Learning to Separate Object Sounds by Watching Unlabeled Video

Ruohan Gao[1], Rogerio Feris[2], and Kristen Grauman[1,3]
[1]The University of Texas at Austin [2]IBM Research [3]Facebook AI Research

ECCV 2018

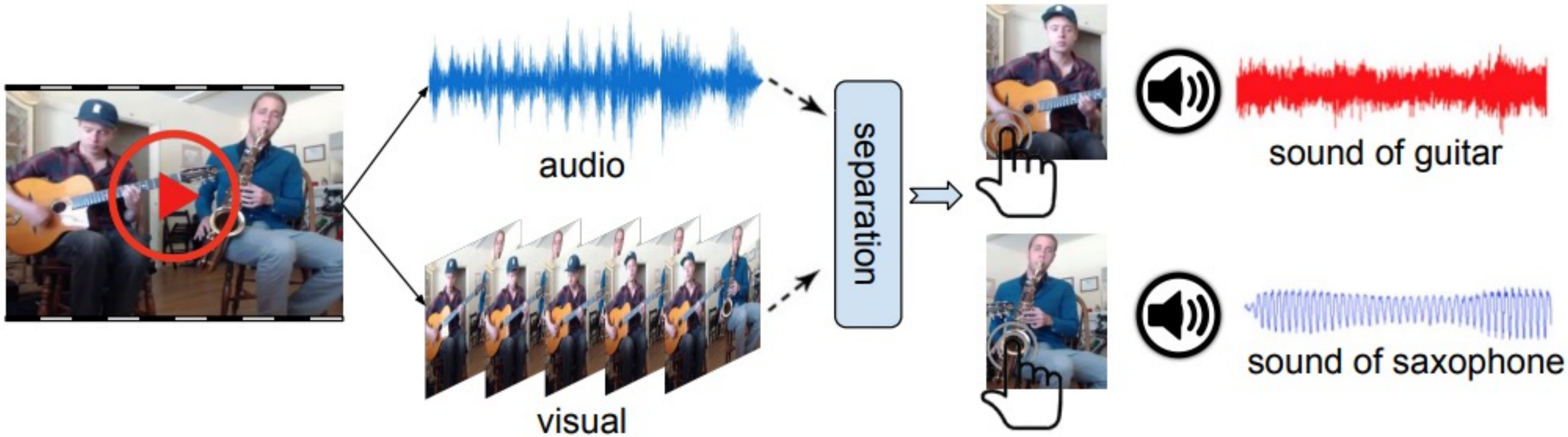
Presenter: Yan-Bo Lin
11-08-2021

Overview

- Introduction
- Motivation
- Proposed framework
- Dataset
- Results
- Conclusion
- Discussion

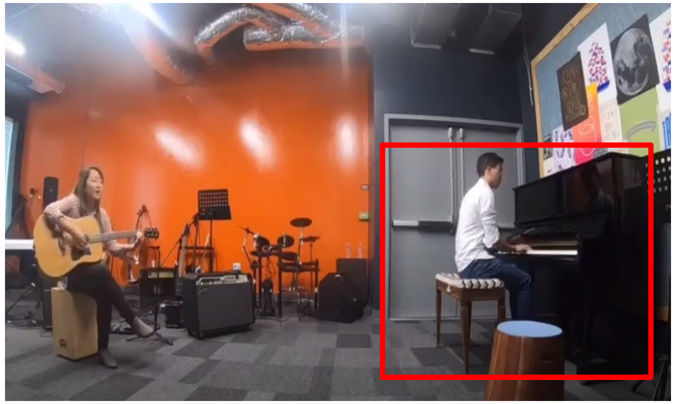
Introduction

- What is **Audio-visual source separation**?
  - Input: a video with audio track.
  - Output: separated sound corresponding to objects



