The Sound of Pixels

Aniruddh Doki, Feihong He, Yue Yang, Ce Zhang

Department of Computer Science, University of North Carolina at Chapel Hill

Paper Battle

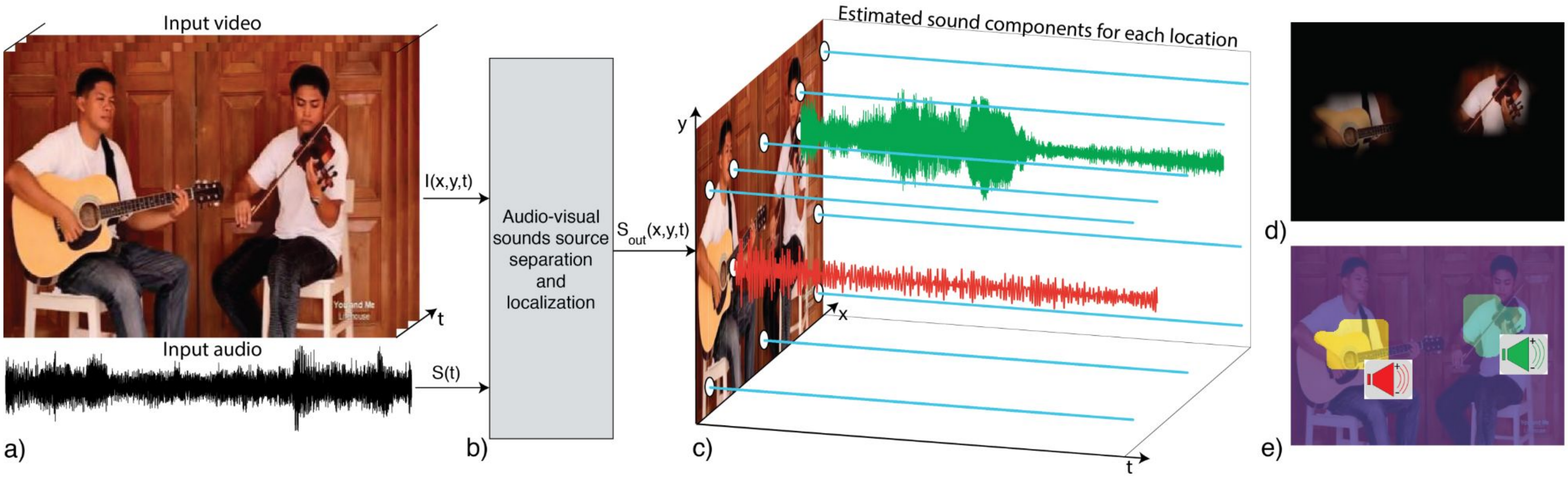
Oct 11st, 2023

Outlines

- Overview
  - Motivation and Introduction
  - Related work
  - Methods
  - Experiments and results

- Battle Part

Motivation and Introduction



http://sound-of-pixels.csail.mit.edu/

Related works

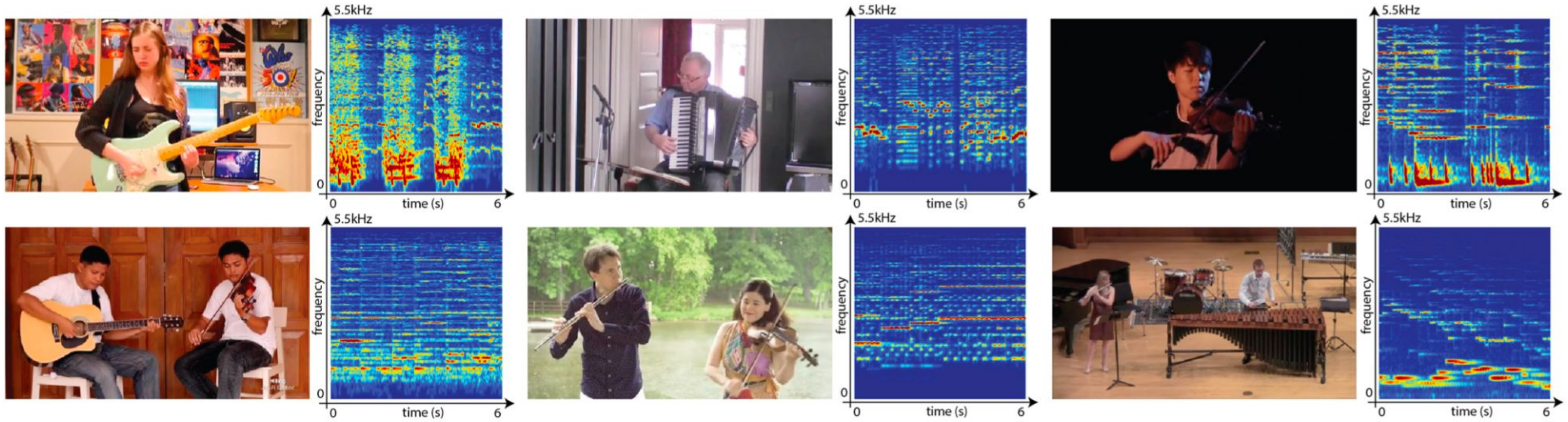
- Well explored
- Complex detection and tracking schemes [1, 2, 3, 4]
- Explicit modeling of motion to sound [5, 6, 7, 8]

