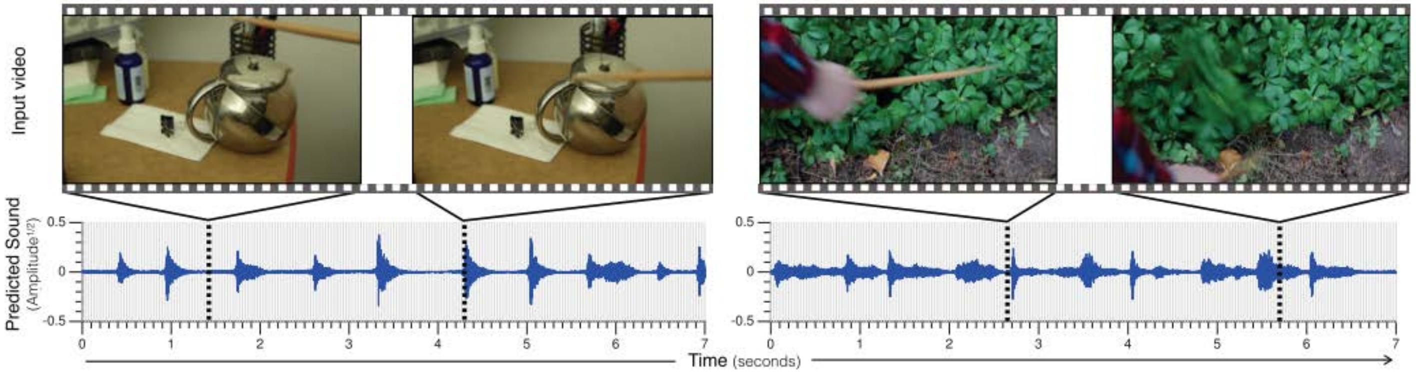
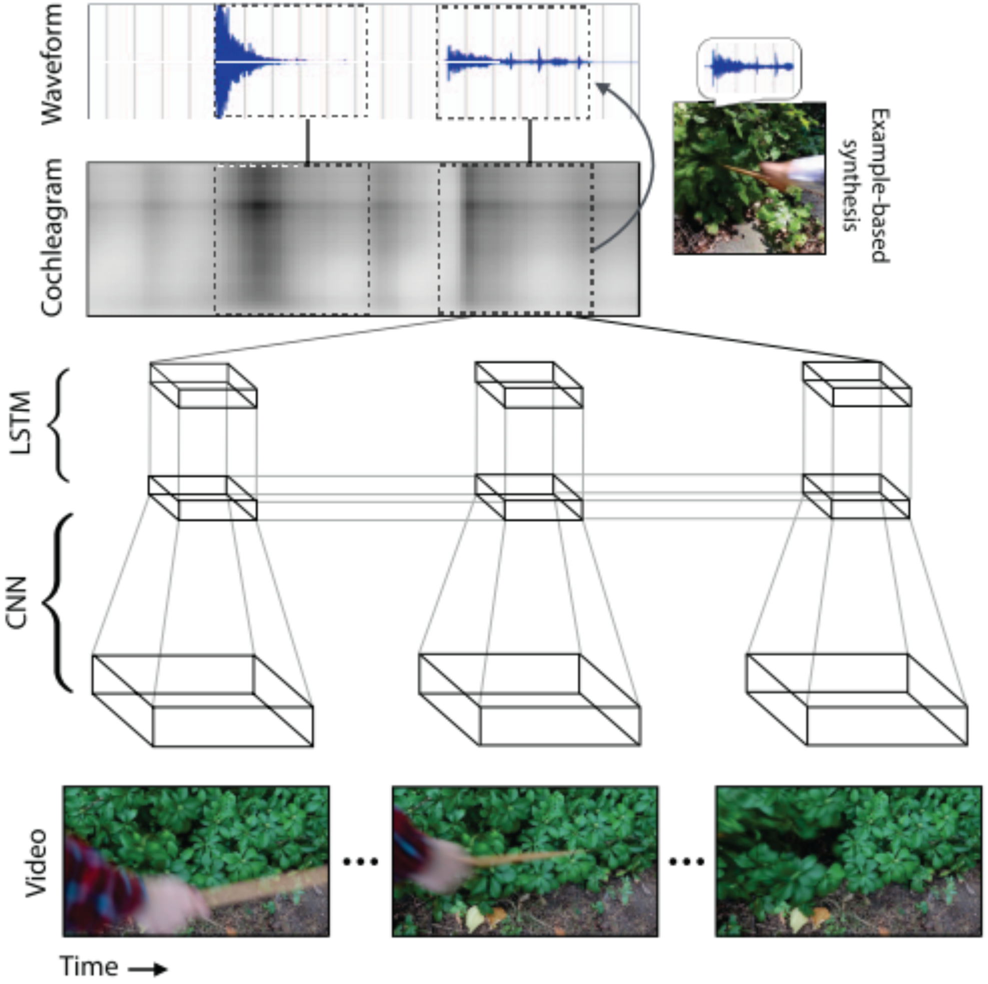
The sound of Pixels

