V-SlowFast Network for Efficient Visual Sound Separation –Supplementary Material

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

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

The supplementary material is arranged as follows: Section A reports the loss and matrix curves of the visual sound source separation on single spectrogram of different temporal resolutions; Section B presents the visual sound separation performance with V-FastSlow framework; Section C provides additional visualization of the source separation and localization; Section D contains additional implementation details.

A. Visual Sound Separation on Single Spectrogram of Different Temporal Resolutions

In Figure A, we display the loss and evaluation matrix curves (training procedure) of the visual sound source separation performance on single spectrogram of different temporal resolutions ($\alpha \in \{1, 2, 4, 8, 16\}$). We observe that the training procedure with larger α converges faster. In addition, the smaller temporal resolution (larger α) the input spectrogram has, the lower evaluation scores of SDR and SIR the models obtain, which reflects the larger separation loss. However, as is shown in Figure A, the SAR score does not follow the same trend. SAR captures only the absence of artifacts, hence can be high even if separation is poor. Thus, we conclude that the SDR and SIR scores measure the separation quality.

B. Visual Sound Separation with V-FastSlow Network

In order to study whether the order of the slow and fast spectrogram matters, we also assess the opposite way (V-FastSlow), where the fast spectropgram appears first and slow spectrogram occurs second. We experimented the V-FastSlow network with $\alpha_f = 1$, $\alpha_s \in \{2, 4, 8, 16\}$ and reported the results in Table A. The V-FastSlow obtains very close performance as the V-SlowFast in terms of the evaluation metrics, number of parameters and operations. Especially when the α_s=2 and 4, the V-SlowFast achieves slightly better performance, e.g. gain of $0.2 \sim 0.4$dB in SDR. Thus, we mainly discuss on the case of V-SlowFast model in the main paper.

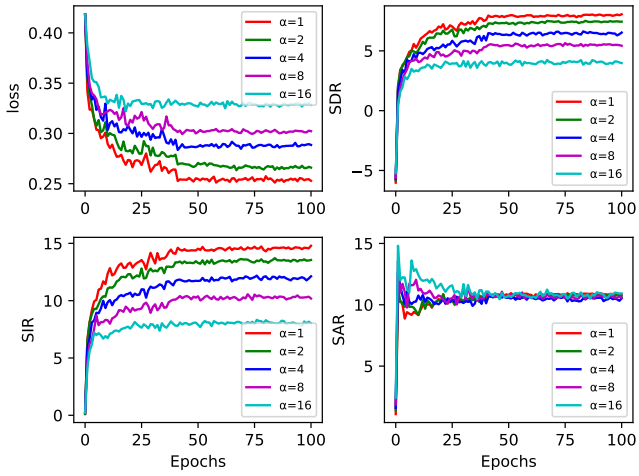


Figure A: Visualization of the loss and matrix curves of the visual sound source separation on single spectrogram of different temporal resolutions ($\alpha \in \{1, 2, 4, 8, 16\}$).

C. Additional Qualitative Results

This section provides additional qualitative visualization of the visual sound source separation and source localization examples.

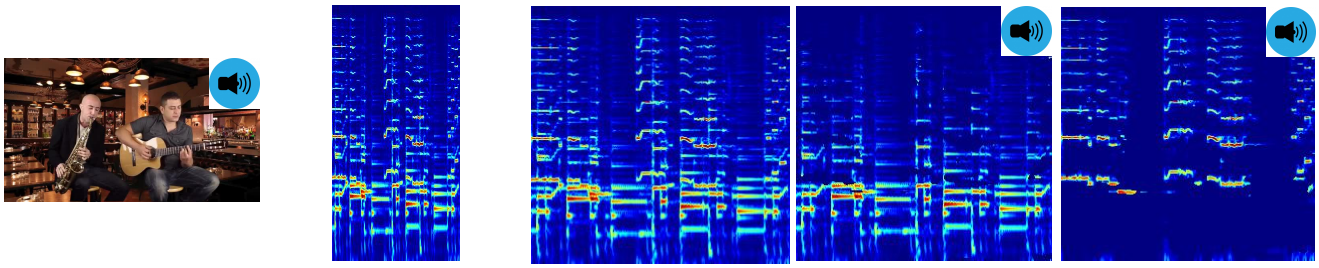
C.1. Visual Sound Separation

Figures C, D and E present additional qualitative visualization of separating mixtures of two sound sources using V-SlowFast network from the MUSIC-21, AVE, and VGG-Sound datasets, respectively. Figure F and G show results of separating mixtures of three and four sound sources from MUSIC-21.

A natural scenario example is shown in Figure B (click to play).

C.2. Sounding Source Localization

Figures H, I and J provide additional qualitative visualization of the sound source localization with the proposed



(a) Input video (b) Slow spectrogram (c) Fast spectrogram (d) Guitar (e) Saxophone

Figure B: Visual sound separation in natural scenario using V-SlowFast (2) network. Use Adobe Acrobat Reader to play.

