Audio-Visual Learning

Chuang Gan

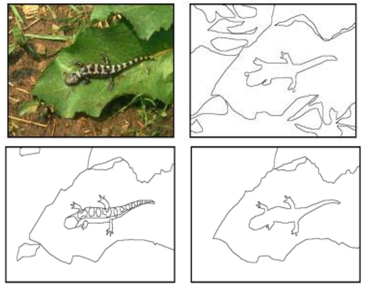
UNIVERSITY OF MASSACHUSETTS

1863

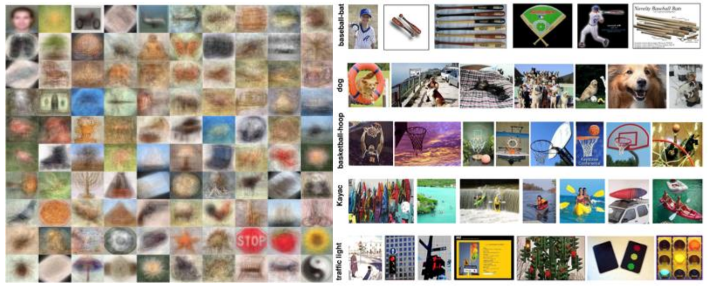
Welcome Visitors!

The Mcgurk Effect

Learning from images and video frames

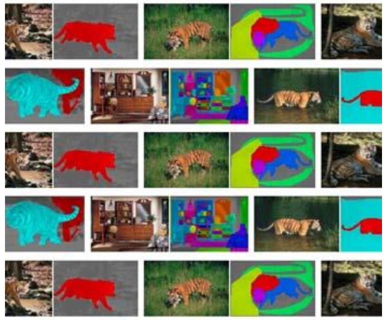


BSD (2001)



Caltech 101 (2004), Caltech 256 (2006)

PASCAL (2007-12)

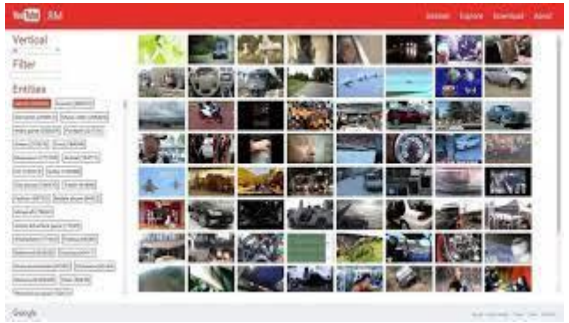


LabelMe (2007)

ImageNet (2009)

SUN (2010)

UCF-101 (2012)



Youtube-8M (2017)

Kinetics (2017)

What can sound give us?

physical interactions

speech

sound of distant object

Can machines connect sight with sound for rich perception?

