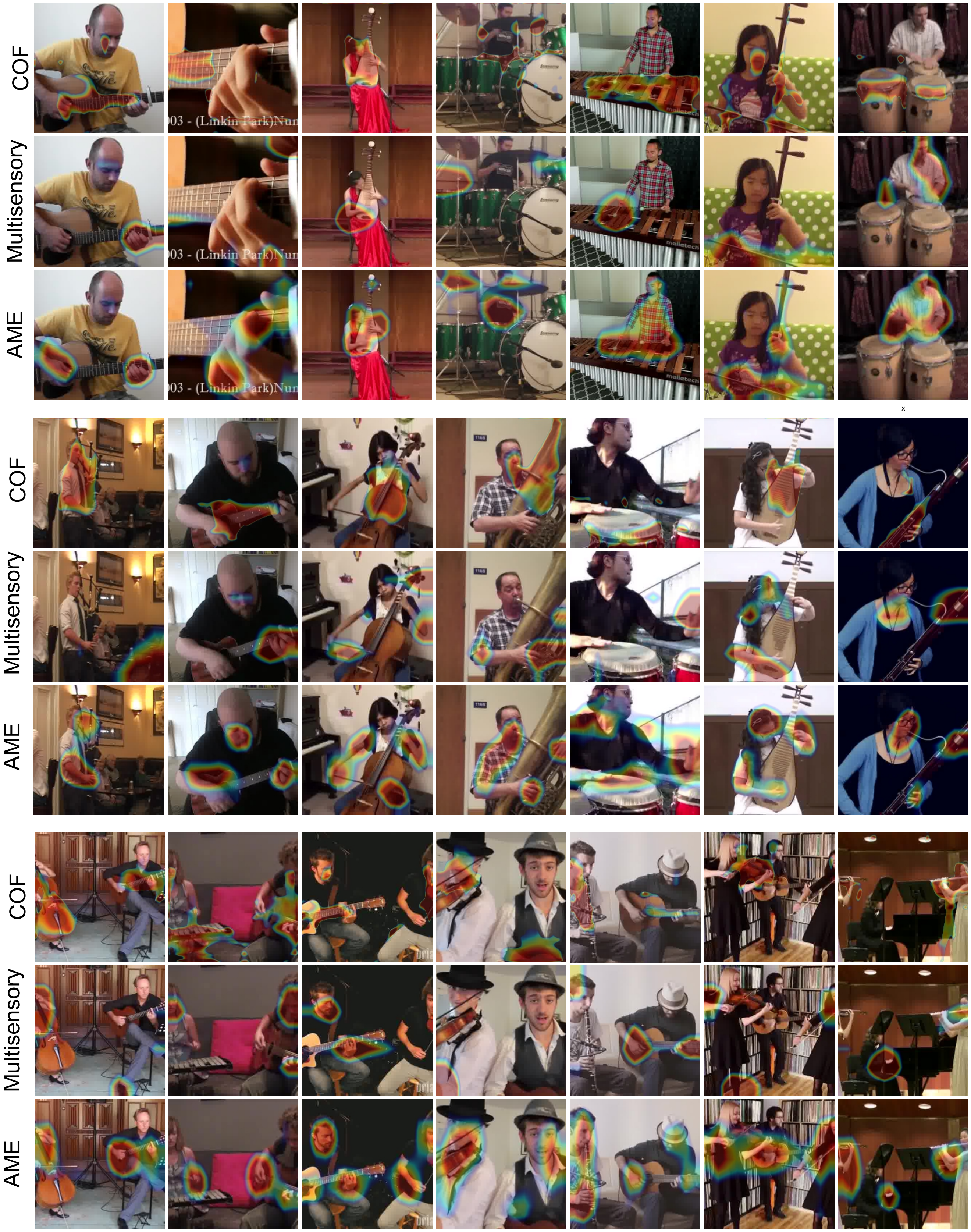

Figure B: Visualization of the CAM responses with MUSIC-21 dataset for AME, COF, and Multisensory.