Fast	Slow	SDR	SIR	SAR	Param (M)	GMACs
α_f=1, U-Net (7-layer)	α_s=2, U-Net (7-layer)	10.26	16.92	12.87	51.69	2.45
α_f=1, U-Net (7-layer)	α_s=4, U-Net (7-layer)	10.11	16.97	12.59	51.69	2.10
α_f=1, U-Net (7-layer)	α_s=8, U-Net (7-layer)	9.94	16.83	12.34	51.69	1.92
α_f=1, U-Net (7-layer)	α_s=16, U-Net (7-layer)	9.58	16.62	11.92	51.69	1.84
α_f=1, U-Net (9-layer)	α_s=2, U-Net (5-layer)	10.11	16.50	12.73	39.74	2.65
α_f=1, U-Net (9-layer)	α_s=4, U-Net (5-layer)	10.04	17.00	12.50	39.74	2.40
α_f=1, U-Net (9-layer)	α_s=8, U-Net (5-layer)	9.96	17.06	12.35	39.74	2.28
α_f=1, U-Net (9-layer)	α_s=16, U-Net (5-layer)	9.71	17.00	12.01	39.74	2.21

Table A: Source separation performance using mixtures of two sources from the MUSIC-21 dataset with V-FastSlow network for α_f=1, $\alpha_s \in \{2, 4, 8, 16\}$. The vision network is *Res-18 + AVGA + Contrast*.

V-SlowFast framework using MUSIC-21, AVE, and VGG-Sound datasets, respectively.

D. Implementation Details

We extract video frames at 8 fps for all datasets and sub-sample audio signal at 11KHz, 22kHz, and 22KHz for MUSIC-21, AVE, and VGG-Sound datasets, respectively. We randomly crop 6-second audio clip and convert the input audio to F-T spectrogram using STFT with a hanning window of size 1022 (MUSIC-21, AVE) and 1498 (VGG-Sound), and a hop lengths of 256 (MUSIC-21), 184 (AVE) and 375 (VGG-Sound).

A single frame (224×224) is forwarded to the vision network. The vision network produces a compact representation $e_v \in \mathbb{R}^{1 \times C_V}$. C_V equals to 21, 28 and 310 for MUSIC-21, AVE, and VGG-Sound datasets, respectively. The dimension of sound features C_S equals to C_V, which represents the category numbers of dataset.

α = 1 represents the full temporal resolution spectrogram. The slow spectrogram network and the fast spectrogram residual network take the low and high temporal resolution spectrograms as input, respectively. Thus, we consider $\alpha_s > 1$, and $\alpha_f < \alpha_s$. The ϕ^{inv} (inverse ϕ) operation inverts the spectrogram into full temporal resolution spectrogram of α_s = 1 or α_f = 1.

The proposed V-SlowFast model is implemented using Pytorch framework. We adopt stochastic gradient descent (SGD) with momentum 0.9, weight decay 1e-4, and batch size 10. The vision network, pre-trained on ImageNet, uses a learning rate of 1e-4, while all other of modules are trained from scratch using a learning rate of 1e-3.