Motivation

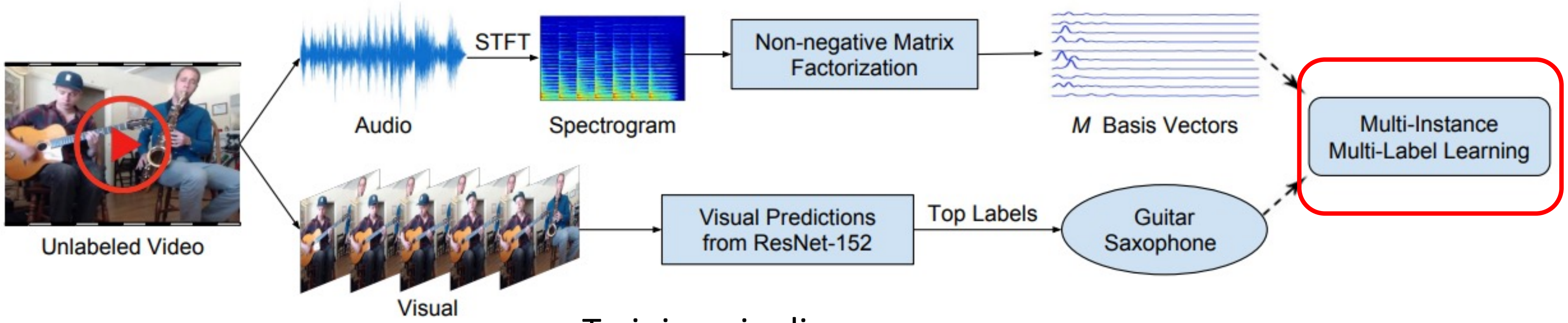
- Limitation of traditional works on audio source separation:
  - Traditional approaches aim to learn audio basis of object sound.
  - Audio source separation requires **clean** single audio source.
- Visual contents from unlabeled video can served as a supervisory signal for audio.



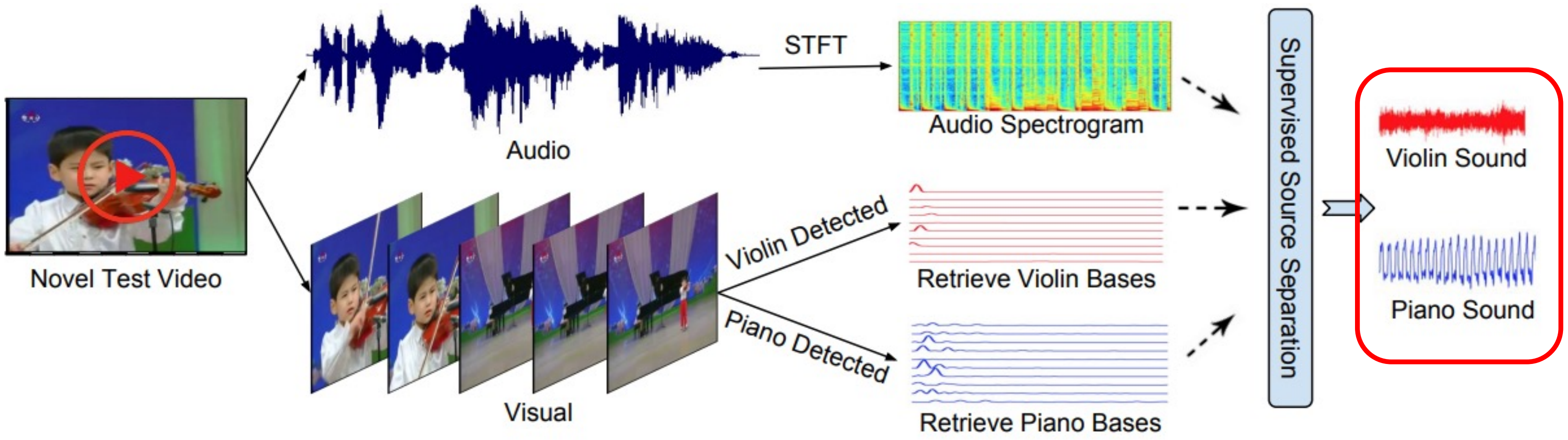
We may find several unlabeled videos containing piano sounds.

Proposed Method

- Noted that pipelines during training and inference time are different.