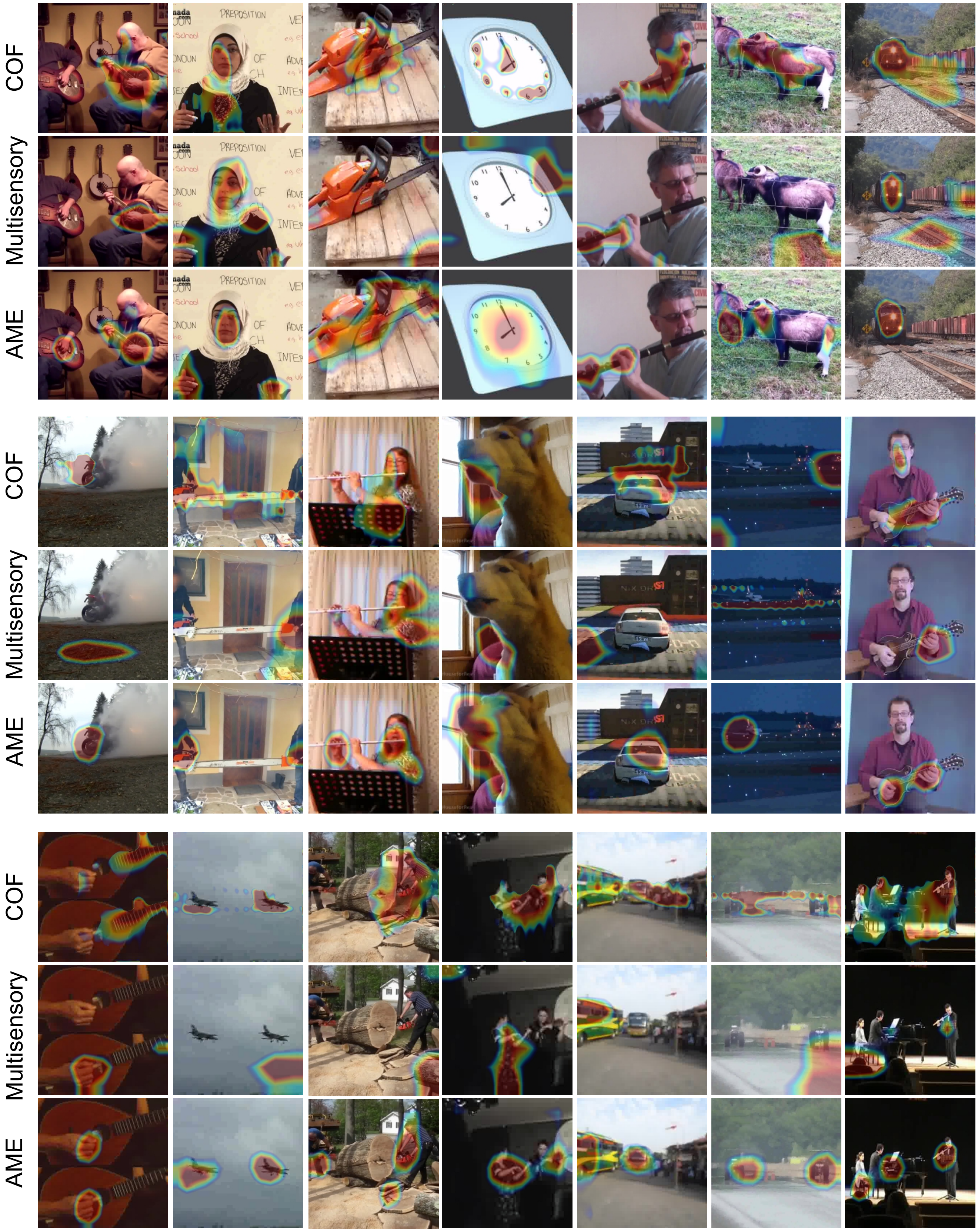

Figure C: Visualization of the CAM responses with AVE dataset for AME, COF, and Multisensory.