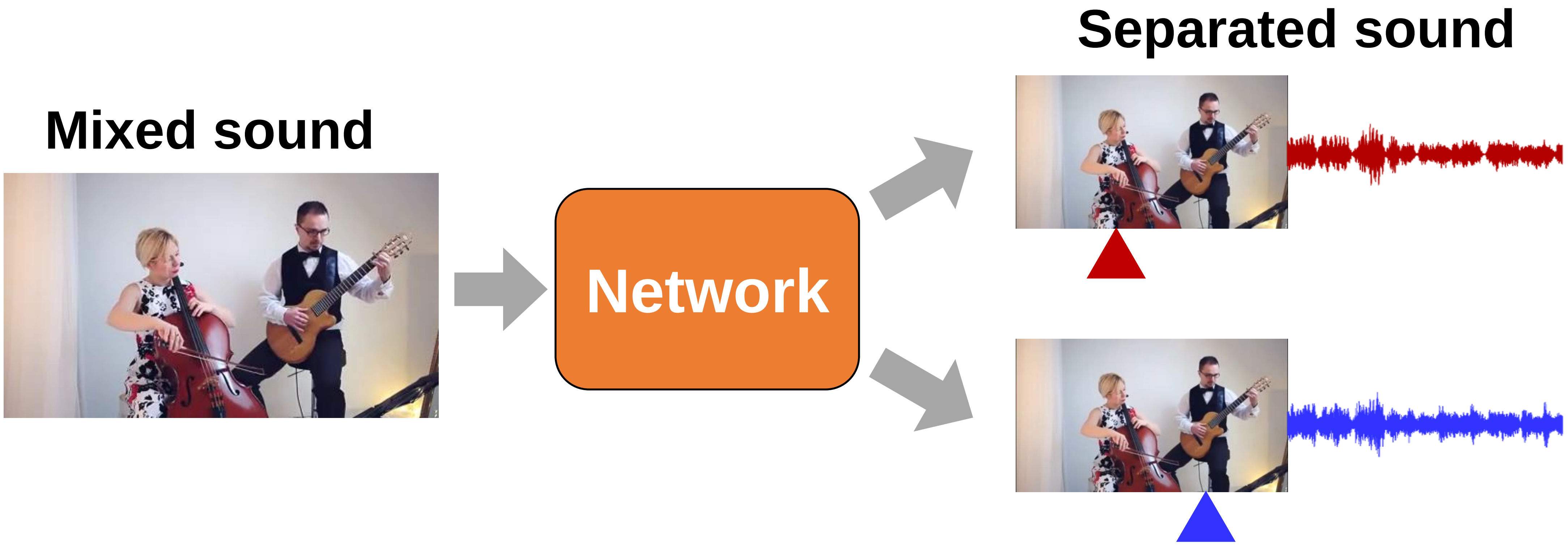
Task: Visual Sound Separation

Given a music performance video…

Mixed sound

Task: Visual Sound Separation

…we aim to separate two sounds played by different instruments.



Mixed sound

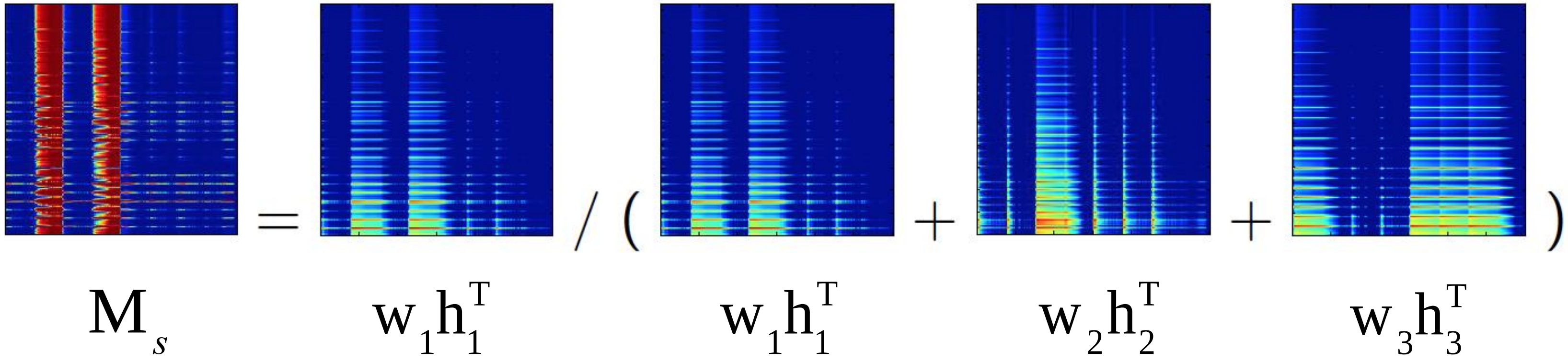
Network

Separated sound

Source Separation: Traditional Approach

- To separate one component out of K:
  - 1. Estimate the mask for the target component

$$\mathbf{M}_s = \frac{\mathbf{w}_1\,\mathbf{h}_1^{\mathrm{T}}}{\sum_{i=1}^{K} \mathbf{w}_i\,\mathbf{h}_i^{\mathrm{T}}}$$