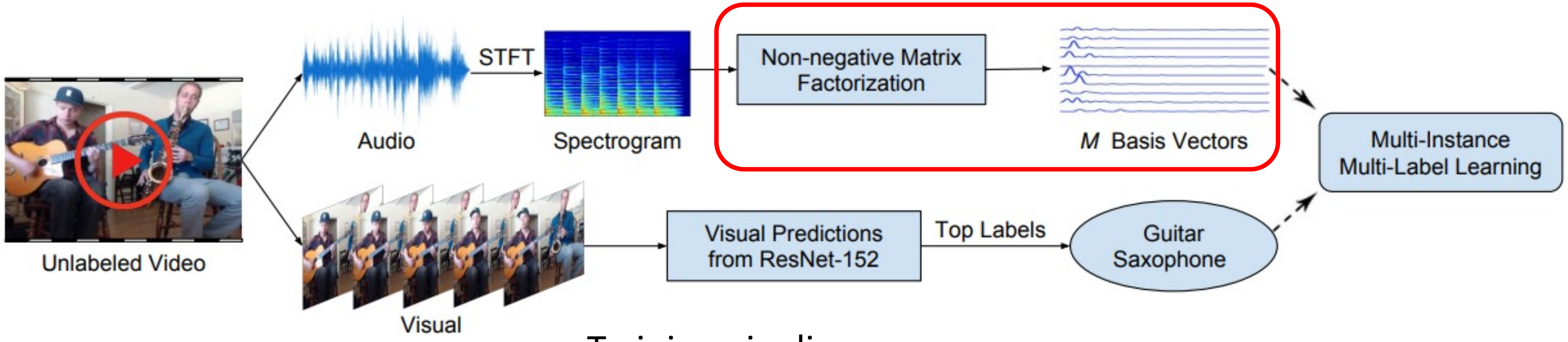
Training pipeline



Inference pipeline

Proposed Method

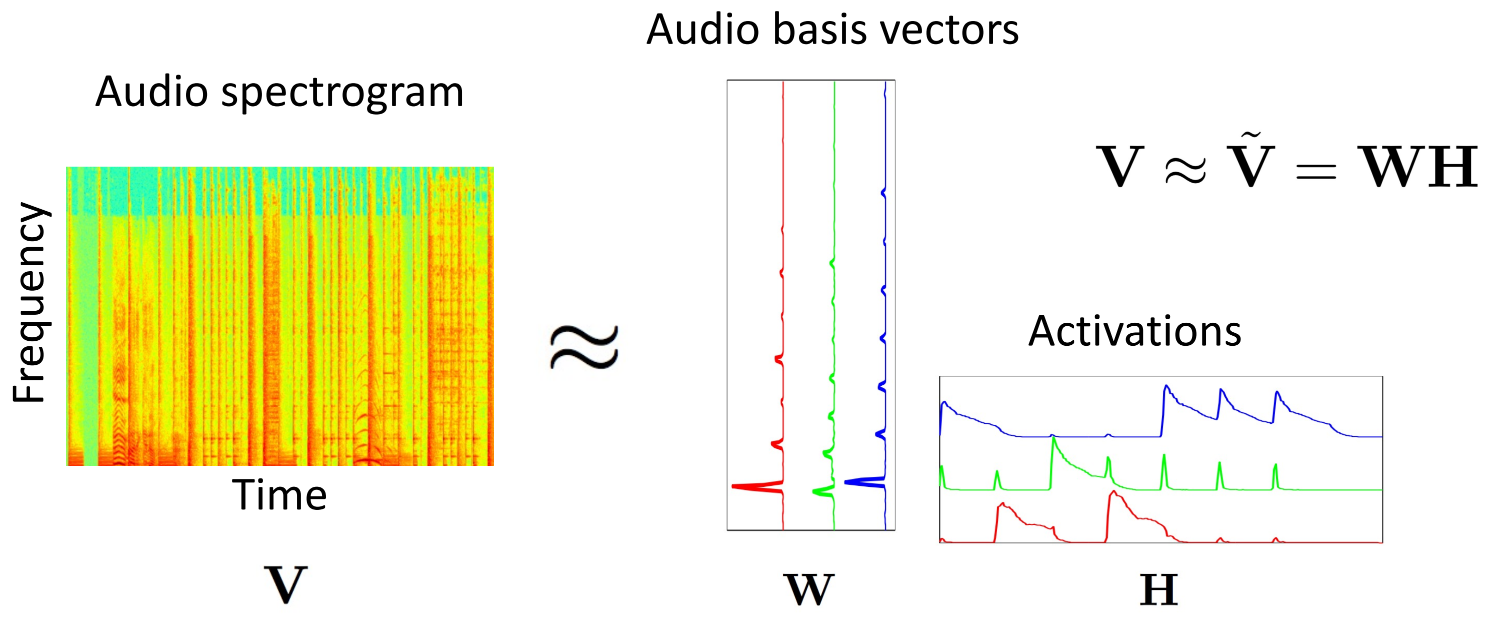
- Non-negative matrix factorization (NMF) aims to decompose audio spectrogram into basis and corresponding weights.