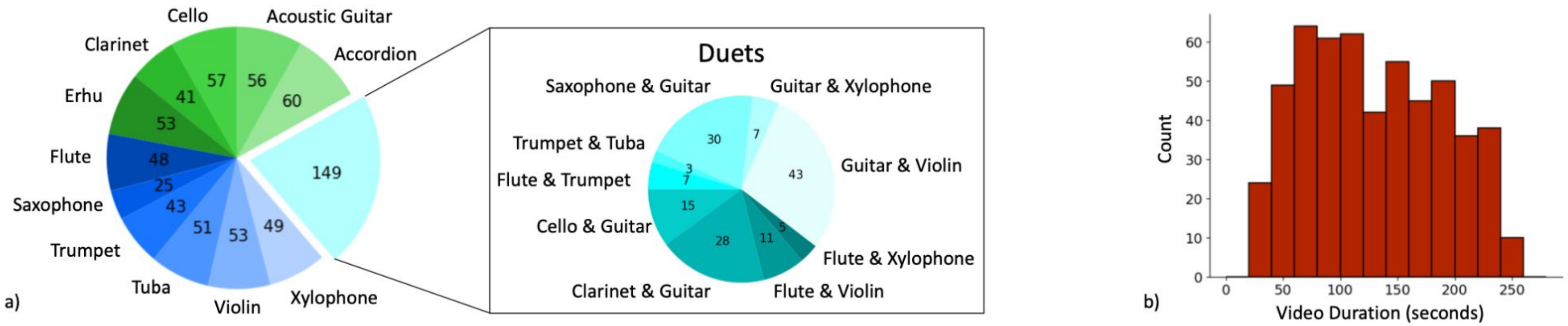
[1]. Vondrick C, Shrivastava A, Fathi A, et al. Tracking emerges by colorizing videos[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 391-408.
[2]. Wang, DeLiang, and Jitong Chen. "Supervised speech separation based on deep learning: An overview." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10 (2018): 1702-1726.
[3]. Wang, Xiaolong, and Abhinav Gupta. "Unsupervised learning of visual representations using videos." *Proceedings of the IEEE international conference on computer vision*. 2015.
[4]. Zhao, Mingmin, et al. "Through-wall human pose estimation using radio signals." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
[5]. Ma, Wei-Chiu, et al. "Single image intrinsic decomposition without a single intrinsic image." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
[6]. Haykin, Simon, and Zhe Chen. "The cocktail party problem." *Neural computation* 17.9 (2005): 1875-1902.
[7]. Mesaros, Annamaria, et al. "DCASE 2017 challenge setup: Tasks, datasets and baseline system." *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*. 2017.
[8]. Nagrani, Arsha, Samuel Albanie, and Andrew Zisserman. "Seeing voices and hearing faces: Cross-modal biometric matching." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Method

MUSIC Dataset: (Multimodal Sources of Instrument Combinations)

- 685 untrimmed videos of musical solos and dutes
- 11 instrument categories
- average duration: 2 min





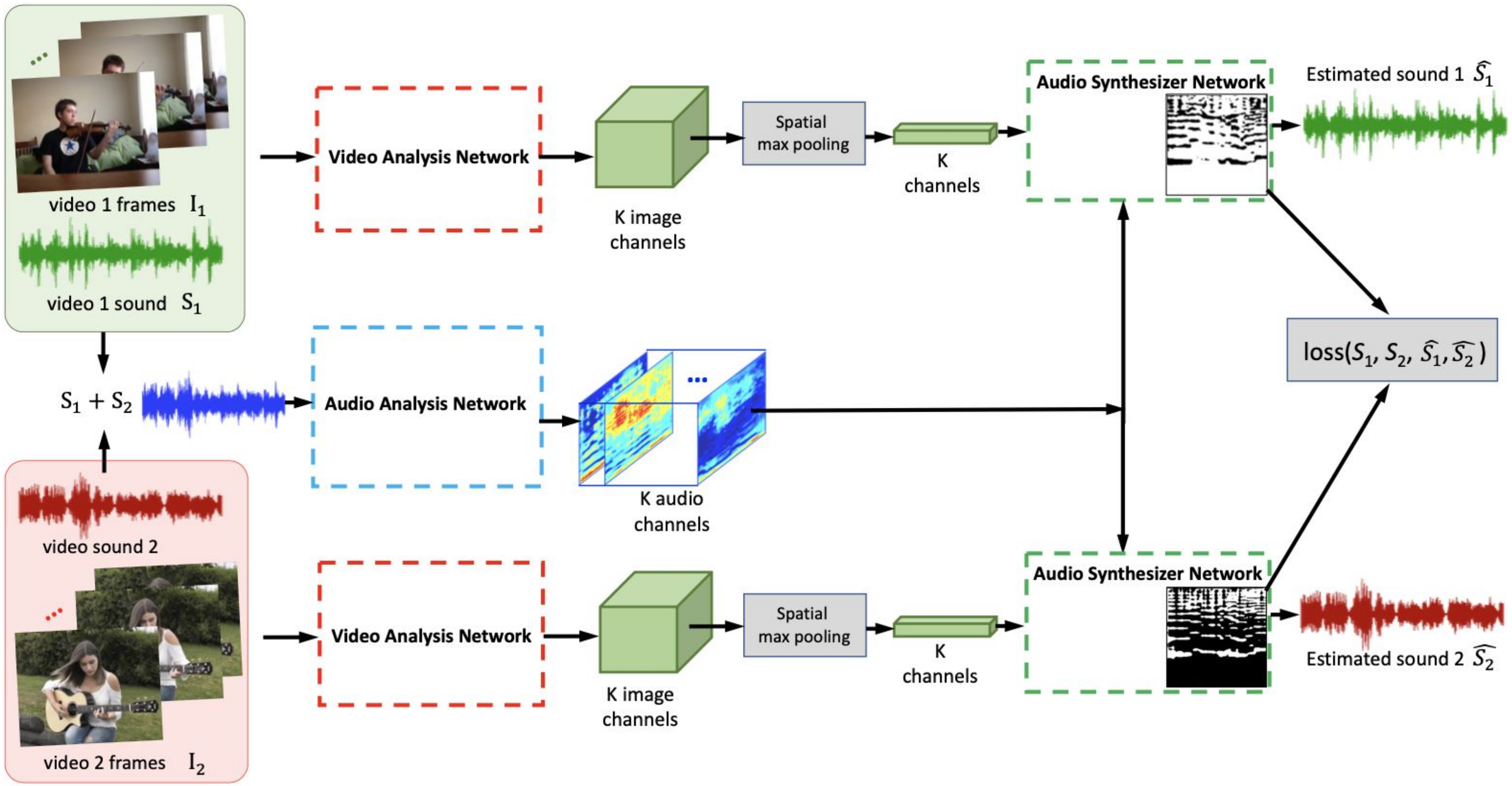
Mix-and-Separate framework for Self-supervised Training