ECCV 2018

Related work

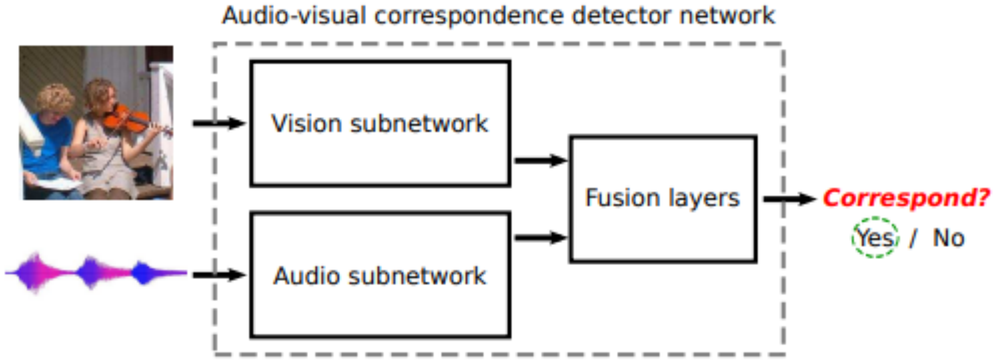
- Visually indicated sounds
  - Auditory is not as sensitive as visual

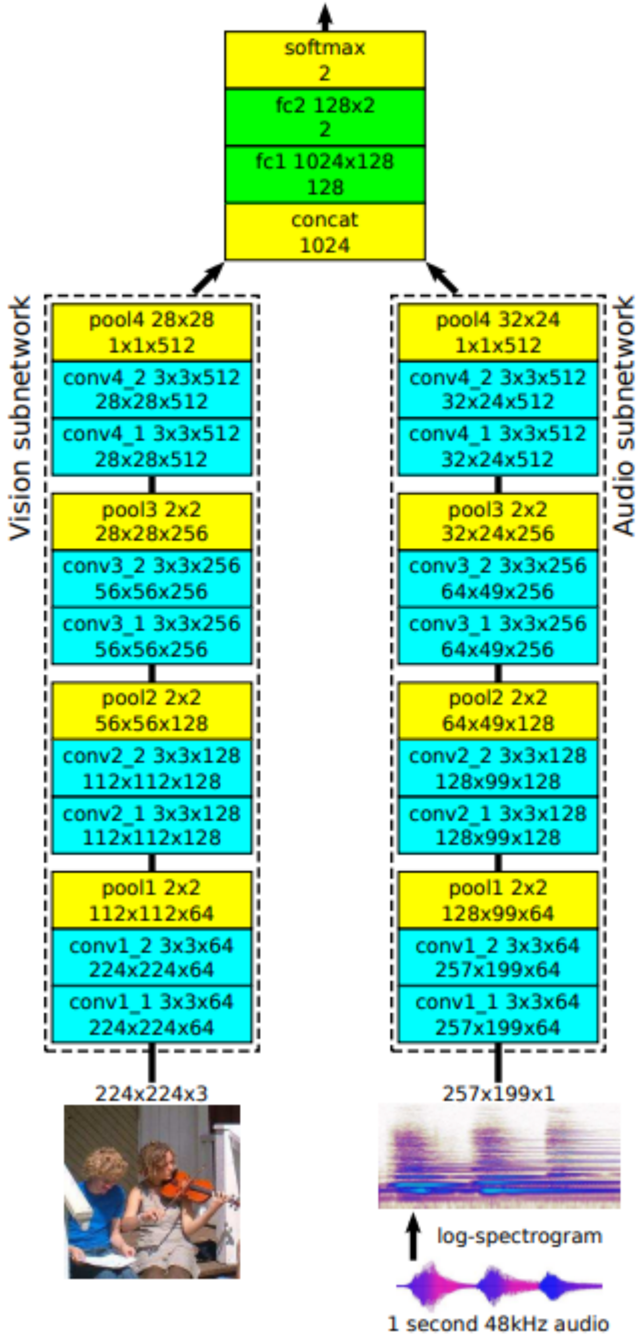




Related work

- Look, Listen and Learn

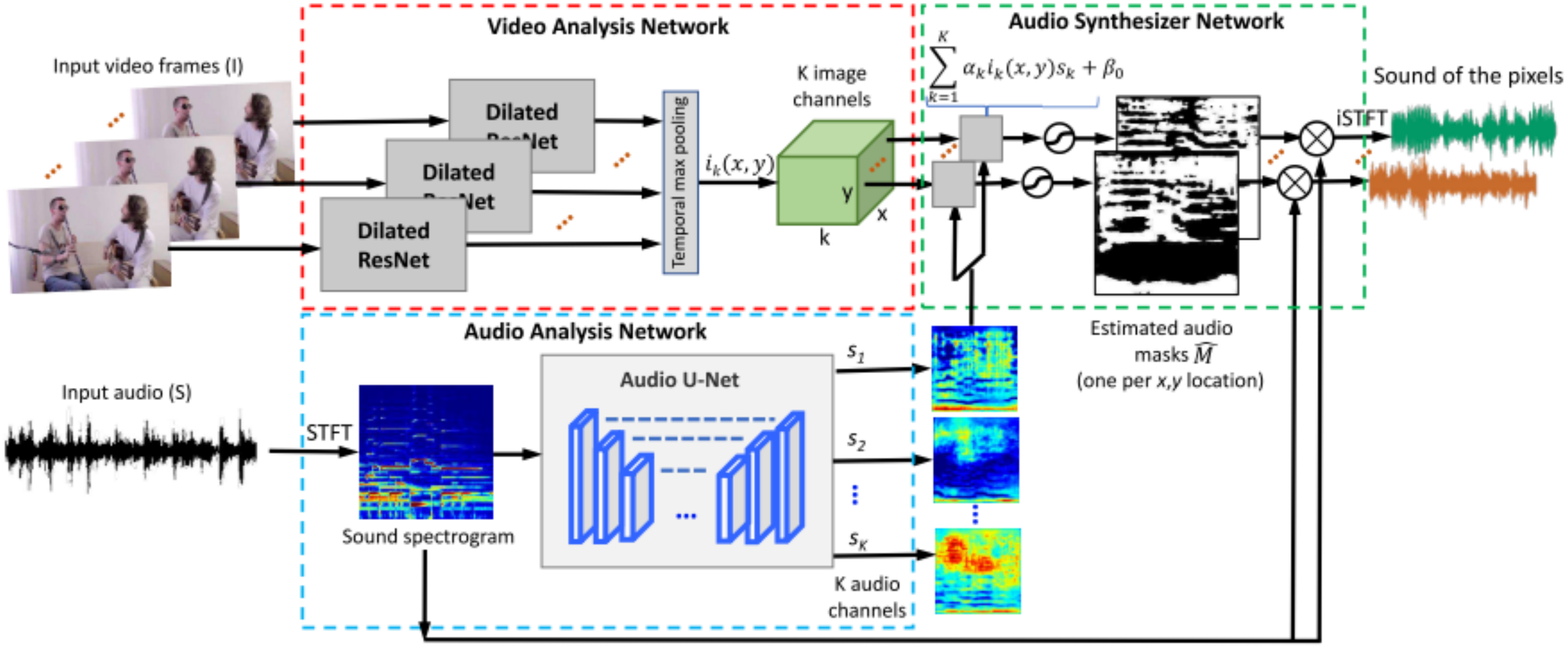




Related work