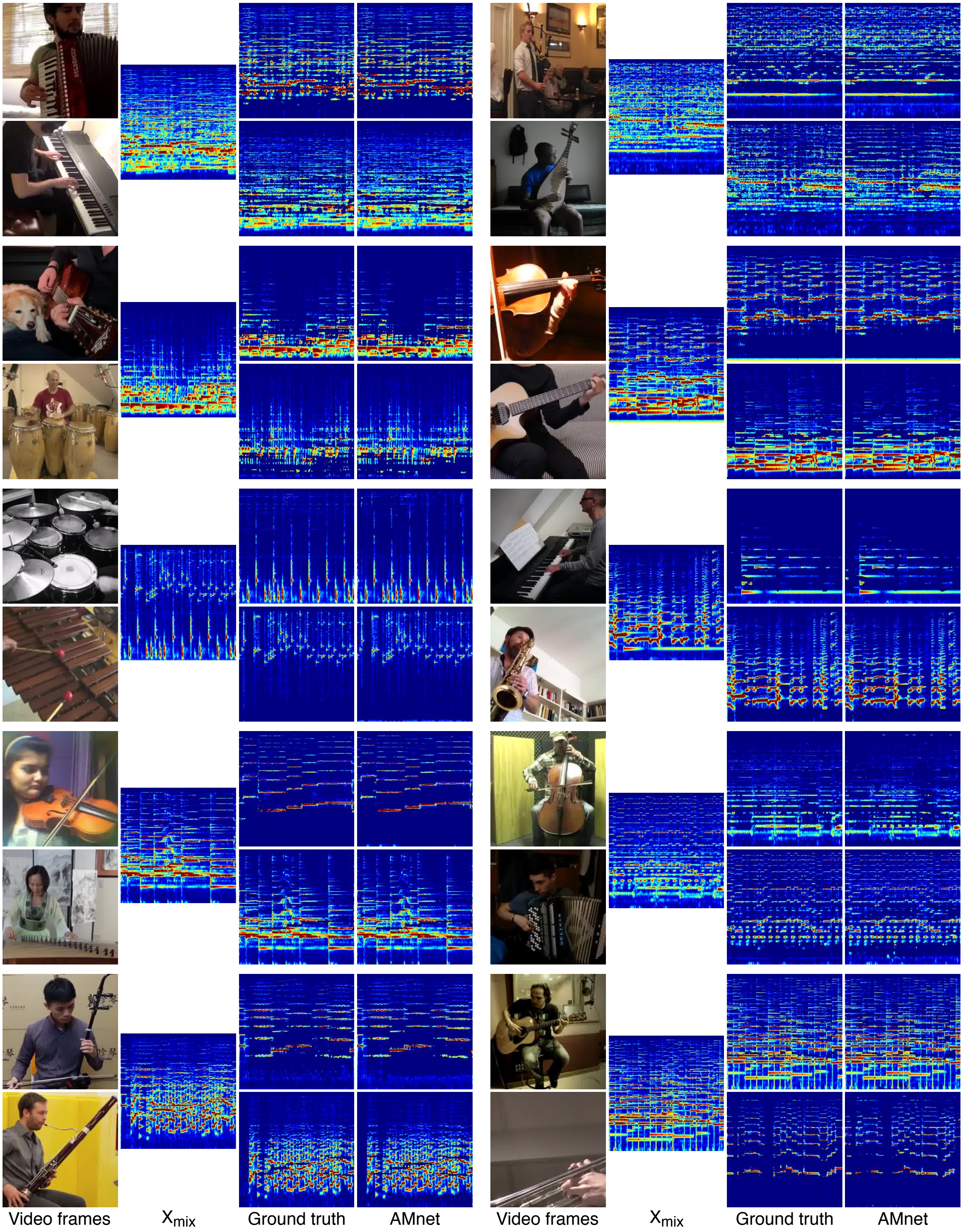

Figure D: Visualization of the sound source separation results using AMnet with mixtures of two different sources from MUSIC-21 dataset.