Training pipeline

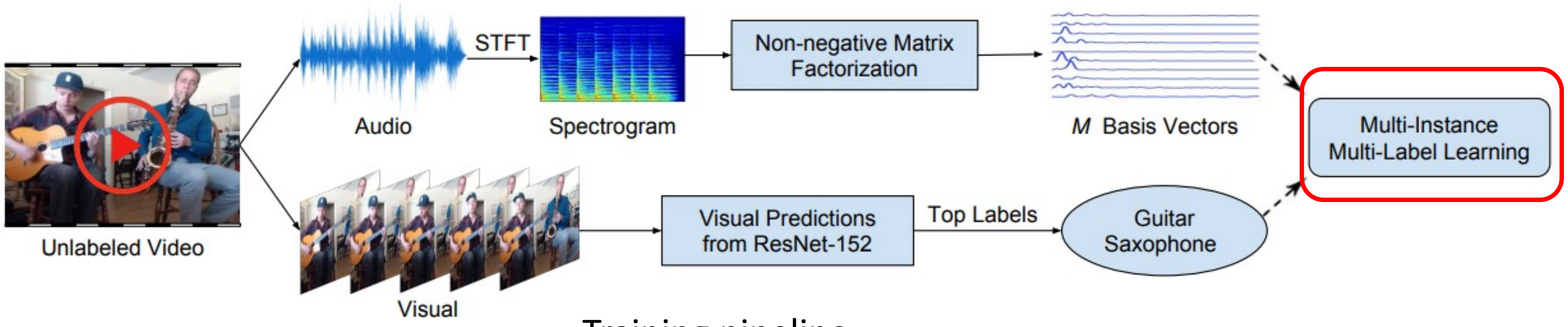
Proposed Method

- Non-negative matrix factorization (NMF) aims to decompose audio spectrogram into basis and corresponding weights.



Proposed Method

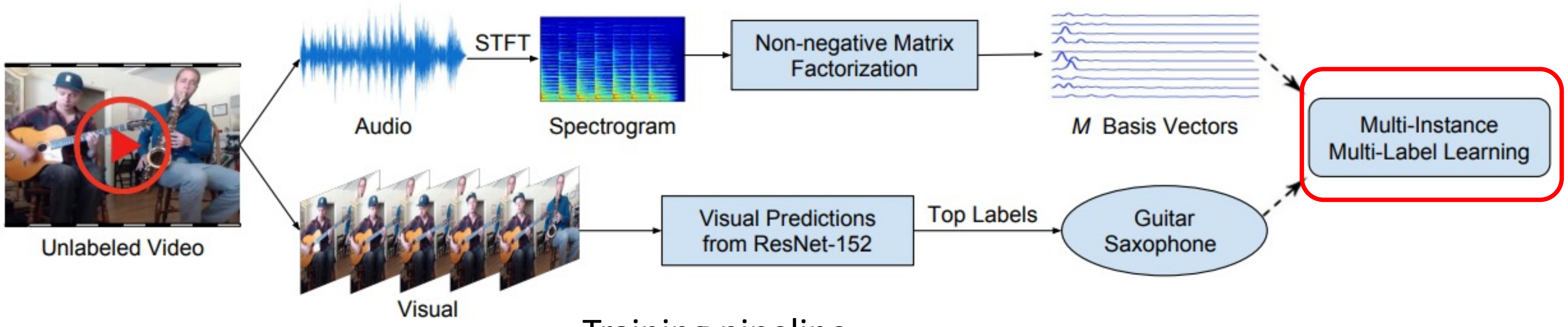
- After obtaining *M* (pre-defined) audio basis (taking W only), proposed method leverage multi-instance learning framework to associate audio-visual information.
- MIL framework can address noise labels from ResNet.