- Input:
  - Artificially create audio mixture (add together)
  - two video frames
- Output:
  - two estimated sounds
- Mask(for each T-F unit):
  - binary: whether the target sound is the dominant component in the mixed sound

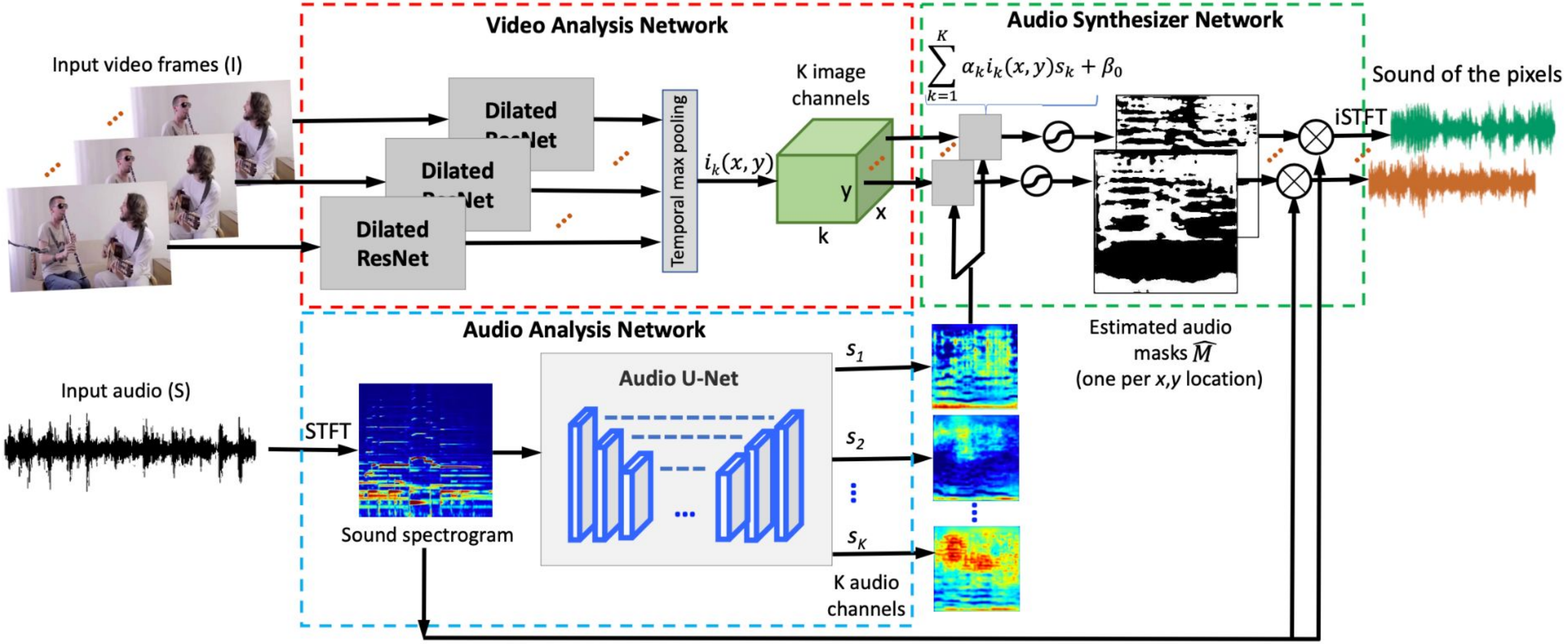
$$M_n(u,v) = [\![S_n(u,v) \geq S_m(u,v)]\!], \quad \forall m = (1, ..., N),$$