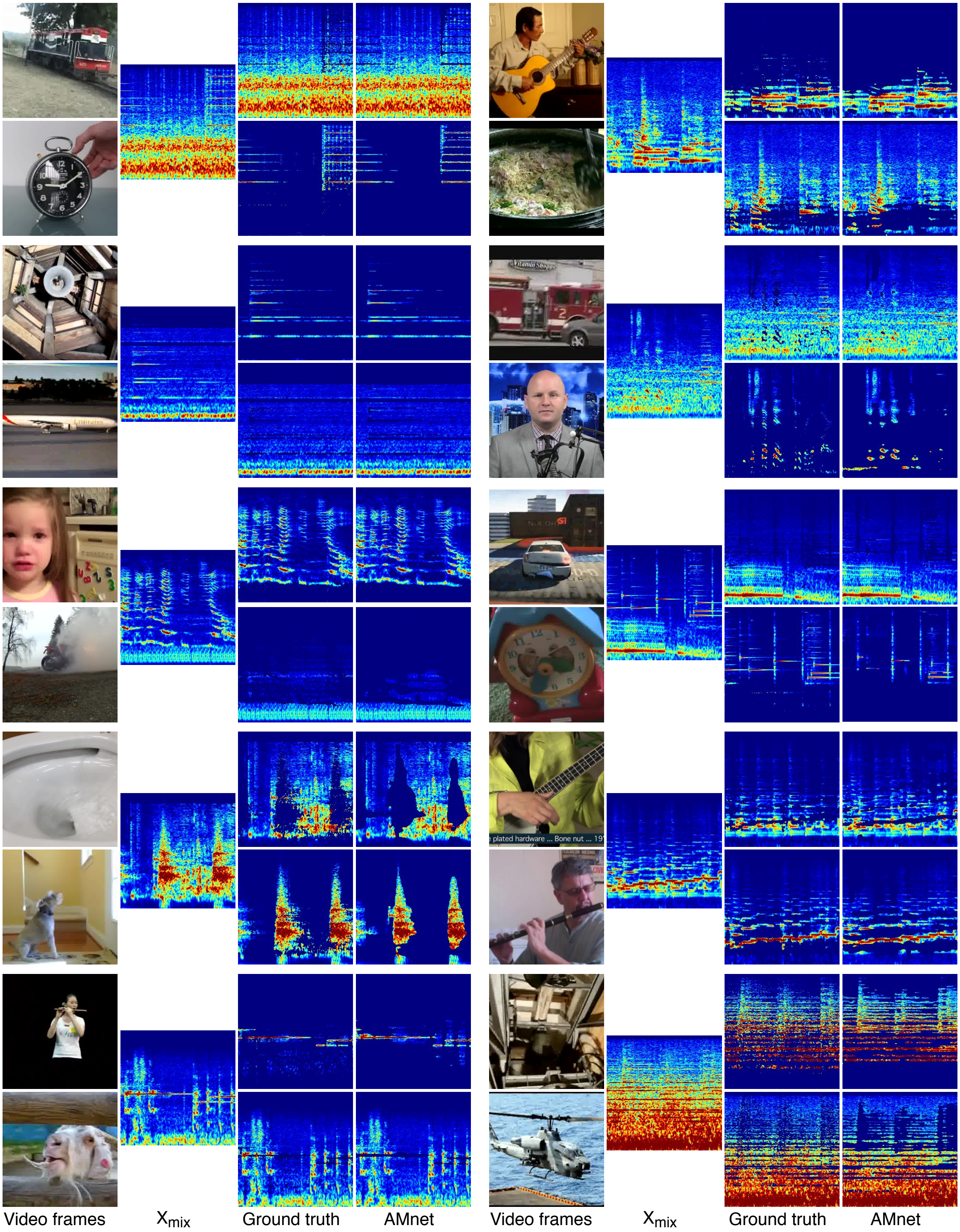

Figure E: Visualization of the sound source separation results using AMnet with mixtures of two different sources from AVE dataset.