Training pipeline

Proposed Method

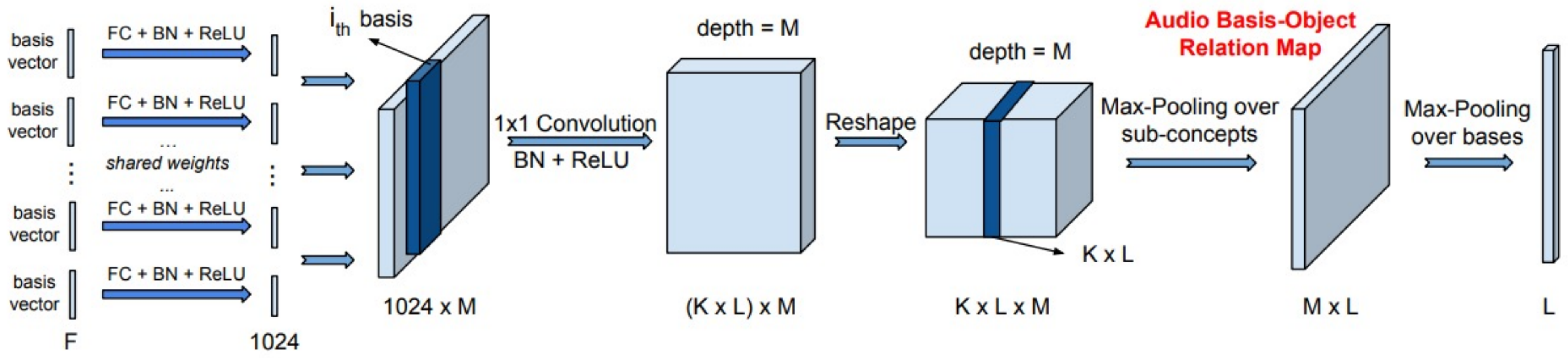
- MIL aims to associate information in **bag-level.**
  - For example, the visual prediction may contains guitar and saxophone. However, the video may contain guitar sound only.
  - In this setting, the positive bag is that at least one audio sound and a object are associated.



Training pipeline

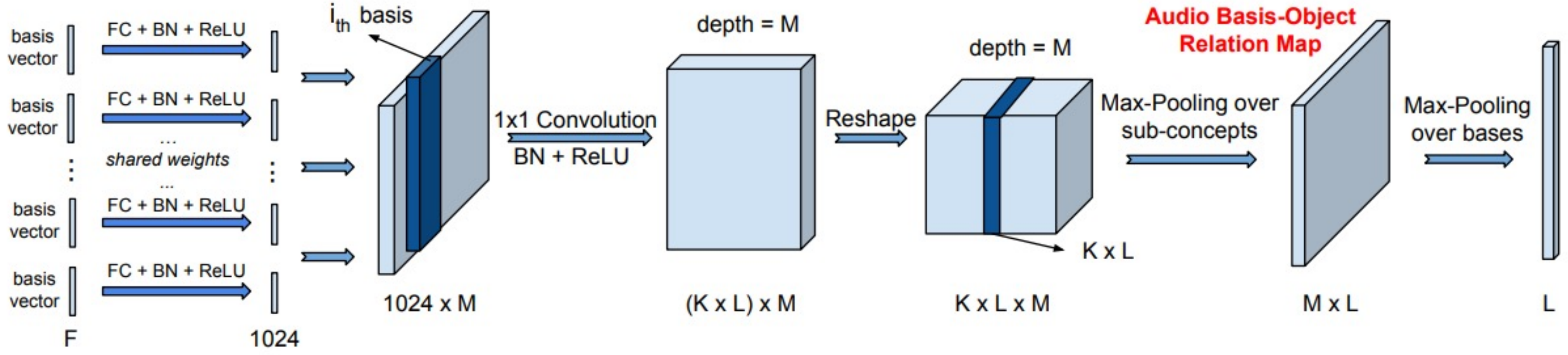
Proposed Method

- MIL aims to associate information in **bag-level.**
  - There are **M** basis vector with 1024-Dimension.
  - 1024-D features are decomposed into **K** sub-concepts with **L** object categories.
  - Max-pooling first apply over sub-concept and then over **M** basis.