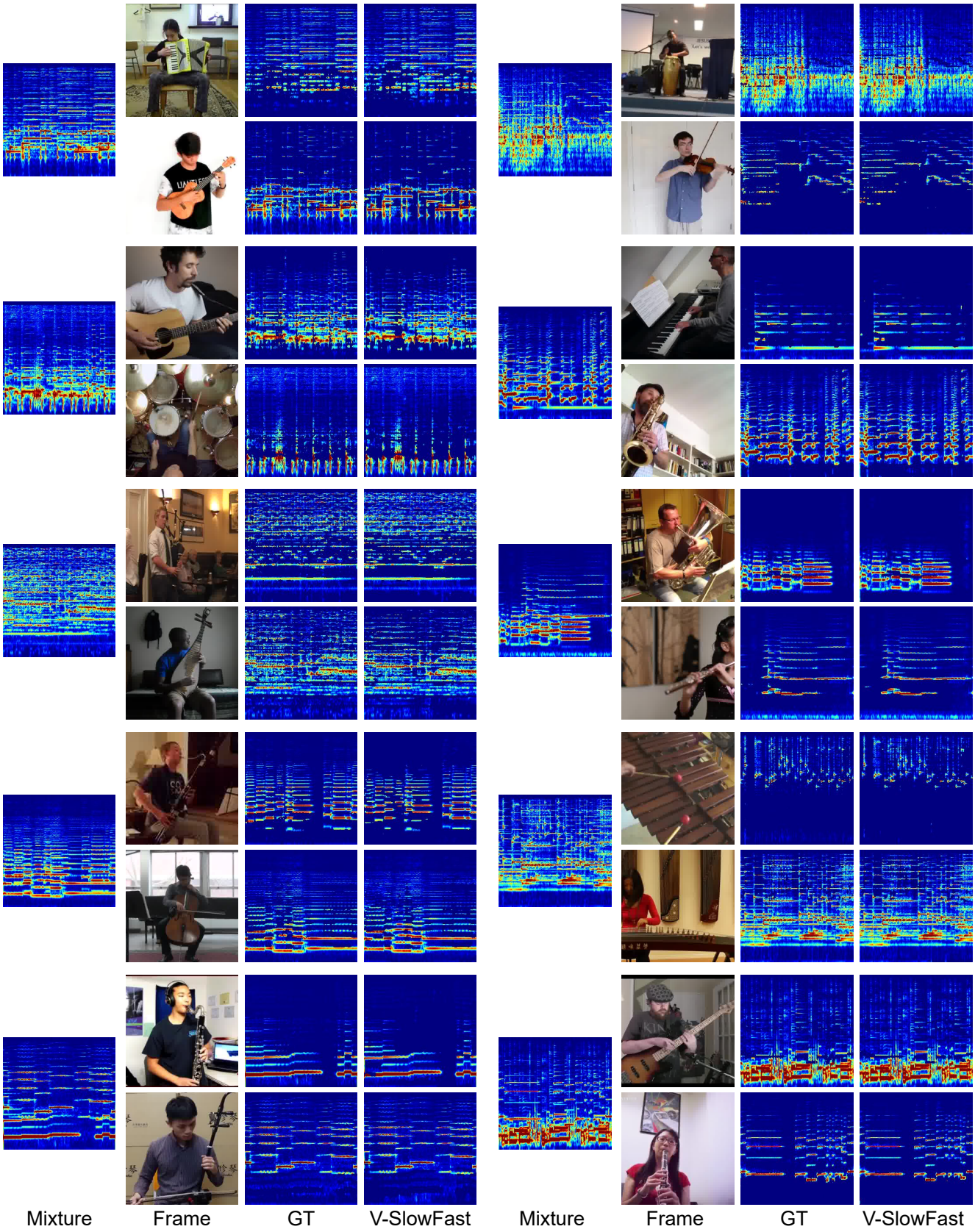


Figure C: Visualization of the source separation results using V-SlowFast network with mixtures of two sources from MUSIC-21 dataset.