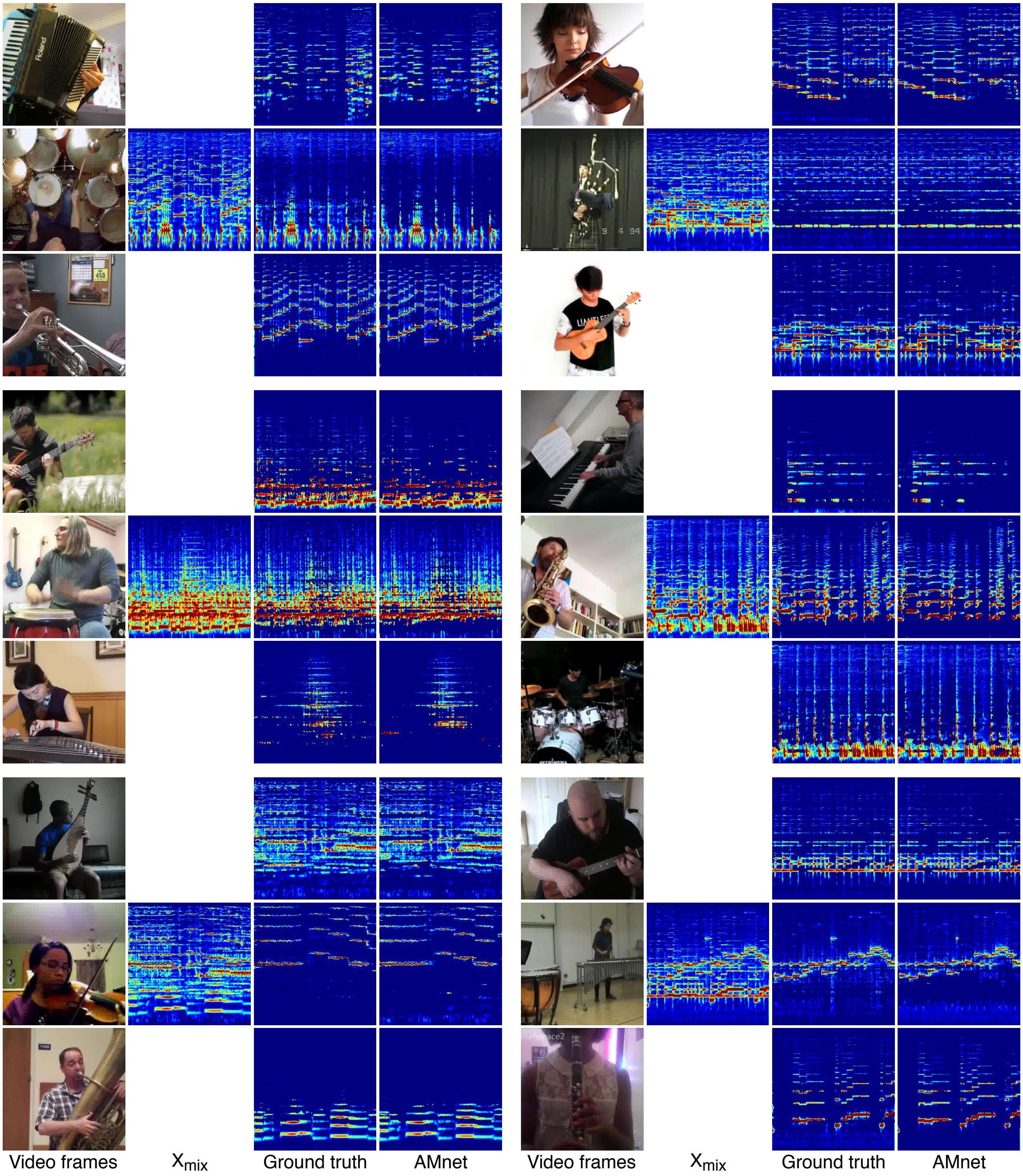

Figure F: Visualization of the sound source separation results using AMnet with mixtures of three different sources from MUSIC-21 dataset.