  - ratio:ground truth mask of a video is calculated as the ratio of the magnitudes of the target sound and the mixed sound

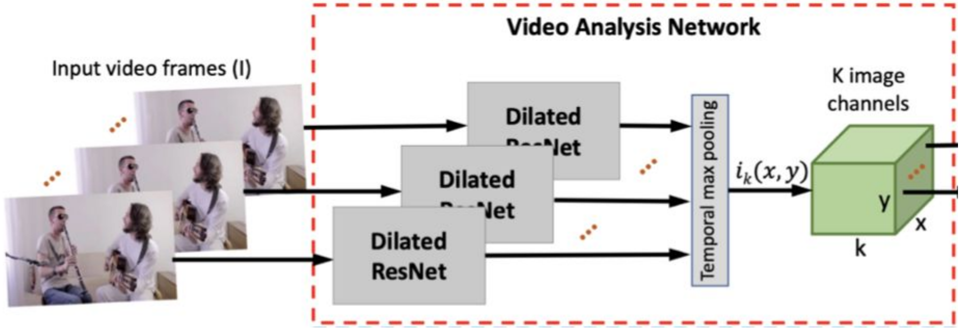
$$M_n(u,v) = \frac{S_n(u,v)}{S_{mix}(u,v)}.$$



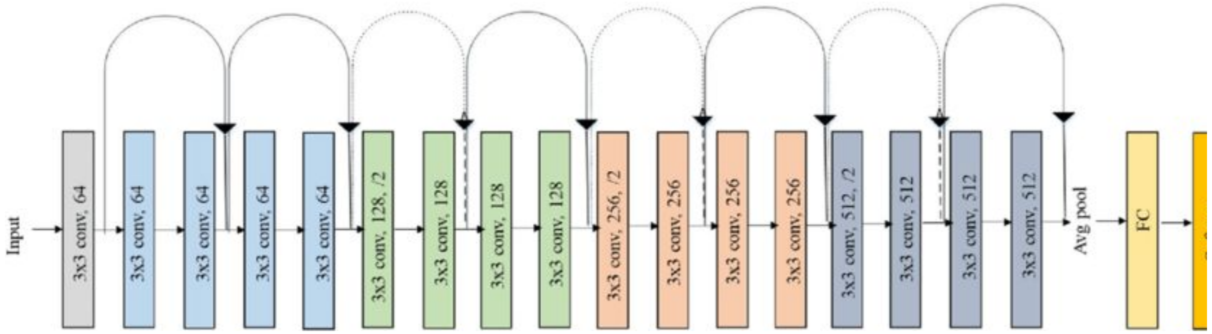
Method: Overview



Video Analysis Network:
Dilated ResNet

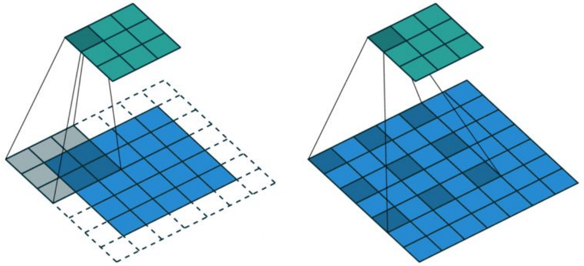


ResNet-18



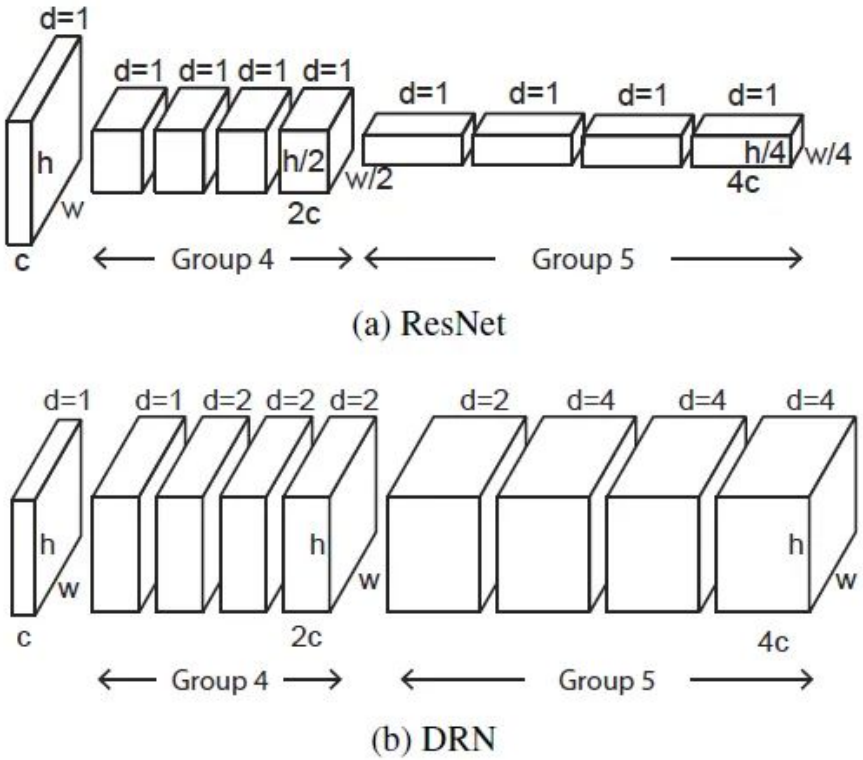
- **Input: T*W*H*3 video frame**
- **output: k image channels**

What is dilated?