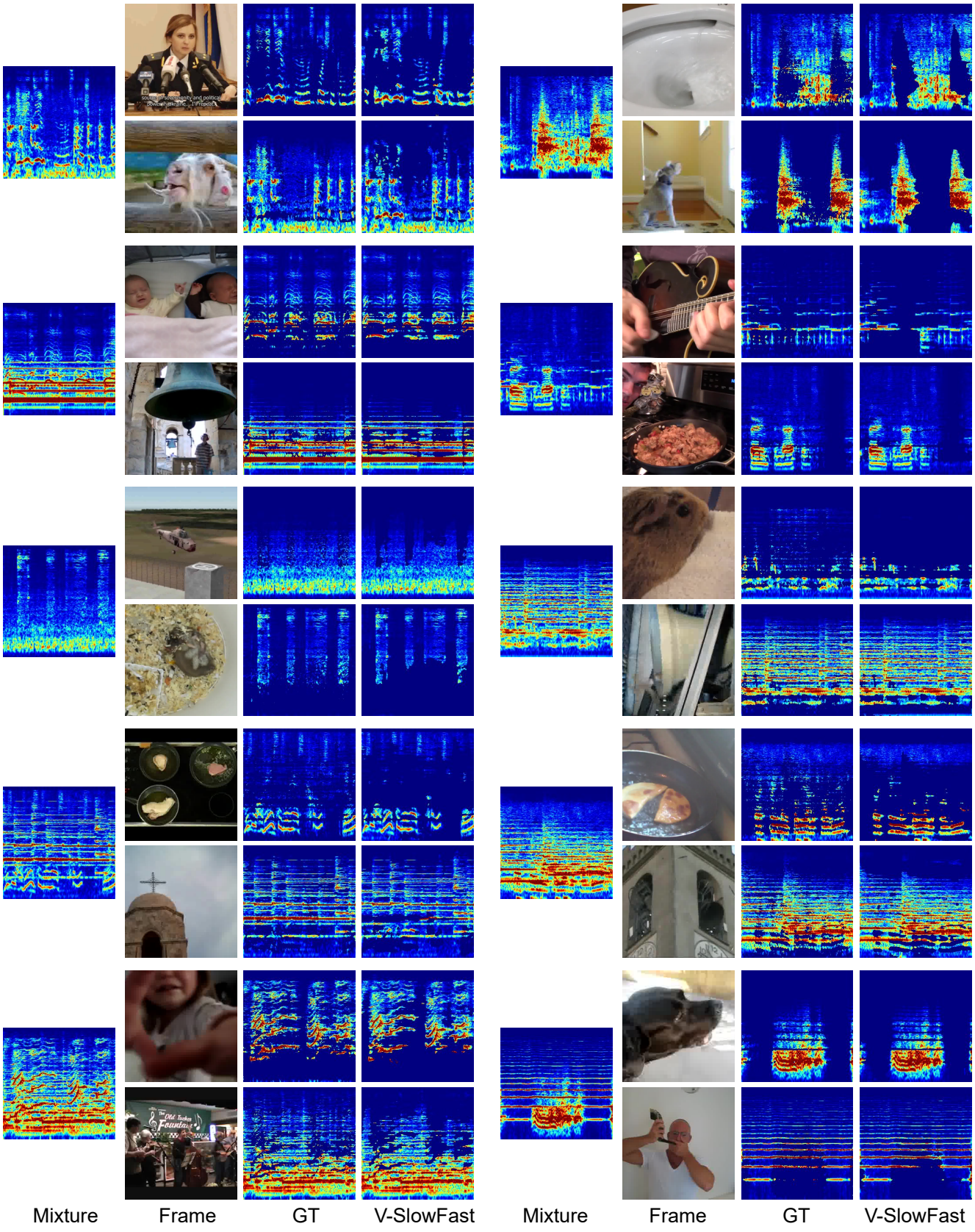


Figure D: Visualization of the source separation results using V-SlowFast network with mixtures of two sources from AVE dataset.

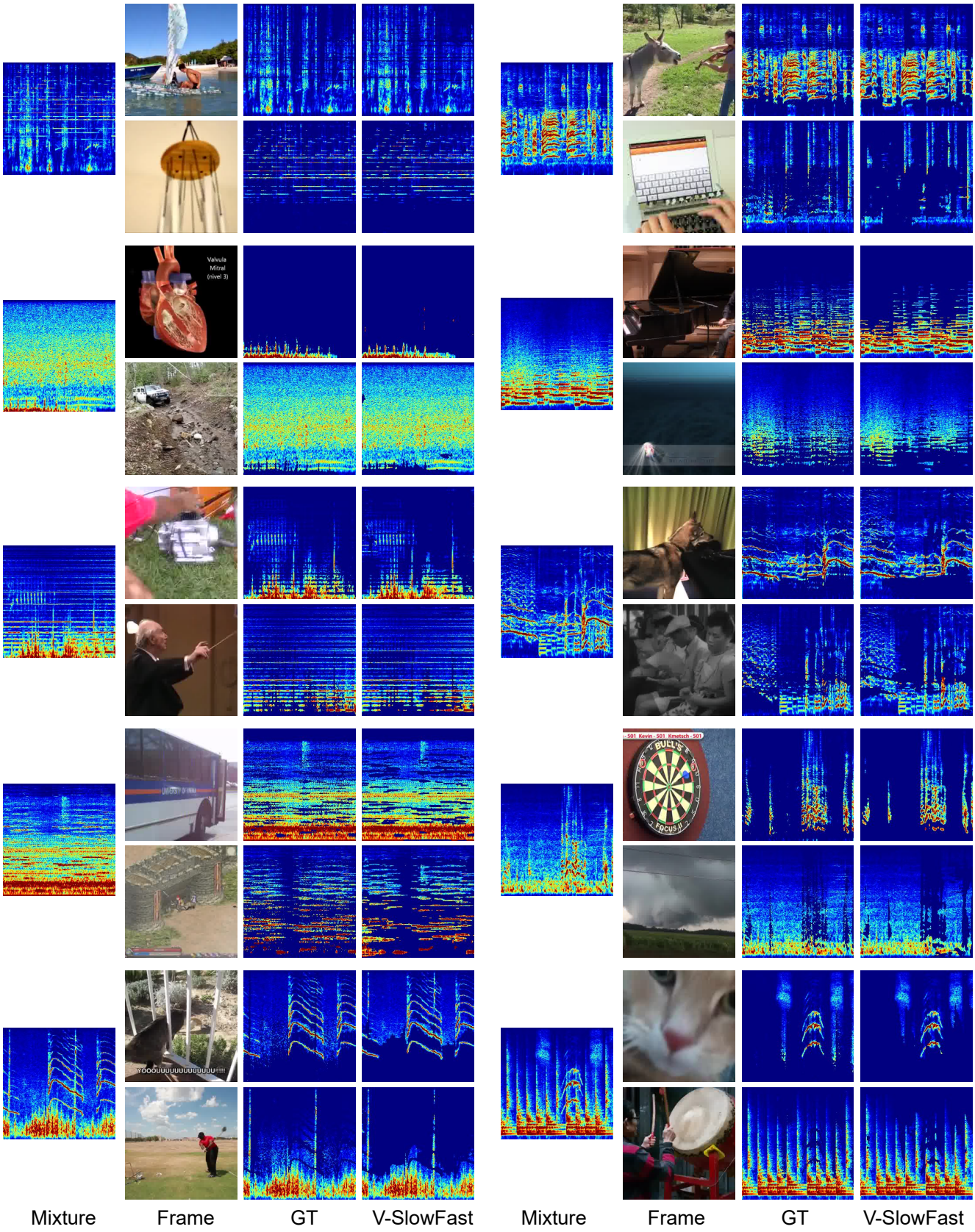


Figure E: Visualization of the source separation results using V-SlowFast network with mixtures of two sources from VGG-Sound dataset.

