ICCV 2019 Seoul, Korea

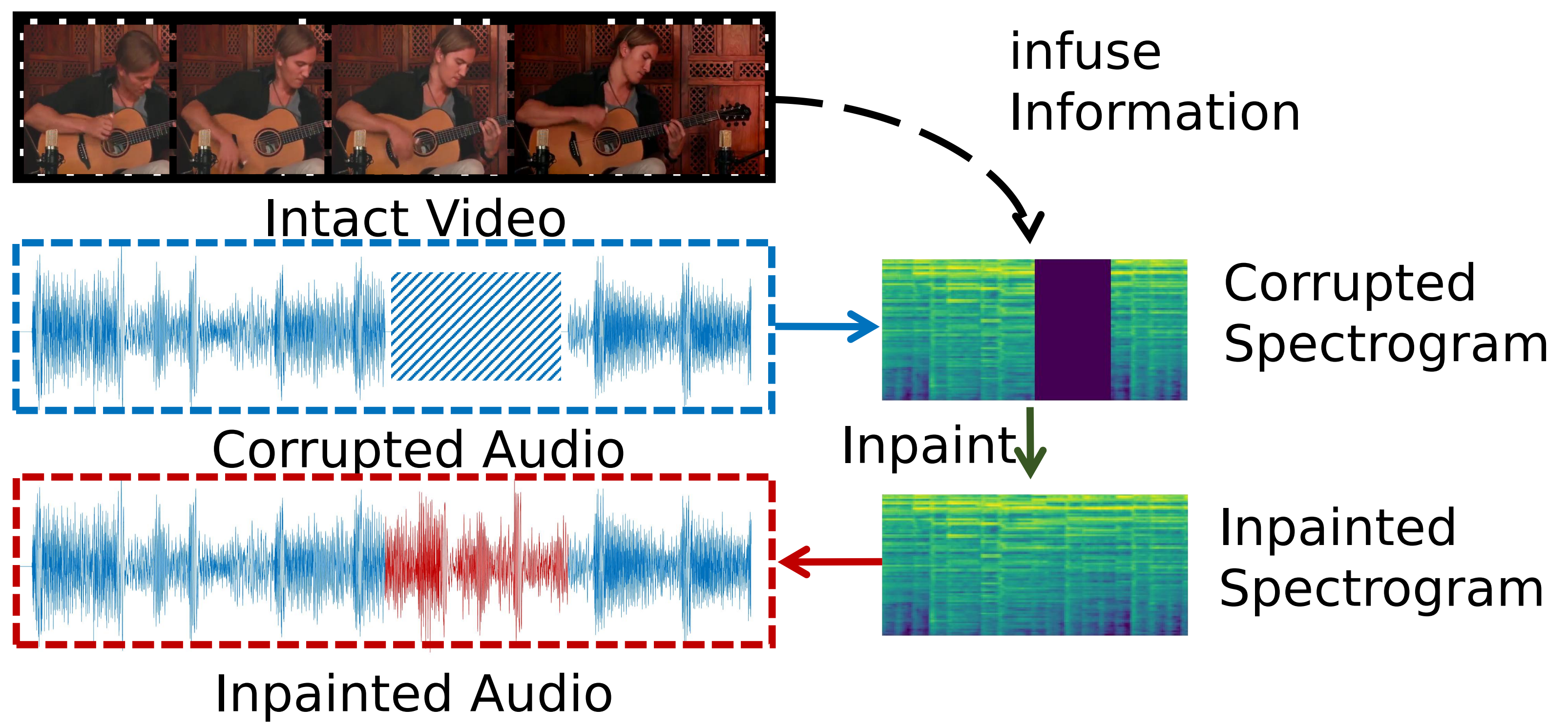
Vision-Infused Deep Audio Inpainting

Hang Zhou[1], Ziwei Liu[1], Xudong Xu[1], Ping Luo[2], Xiaogang Wang[1]

[1]The Chinese University of Hong Kong [2]The University of Hong Kong

Problem Description:

Synthesizing missing audio segments that correspond to their accompanying videos.