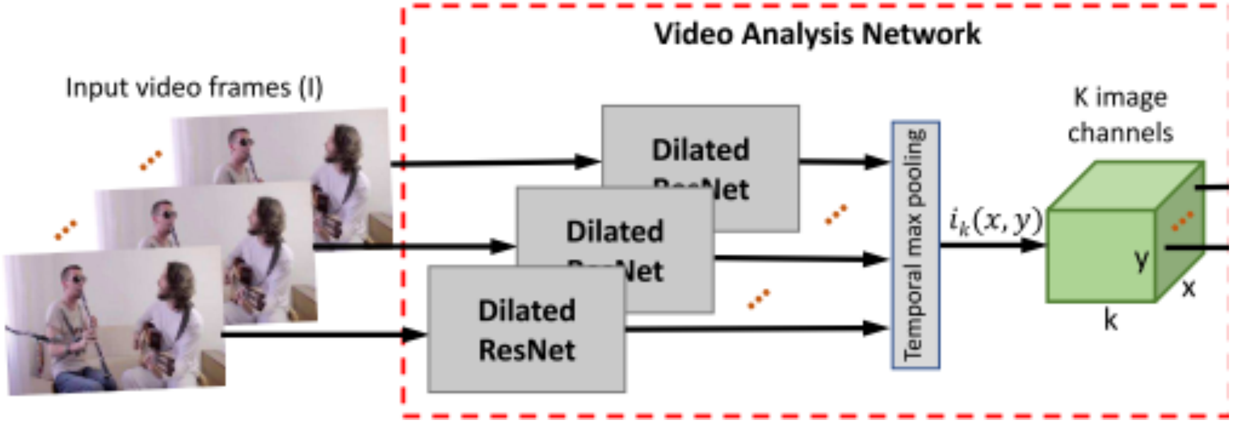
- Sound source separation
  - https://sisec18.unmix.app/#/
- Demo
  - http://sound-of-pixels.csail.mit.edu/