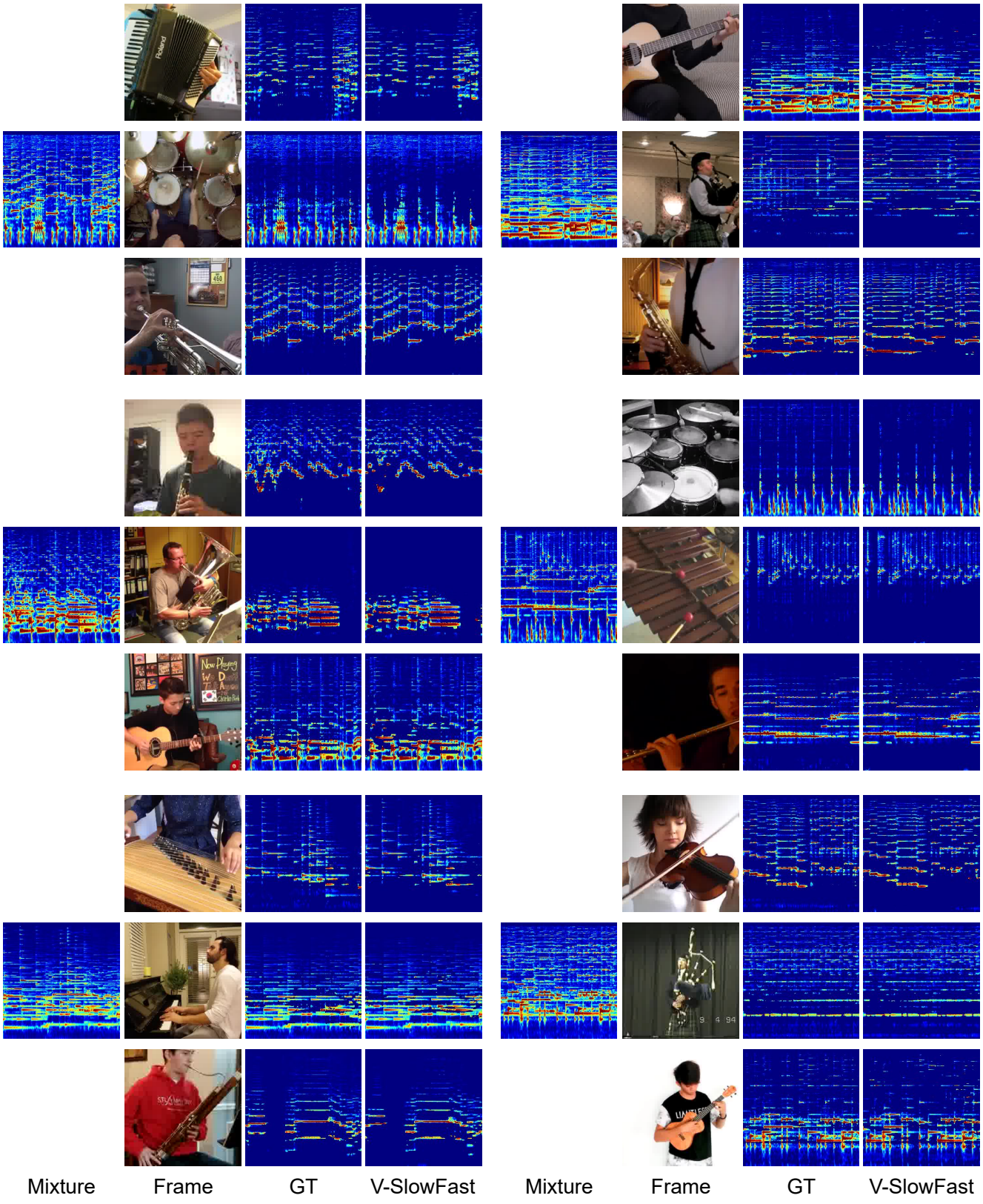


Figure F: Visualization of the source separation results using V-SlowFast network with mixtures of three sources from MUSIC-21 dataset.