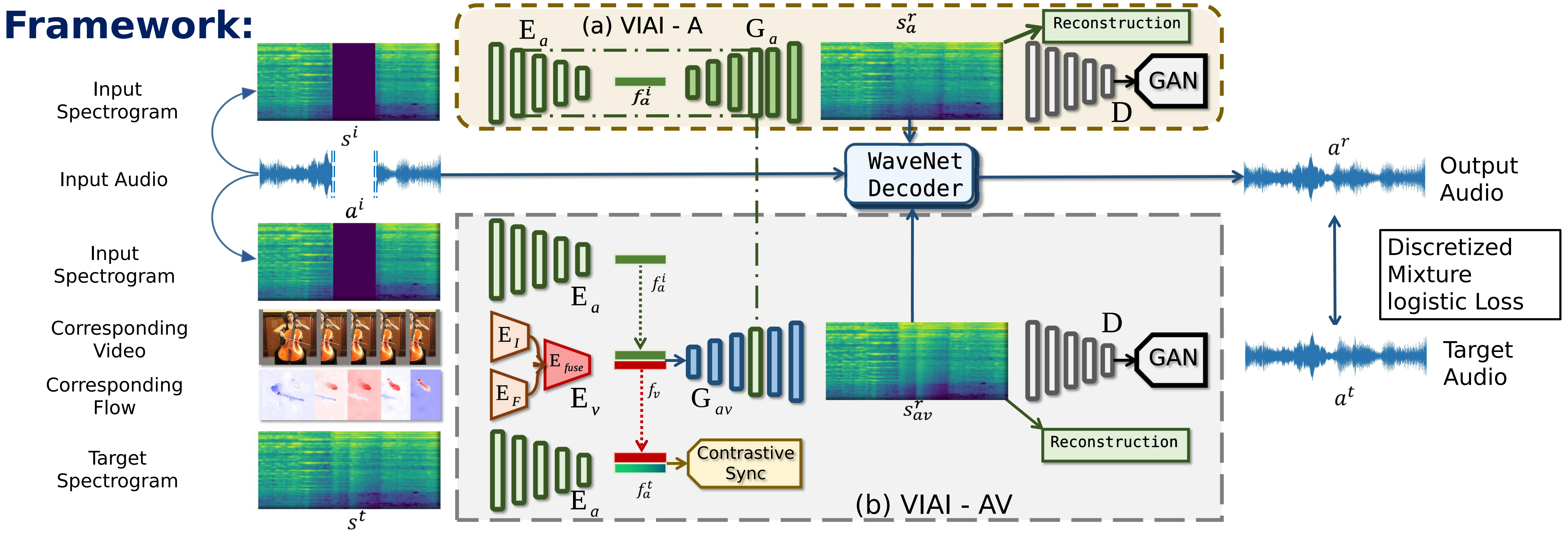


- It is easy to operate on spectrograms instead of raw audios.
- A visual-audio joint feature space needs to be learned.

Contribution:

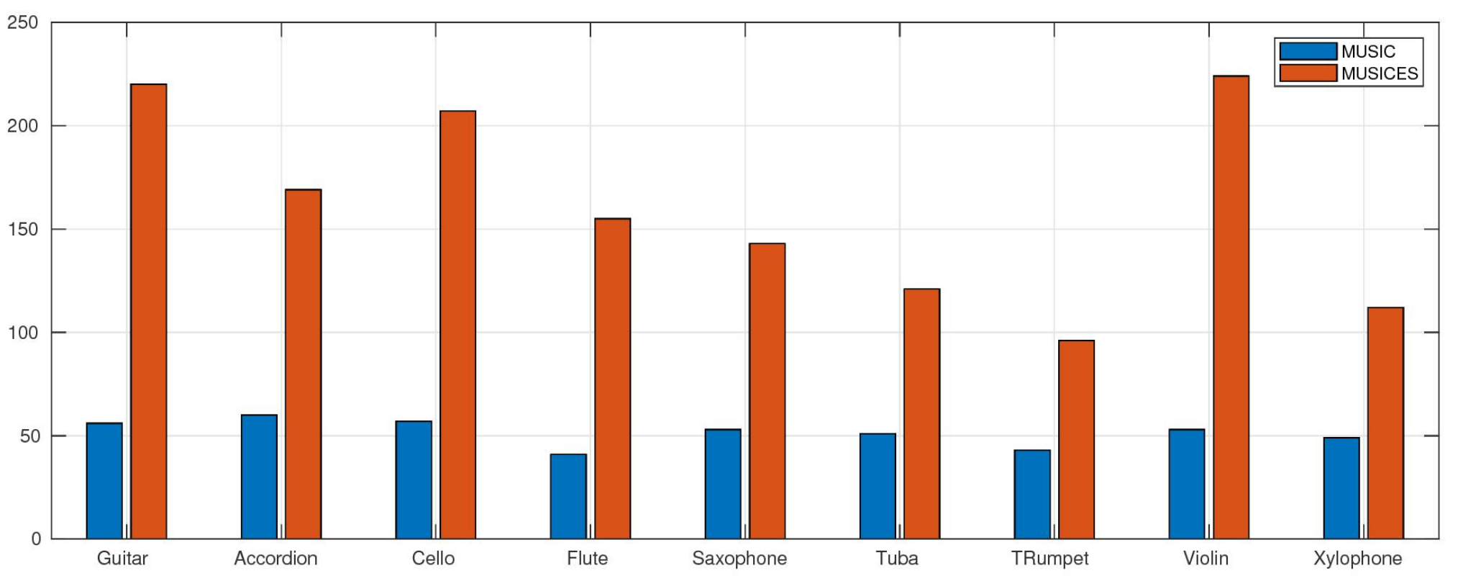
- Propose a novel framework for audio inpainting inspired by image inpainting.
- Design the first system targeting video-associated audio inpainting.
- Extend the original MUSIC dataset to a richer version, named MUSICES.