Model architecture



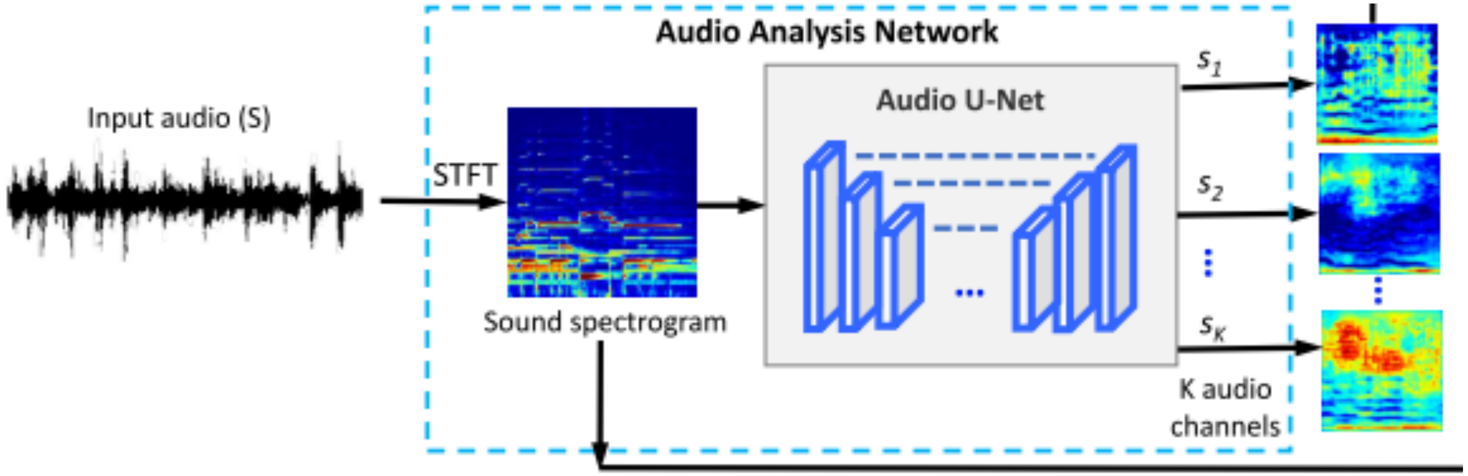
Model architecture

- Video analysis network
  - Input: T by H by W by 3
  - Output: T by (H/16) by (W/16) by K
  - Extract per-frame features
    and apply temporal pooling / sigmoid, denoted as $i_k(x, y)$
  - Dilated ResNet is used (can be replaced with other model)



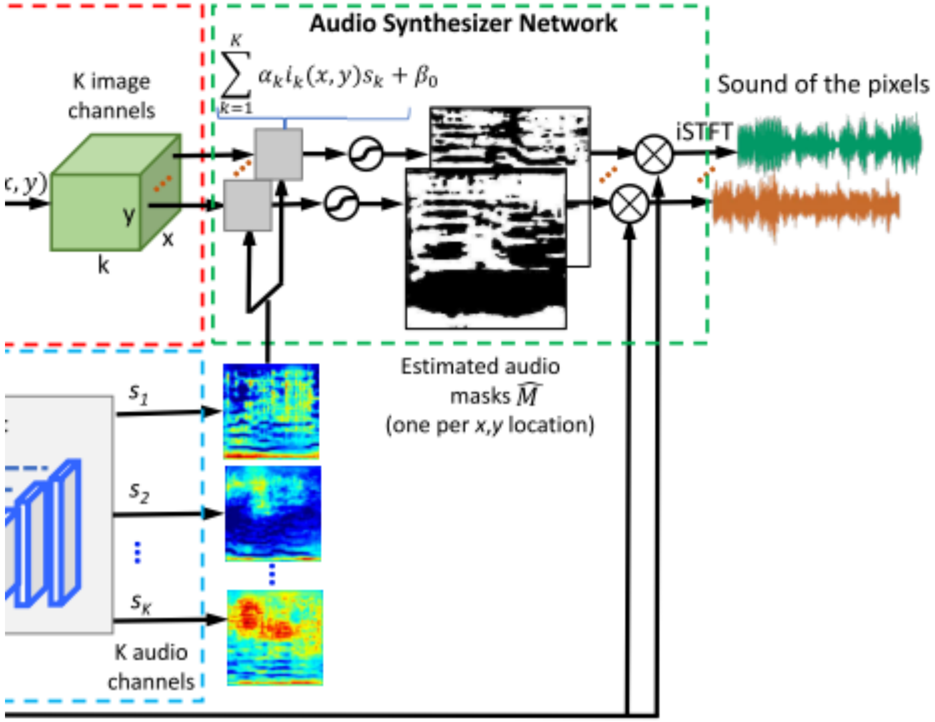
Model architecture

- Audio analysis network
  - Split input audio (log-spectrogram) into K components s_k
  - Based on U-net architecture