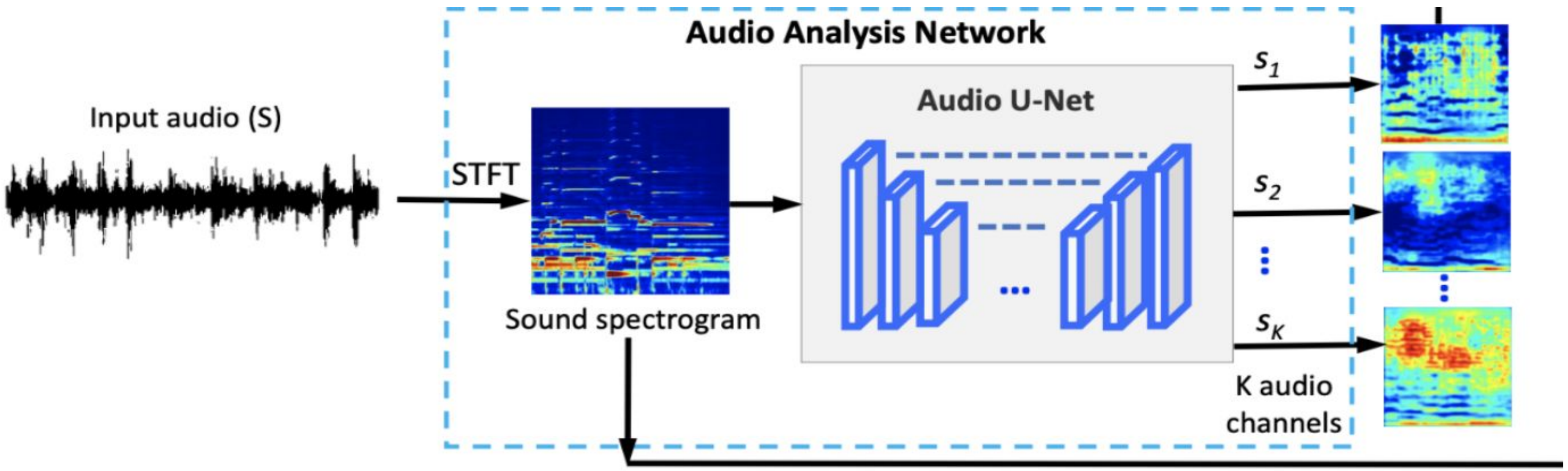


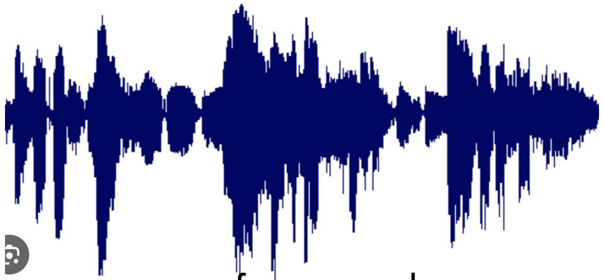
Standard Convolution (l=1) (Left) Dilated Convolution (l=2) (Right)

more modification will be in experiment part~



Audio Analysis Network: (Input)
Input audio–STFT→Sound spectrogram

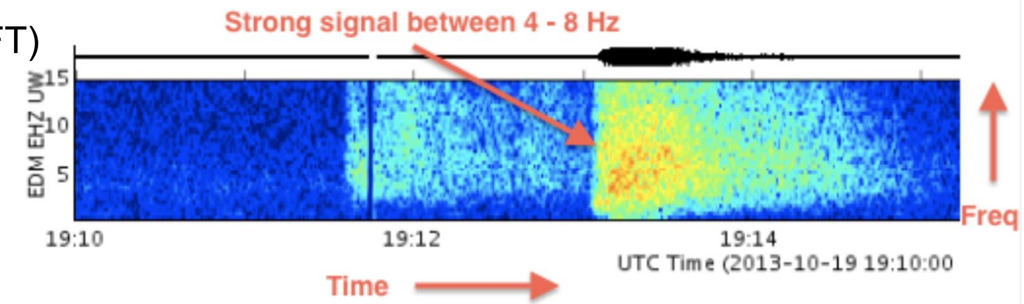




waveform sound

Short Time Fourier Transform (STFT)