Proposed Method

- MIL aims to associate information in **bag-level.**
  - The loss encourage scores of the correct classes larger than incorrect ones by a margin of 1.
  - The classes are predicted from ResNet.

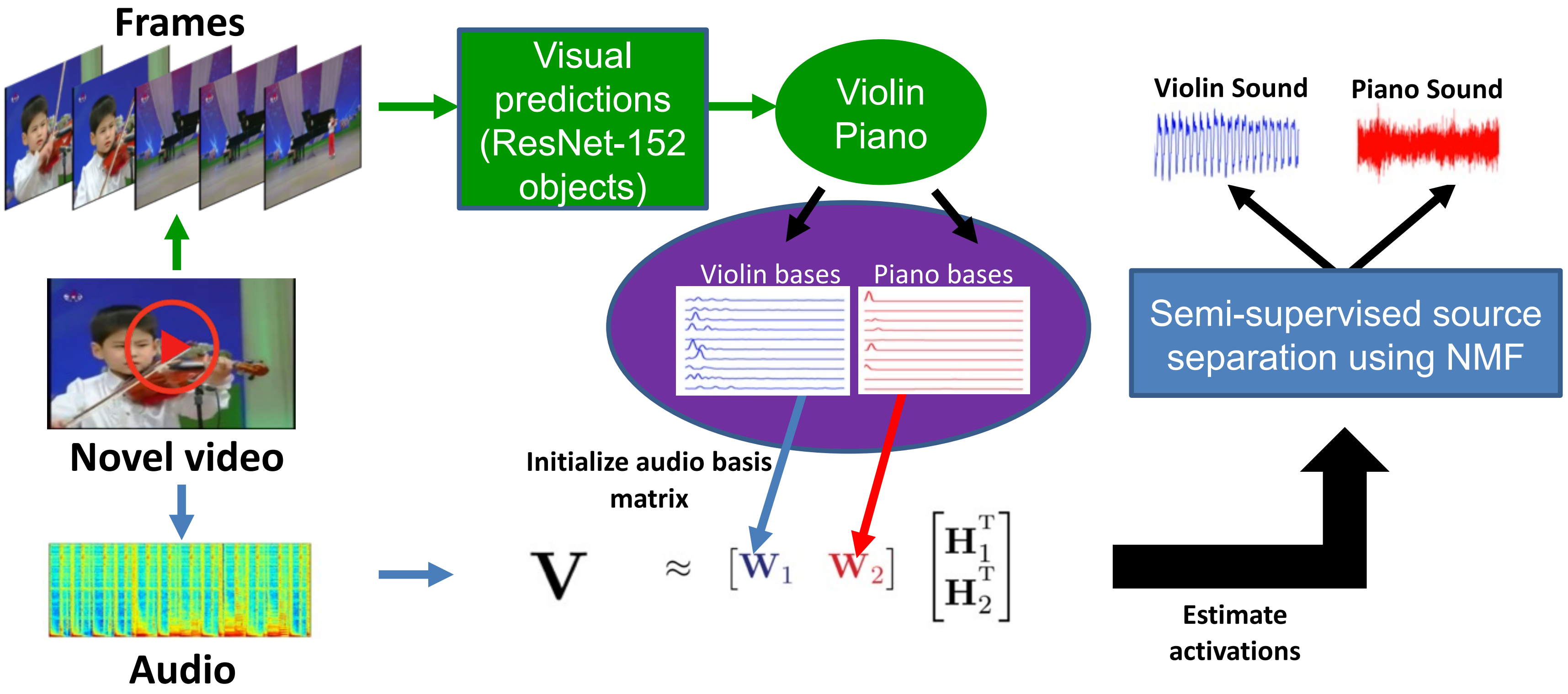


Apply multi-label hinge loss $$\mathcal{L}(A, \mathcal{V}) = \frac{1}{L} \sum_{i=1, i \neq \mathcal{V}_j}^{L} \sum_{j=1}^{|\mathcal{V}|} \max[0, 1 - (A_{\boxed{\mathcal{V}_j}} - A_{\boxed{i}})]$$