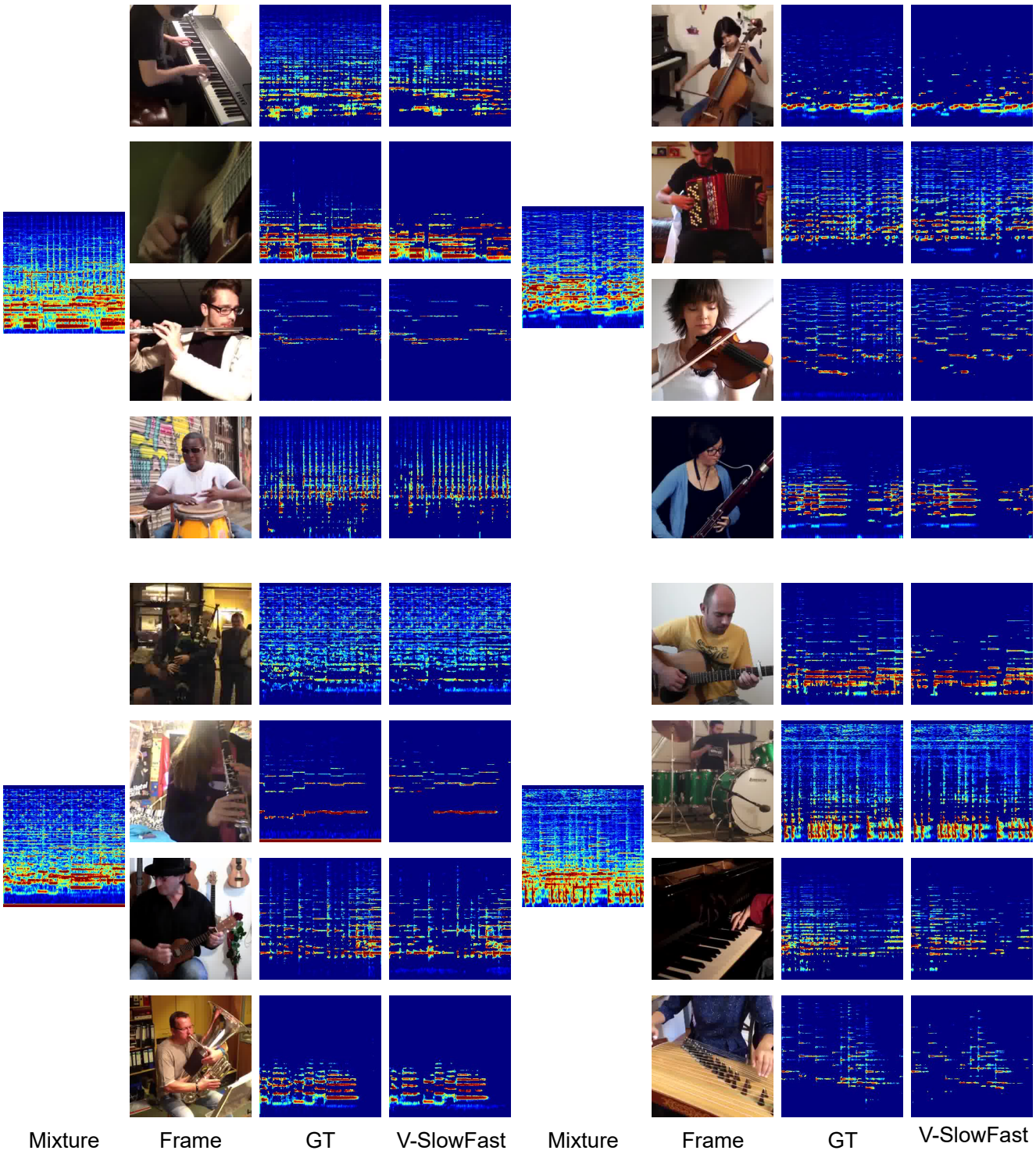


Figure G: Visualization of the source separation results using V-SlowFast network with mixtures of four sources from MUSIC-21 dataset.