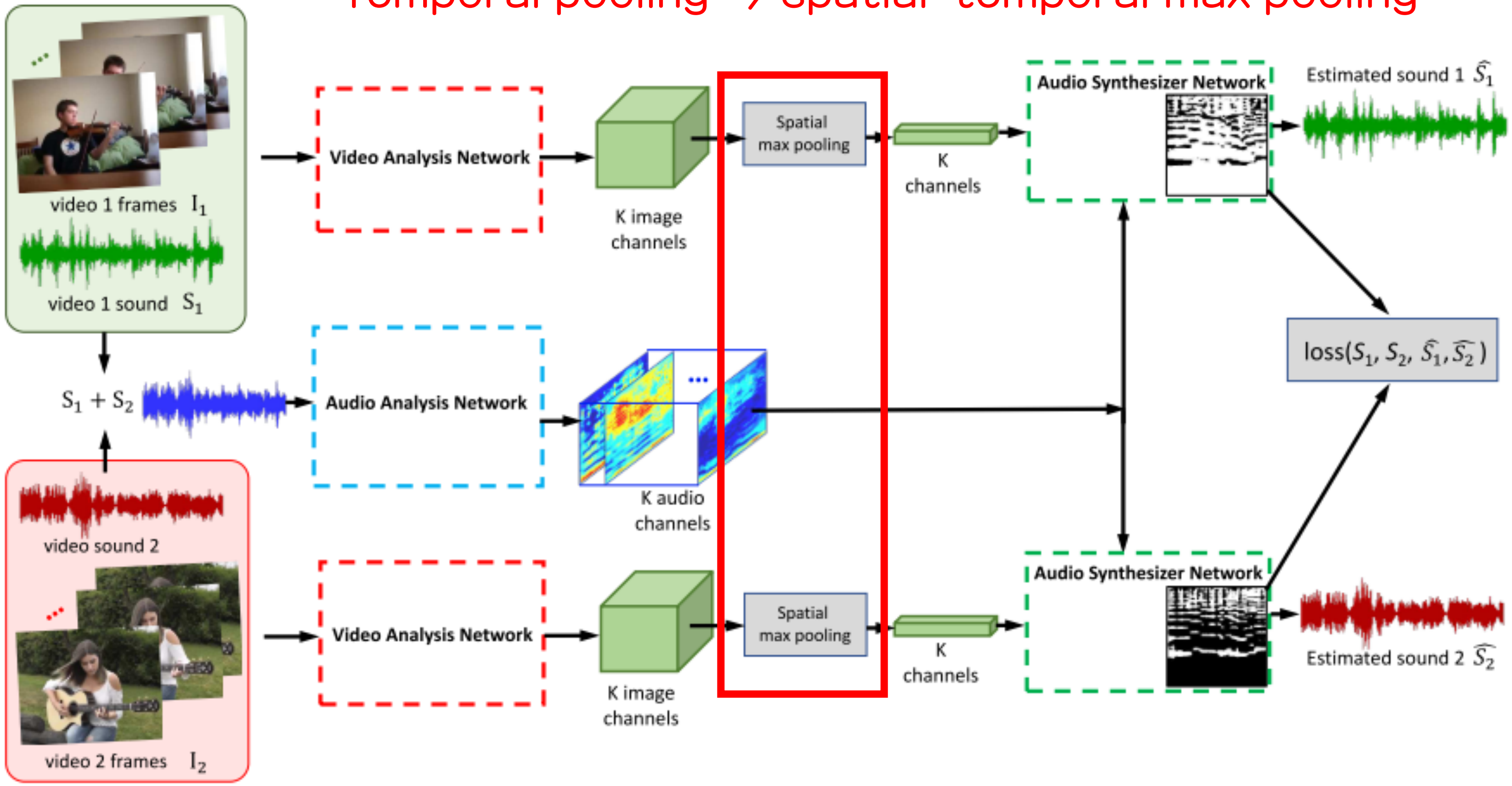
Model architecture

- Sound synthesizer network
  - The output sounds could be separated from masks
  - A mask $M(x,y) = \sigma(\sum_{k=1}^{K} \alpha_k i_k(x,y) s_k + \beta_0)$