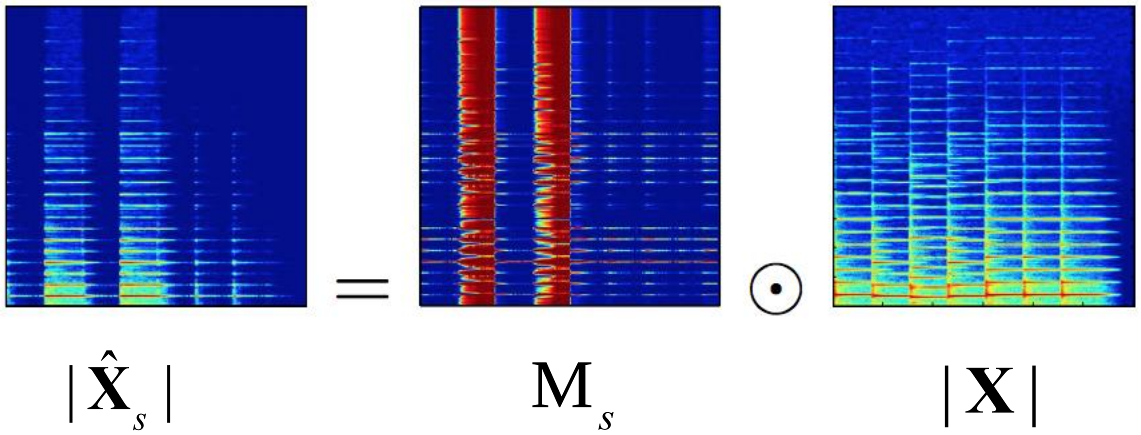


Source Separation: Traditional Approach

❑To separate one component out of K:

- 2. Masking on the input spectrogram to separate the component

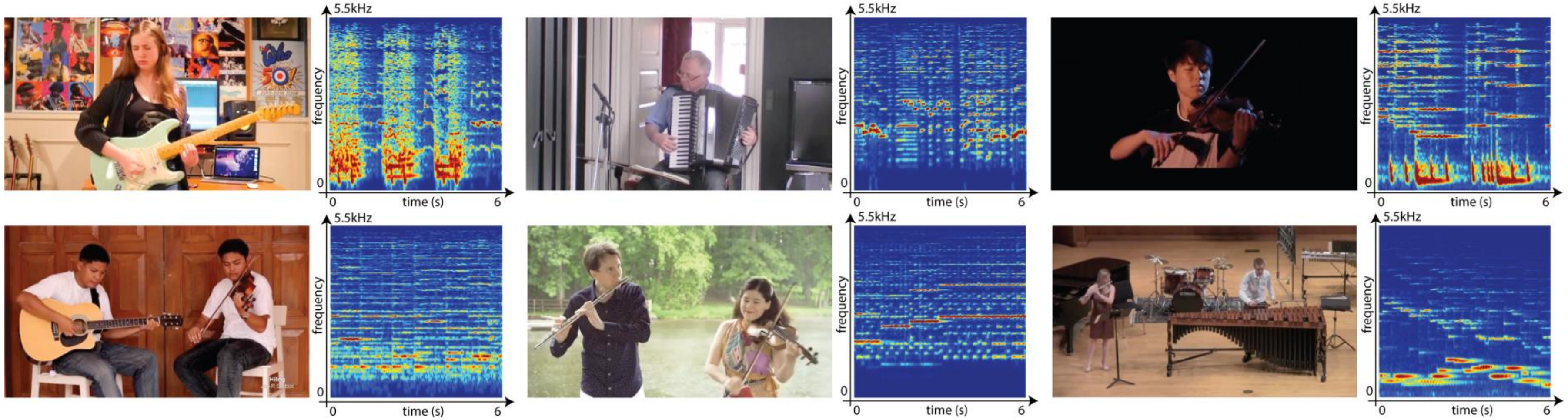
$$|\hat{\mathbf{X}}_s| = \mathbf{M}_s \odot |\mathbf{X}|$$