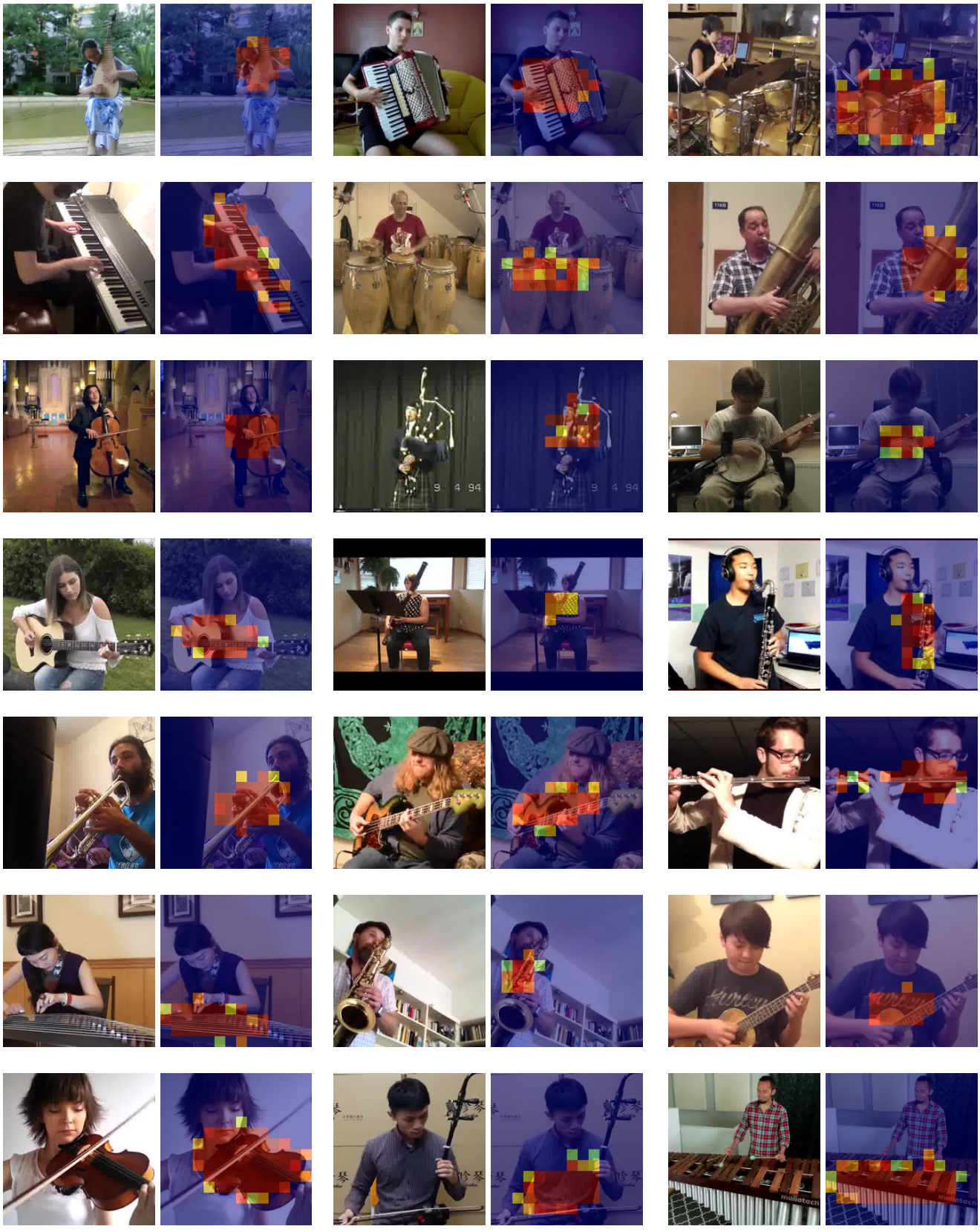


Figure H: Visualization of the sound source localization using V-SlowFast network from MUSIC-21 dataset.