Training

Temporal pooling -> spatial-temporal max pooling



Training

- Objective
  - Binary mask: per-pixel sigmoid cross entropy loss

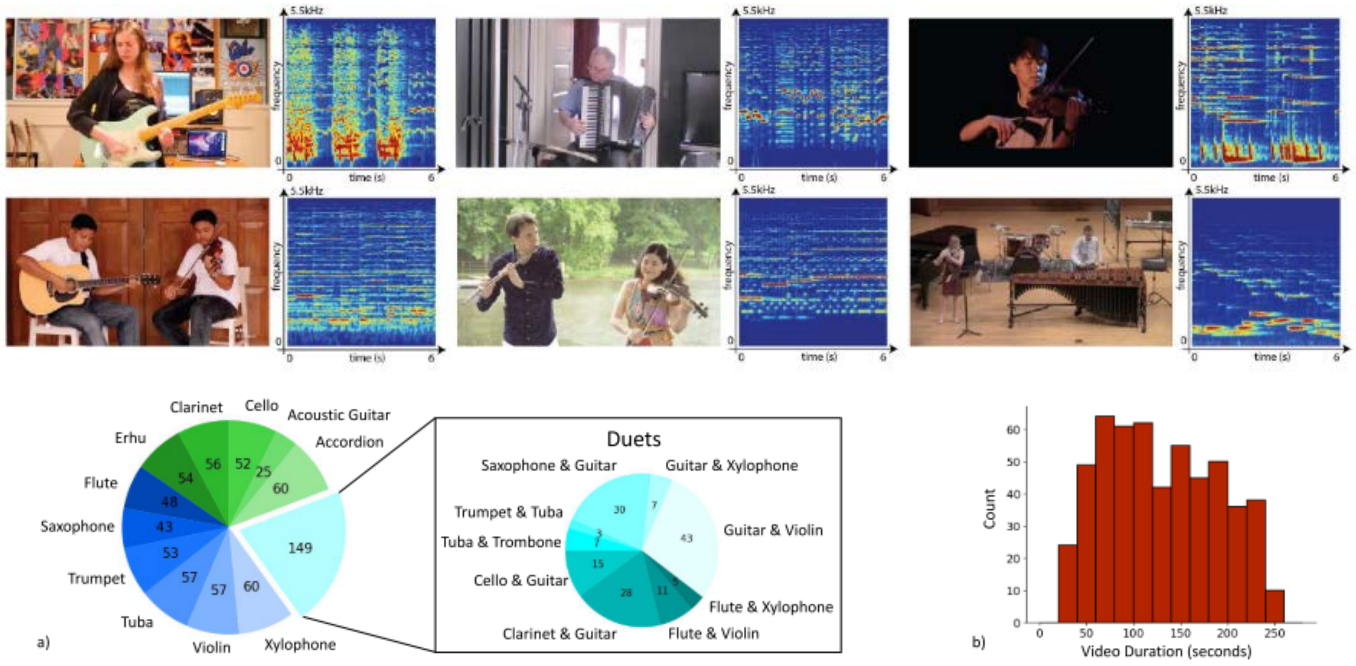
$$M_n(u,v) = [\![s_n(u,v) \geq s_m(u,v)]\!], \forall m = (1, \ldots, N)$$