Framework:



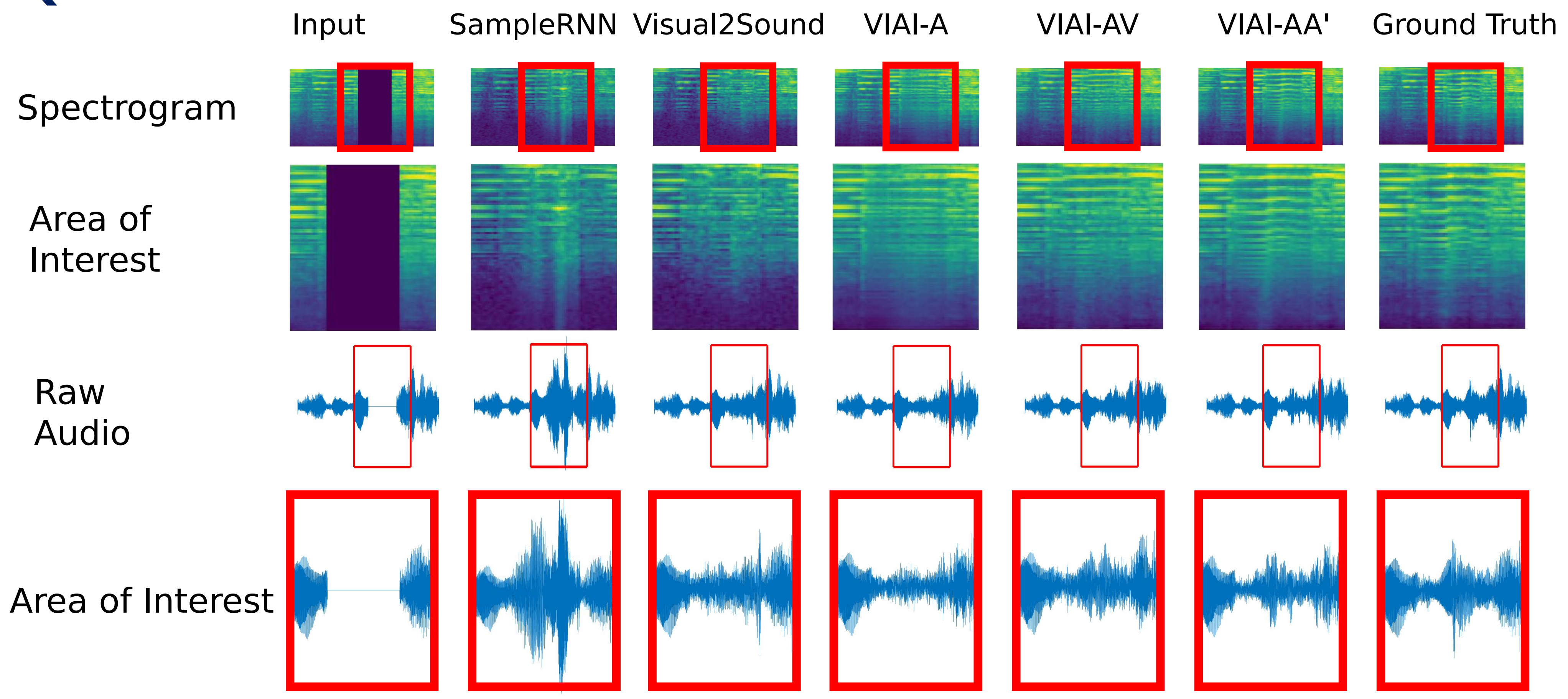
- **Audio Inpainting as Spectrogram Inpainting:** Audio-only inpainting system VIAI-A.
- **Joint Visual-Audio Spectrogram Inpainting:** Audio-Visual inpainting system VIAI-AV.
- **Modified WaveNet Decoder.**

Dataset:



- **MUSICES Dataset.**
- **9 major classes.** Enrich the original MUSIC dataset on 9 major classes for solo videos to approximately triple its size.
- **Shot-detection for video pre-processing.**

Qualitative Results:



Quantitative Results:

- Evalution with the class **cello.**

Score \ Approach	SampleRNN [32]	Visual2Sound [53]	bi-SampleRNN	bi-Visual2Sound	VIAI-A	**VIAI-AV**	**VIAI-AA’**
PSNR	9.1	10.2	12.8	13.6	22.2	**23.2**	**26.6**
SSIM	0.33	0.35	0.38	0.41	0.61	**0.64**	**0.75**
SDR	4.89	3.70	4.20	4.72	6.54	**6.63**	**6.89**
OPS	51.1	51.3	51.2	52.2	52.4	**56.3**	**56.7**

Table 1. Quantitative results. The upper half are the evaluations of spectrograms and the lower half are the evaluation of audios. The maximum of OPS is 100. Larger values are better among these metrics.

Table 2. Users’ Mean Opinion Scores. Lager is higher, with the maximum value to be 5.

MOS on \ Approach	SampleRNN [32]	Visual2Sound [53]	VIAI-A	**VIAI-AV**	**VIAI-AA’**
Audio Quality	2.51	2.20	3.05	**3.93**	**4.35**
Audio-Visual Coherence	2.22	2.23	3.02	**3.96**	**4.40**
Similarity with Target	2.35	2.20	2.97	**4.01**	**4.46**

Code and models: github.com/Hangz-nju-cuhk/Vision-Infused-Audio-Inpainter-VIAI