Our Ideas: Learning from Music Videos

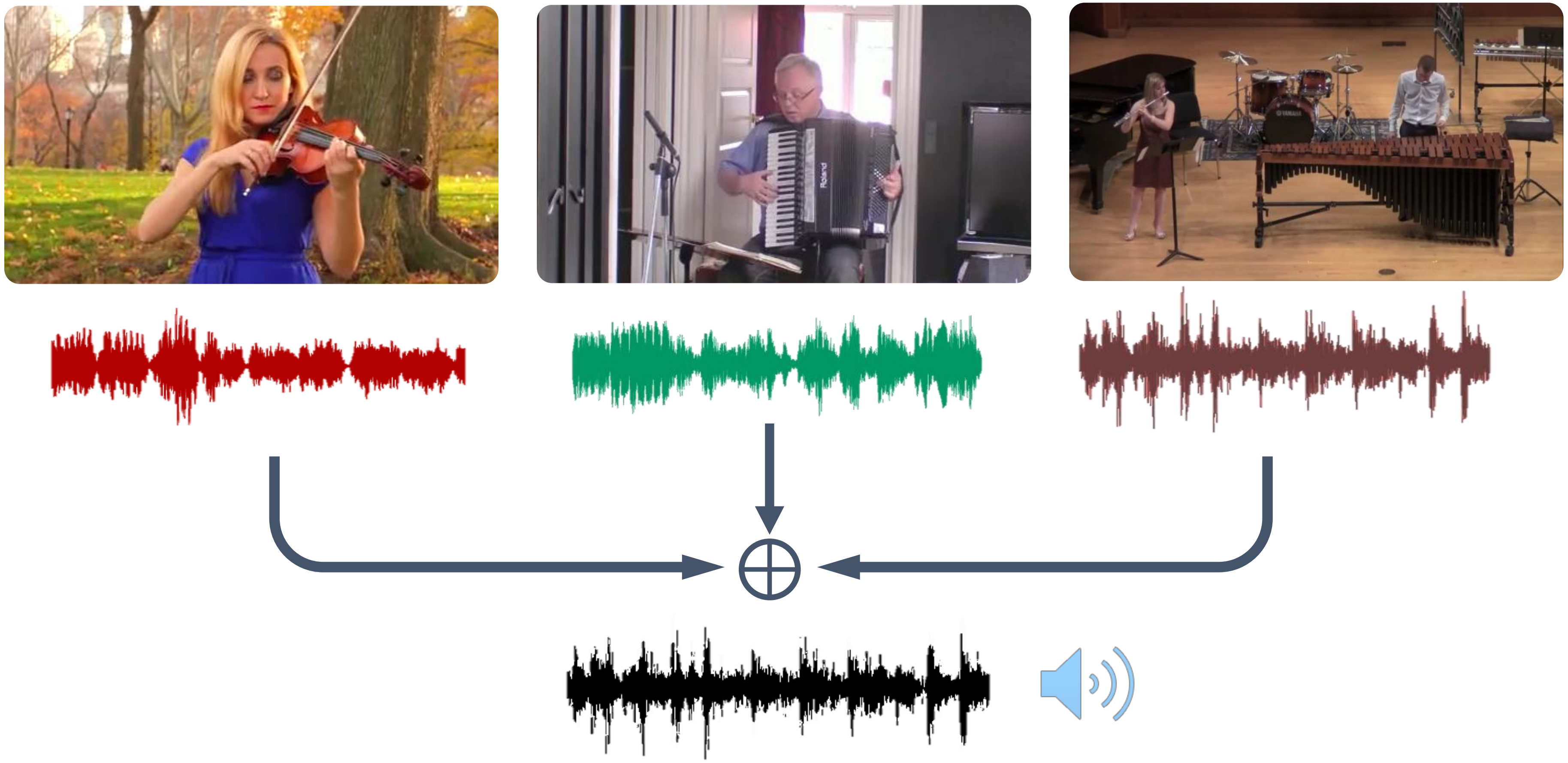
□ Internet music videos

- Keyword search without labeling
- >20 kinds of commonly seen musical instruments
- >1000 solos and duets