$A \in \mathbb{R}^L$

Proposed Method (Inference)

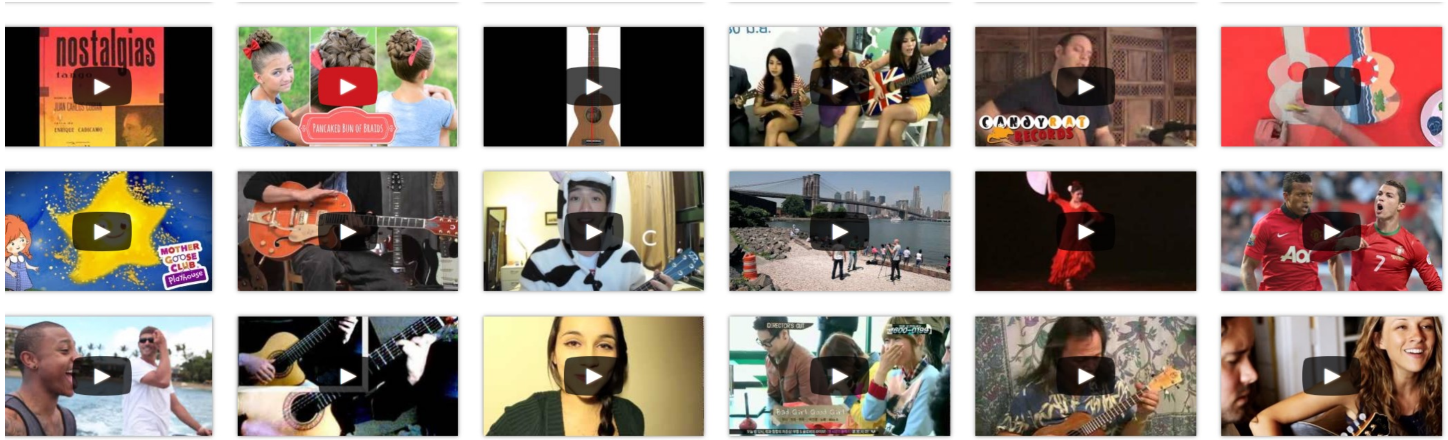
- Given a video, the proposed model leverages learned **W** and **H** to separate sounds.
  - Specifically, W is fixed and applied for all videos. H is estimated from given a video.



Examples adapted from Ruohan's slide

Experiment

- Dataset:
  - AudioSet-Unlabeled is adapted from audioset with filtering pre-defined labels. ~100k videos
  - AudioSet-SingleSource is for evaluation. All videos are single source video. ~23 videos.
  - AV-Bench is toy example with 3 videos (Violin Yanni, Wooden Horse, and Guitar).



Example of audioset

Experiment

- Results and metrics:
  - Given a mixed source from two single sources, the model aims to separate these **two** sources.
  - The results are reported in **SDR**. Higher is better.

Use the GT-labels to find audio basis

	Instrument Pair	Animal Pair	Vehicle Pair	Cross-Domain Pair
Upper-Bound	2.05	0.35	0.60	2.79
K-means Clustering	-2.85	-3.76	-2.71	-3.32
MFCC Unsupervised [72]	0.47	-0.21	-0.05	1.49
Visual Exemplar	-2.41	-4.75	-2.21	-2.28
Unmatched Bases	-2.12	-2.46	-1.99	-1.93
Gaussian Bases	-8.74	-9.12	-7.39	-8.21
Ours	**1.83**	**0.23**	**0.49**	**2.53**

Experiment

- Results and metrics:
  - Given a mixed source from two single sources, the model aims to separate these **two** sources.
  - The results are reported in **SDR**. Higher is better.