512 × 256 Time-Frequency (T-F) representation of the sound

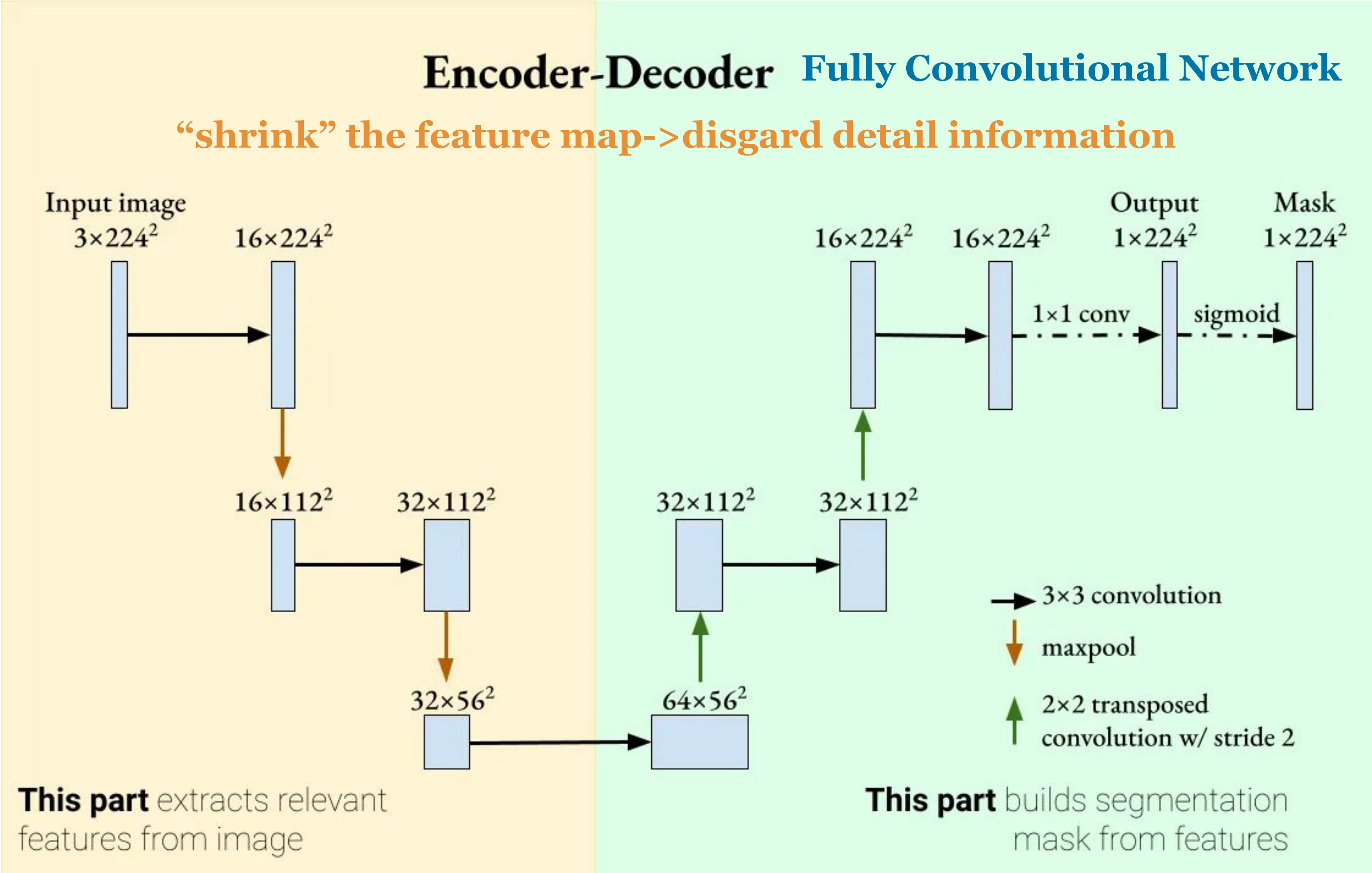


log transformation

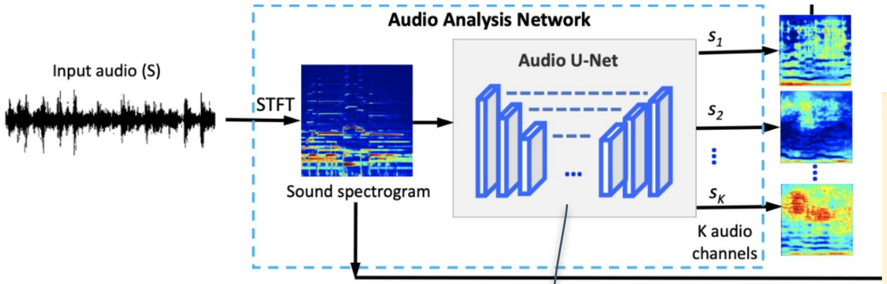
Input sound spectrogram

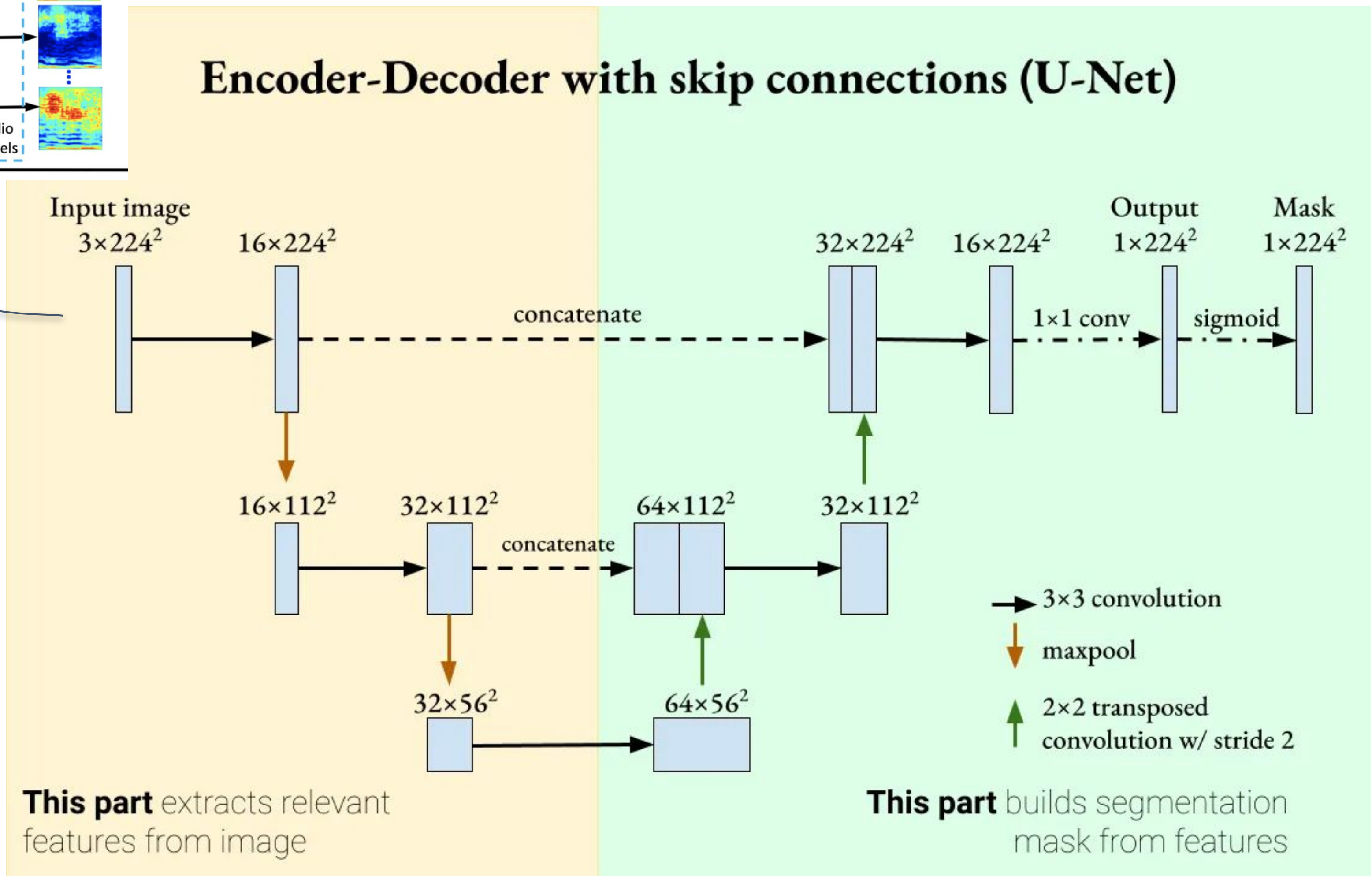
A spectrogram is a figure which represents the spectrum of frequencies of a recorded audio over time.

Audio Analysis Network: U-Net (task of semantic segmentation)



Audio Analysis Network: U-Net (task of semantic segmentation)





- **Keep both detail and general information**
- **Input: Audio Spectrogram**
- **Output: K audio channels**

Experiments

Sound Separation:

Given two videos and the mixture of the two corresponding audios, separate the audios from the mixture.

Visual Grounding of Sounds:

- Which pixels are making sounds?
- What sounds do these pixels make?
- Is the sound coming from this pixel?