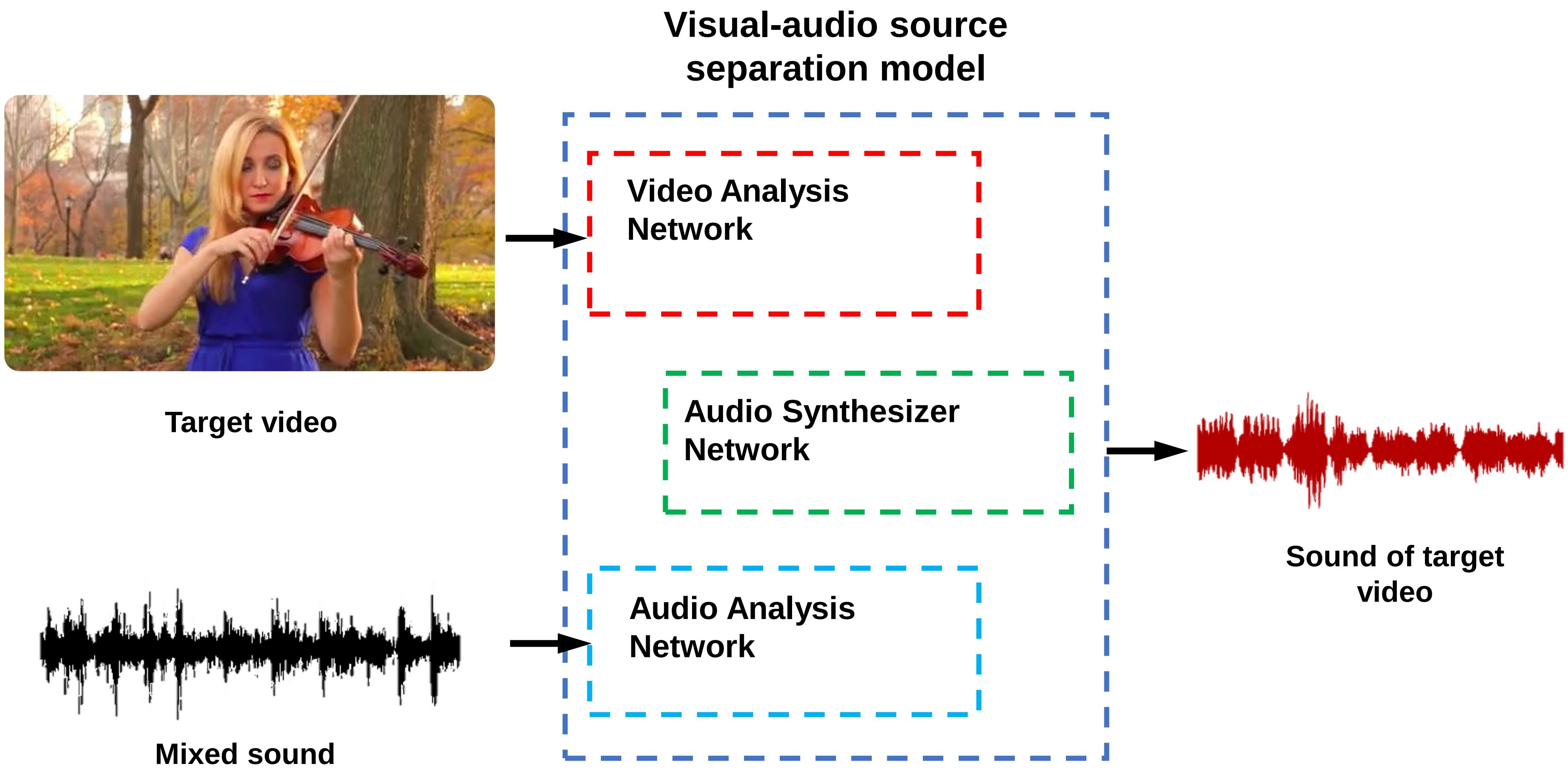
Zhao, Gan et al. “The Sound of Pixels.” ECCV 2020.

Mixing the Sound



Zhao, Gan et al. “The Sound of Pixels.” ECCV 2020.