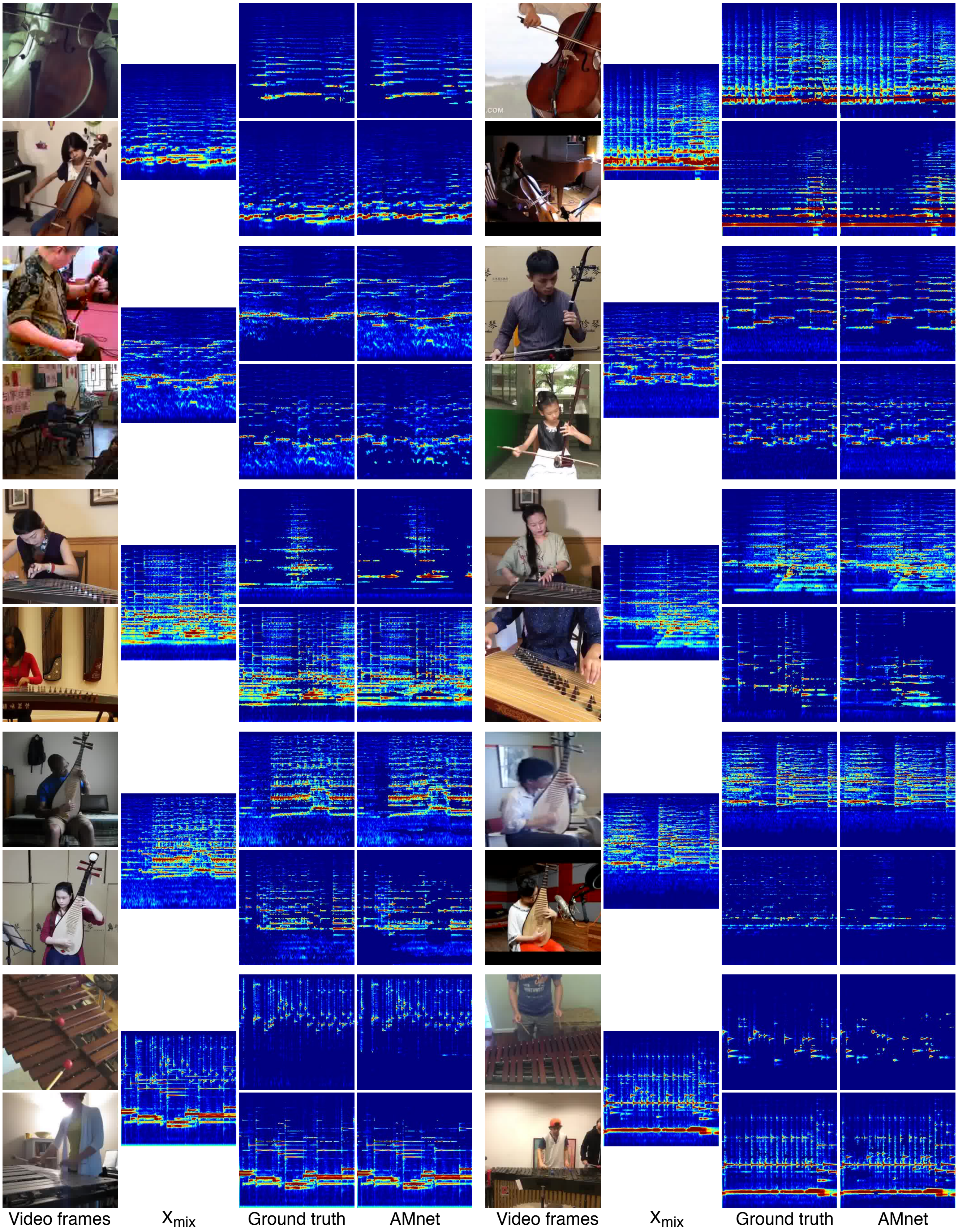

Figure G: Visualization of the sound source separation results using AMnet with mixtures of two same type sources from MUSIC-21 dataset.