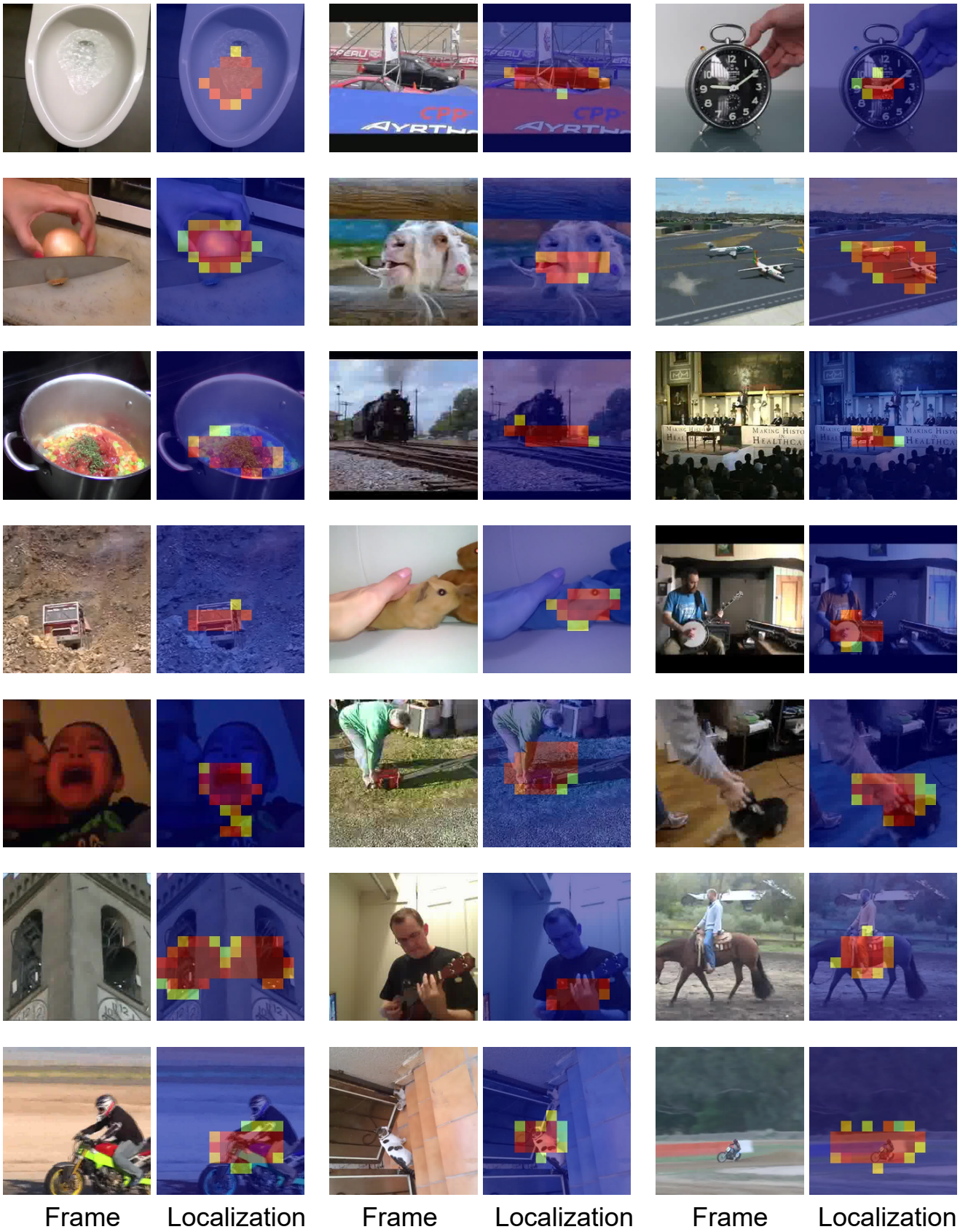


Figure I: Visualization of the sound source localization using V-SlowFast network from AVE dataset.

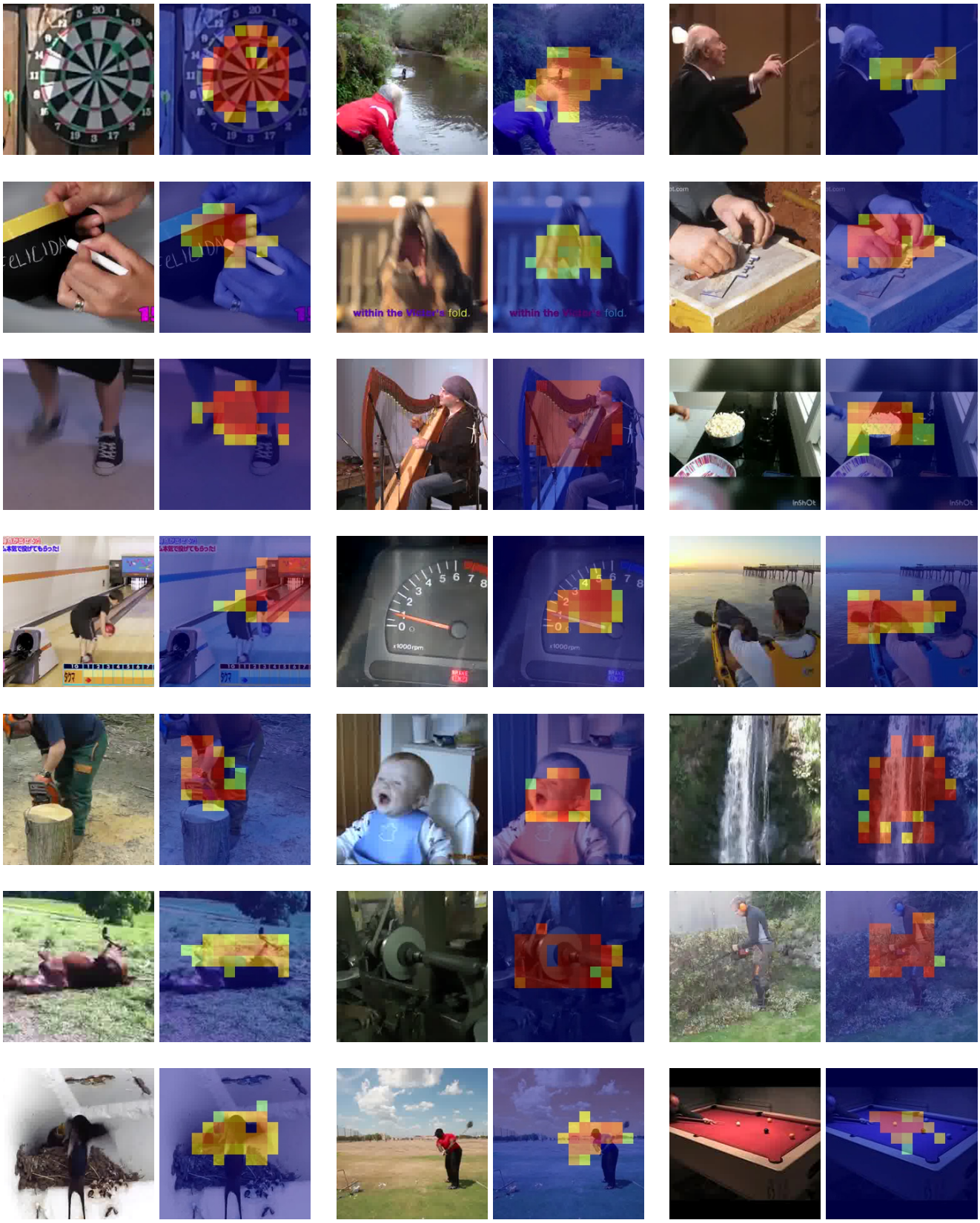


Figure J: Visualization of the sound source localization using V-SlowFast network from VGG-Sound dataset.