  - Ratio mask: per-pixel L1 loss

$$M_n(u,v) = \frac{s_n(u,v)}{s_{mix}(u,v)}$$

Dataset

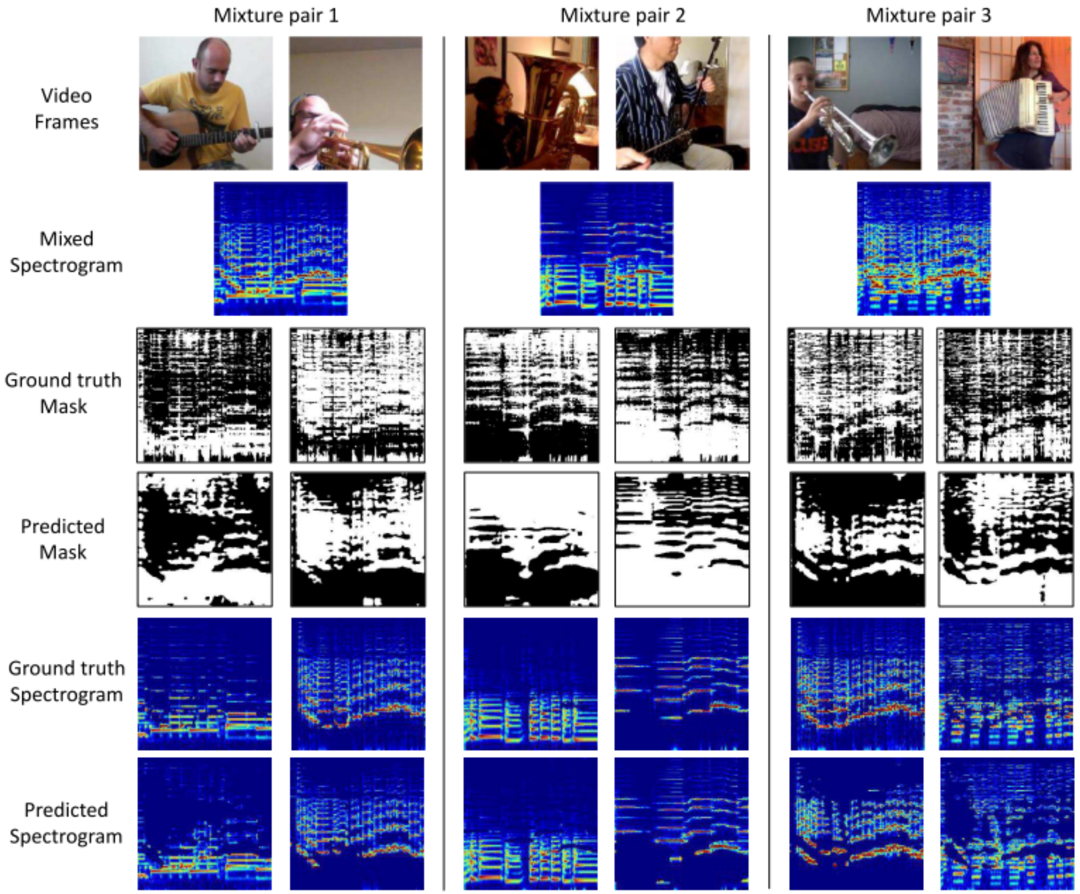
- MUSIC (Multimodal Sources of Instrument Combinations)
  - Videos from Youtube
  - 714 untrimmed videos of musical solos and duets



Dataset

- Preprocessing
  - Sampling rate is 11kHz, the highest frequency is 5.5kHz (resampled from higher sampling rate signal)
  - Length of each audio sample is approximately 6s
  - STFT parameter: window size 1024, hop length 256
  - STFT output shape is 512 by 256,
    and resample on a log-frequency scale to obtain a 256 by 256

Experiments



	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.

- Source separation
  - Normalized Signal-to-Distortion Ratio (NSDR)
  - Signal-to-Interference Ratio (SIR)
  - Signal-to-Artifact Ratio (SAR)

Experiments

- Visual grounding of sounds
  - Sound localization
    - “Which pixels are making sounds?”
    - Calculate the sound energy (or volume) of each pixel in the image

Experiments

- Visual grounding of sounds
  - Clustering of sounds
    - “What sounds do these pixels make?”
    - Apply PCA on log-spectrogram (output dimension is 3, RGB)