	Instrument Pair	Animal Pair	Vehicle Pair	Cross-Domain Pair
Upper-Bound	2.05	0.35	0.60	2.79
K-means Clustering	-2.85	-3.76	-2.71	-3.32
MFCC Unsupervised [72]	0.47	-0.21	-0.05	1.49
Visual Exemplar	-2.41	-4.75	-2.21	-2.28
Unmatched Bases	-2.12	-2.46	-1.99	-1.93
Gaussian Bases	-8.74	-9.12	-7.39	-8.21
Ours	**1.83**	**0.23**	**0.49**	**2.53**

Use the sound from other videos to guide NMF (e.g., two video contains guitars.)

Experiment

- Results on audio-visual denoising on AV-Bench in Normalized SDR.

	Wooden Horse	Violin Yanni	Guitar Solo	Average
Sparse CCA (Kidron et al. [47])	4.36	5.30	5.71	5.12
JIVE (Lock et al. [55])	4.54	4.43	2.64	3.87
Audio-Visual (Pu et al. [62])	8.82	5.90	**14.1**	9.61
Ours	**12.3**	**7.88**	11.4	**10.5**

Experiment

- Demo video.
  - Train on 100,000 unlabeled multi-source video clips, then separate audio for novel video

Conclusion

- This paper leverages unlabeled videos to perform source separation.
- MIL learning can effectively associate audio and visual information in such noise videos.

Discussion

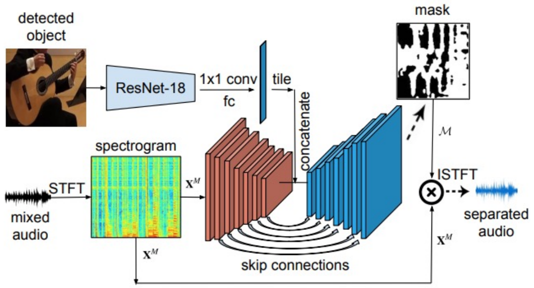
- Is NMF a good way to separate sounds?

- Does proposed method truly leverage unlabeled video?

- Limitation from object labels.

Discussion

- Is NMF a good way to separate sounds?



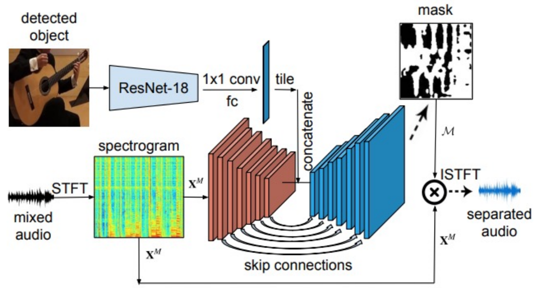
- Does proposed method truly leverage unlabeled video?

- Limitation from object labels.

Gao R, Grauman K. Co-separating sounds of visual objects. In ICCV 2019.

Discussion

- Is NMF a good way to separate sounds?

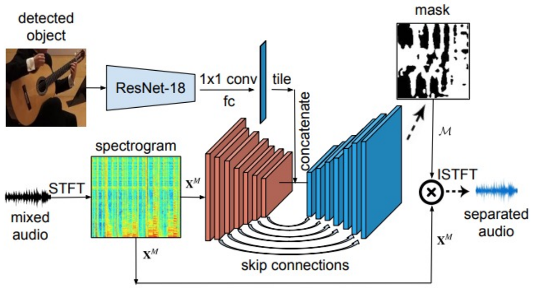


- Does proposed method truly leverage unlabeled video?
  - It is based on some assumptions: objects present in video; some videos are filtered.
- Limitation from object labels.

Gao R, Grauman K. Co-separating sounds of visual objects. In ICCV 2019..

Discussion

- Is NMF a good way to separate sounds?



- Does proposed method truly leverage unlabeled video?
  - It is based on some assumptions: objects present in video; some videos are filtered.
- Limitation from object labels.

Gao R, Grauman K. Co-separating sounds of visual objects. In ICCV 2019.
Zhao H, Gan C, Ma WC, Torralba A. The sound of motions. In ICCV 2019.