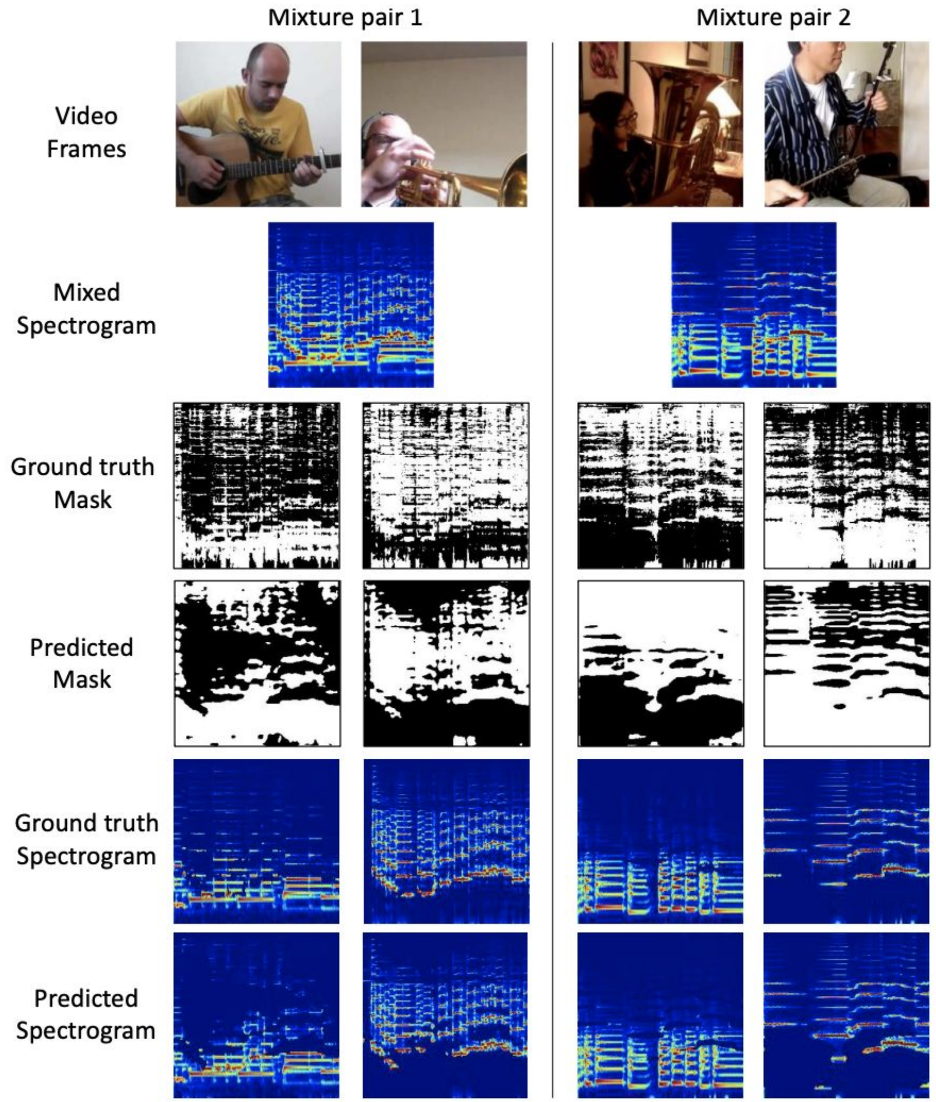
Experiments

Sound Separation:

Given two videos and the mixture of the two corresponding audios, separate the audios from the mixture.

	NMF [42]	DeepConvSep [7]	Spectral Regression	Ratio Mask Linear scale	Ratio Mask Log scale	Binary Mask Linear scale	Binary Mask Log scale
NSDR	3.14	6.12	5.12	6.67	8.56	6.94	**8.87**
SIR	6.70	8.38	7.72	12.85	13.75	12.87	**15.02**
SAR	10.10	11.02	10.43	13.87	**14.19**	11.12	12.28

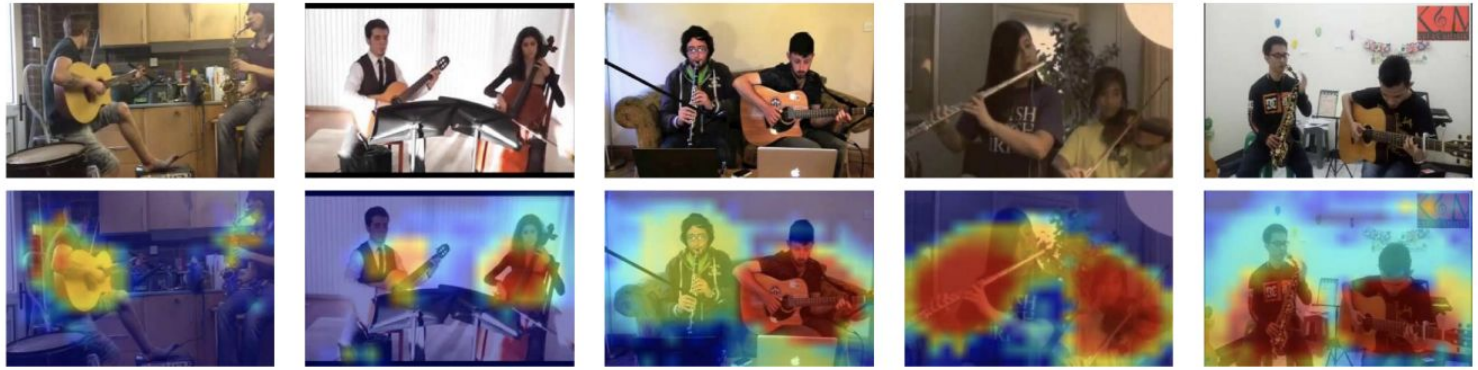
Table 1. Model performances of baselines and different variations of our proposed model, evaluated in NSDR/SIR/SAR. Binary masking in log frequency scale performs best in most metrics.



Experiments

Visual Grounding of Sounds:

Which pixels are making sounds?



What sounds do these pixels make?

Is the sound coming from this pixel?
- Select 256 pixel positions (50% on instruments and 50% on background objects)
- Generate sound from those pixels
- Ask Amazon AMT workers: ‘Yes’ if they hear

Model	Yes(%)
Spectral Regression	39.06
Ratio Mask	54.68
Binary Mask	67.58

Thank you!