Vision to Rescue for Self-supervised Prediction



Visual-audio source separation model

Video Analysis Network

Audio Synthesizer Network

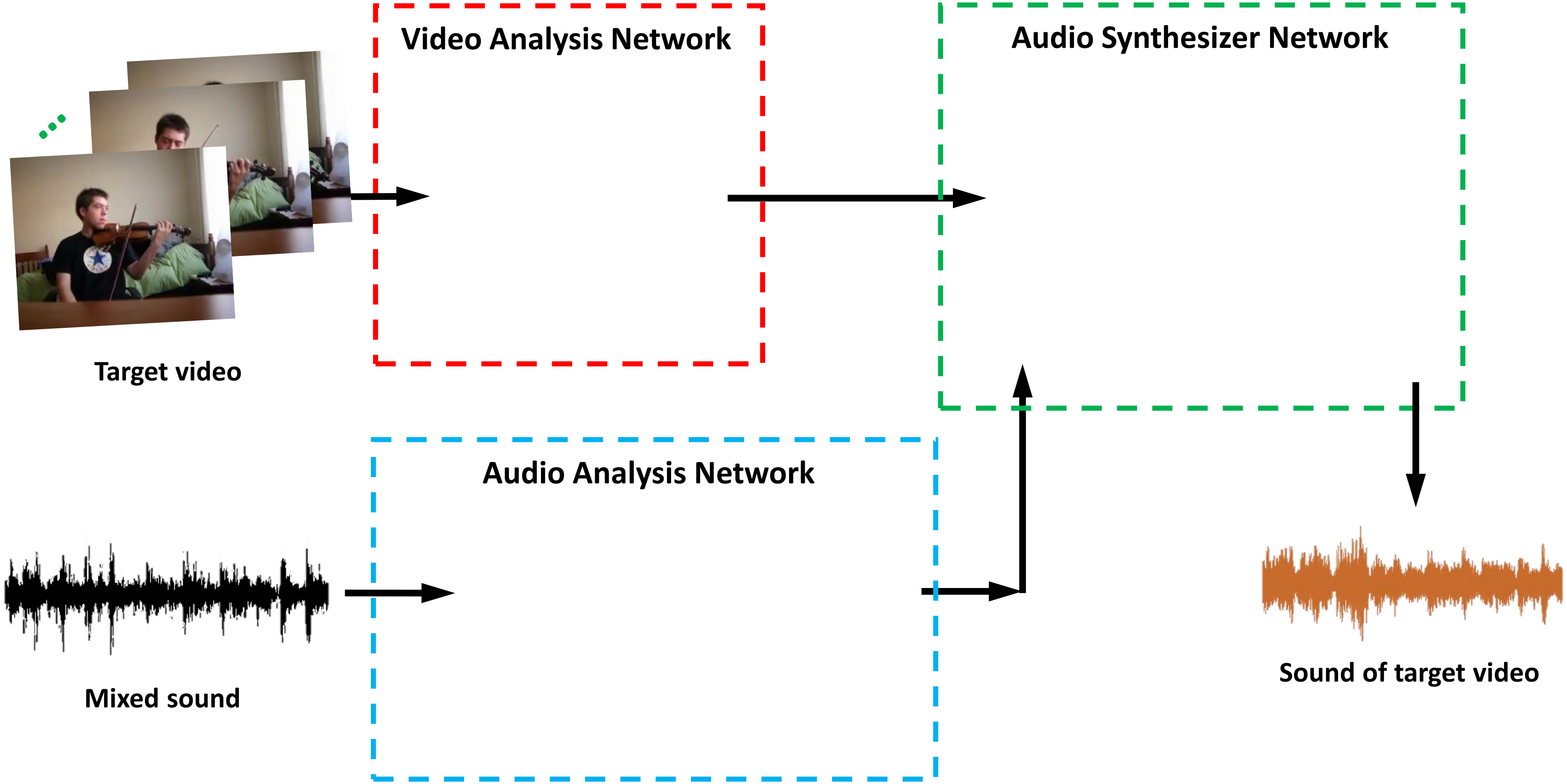
Audio Analysis Network

Target video

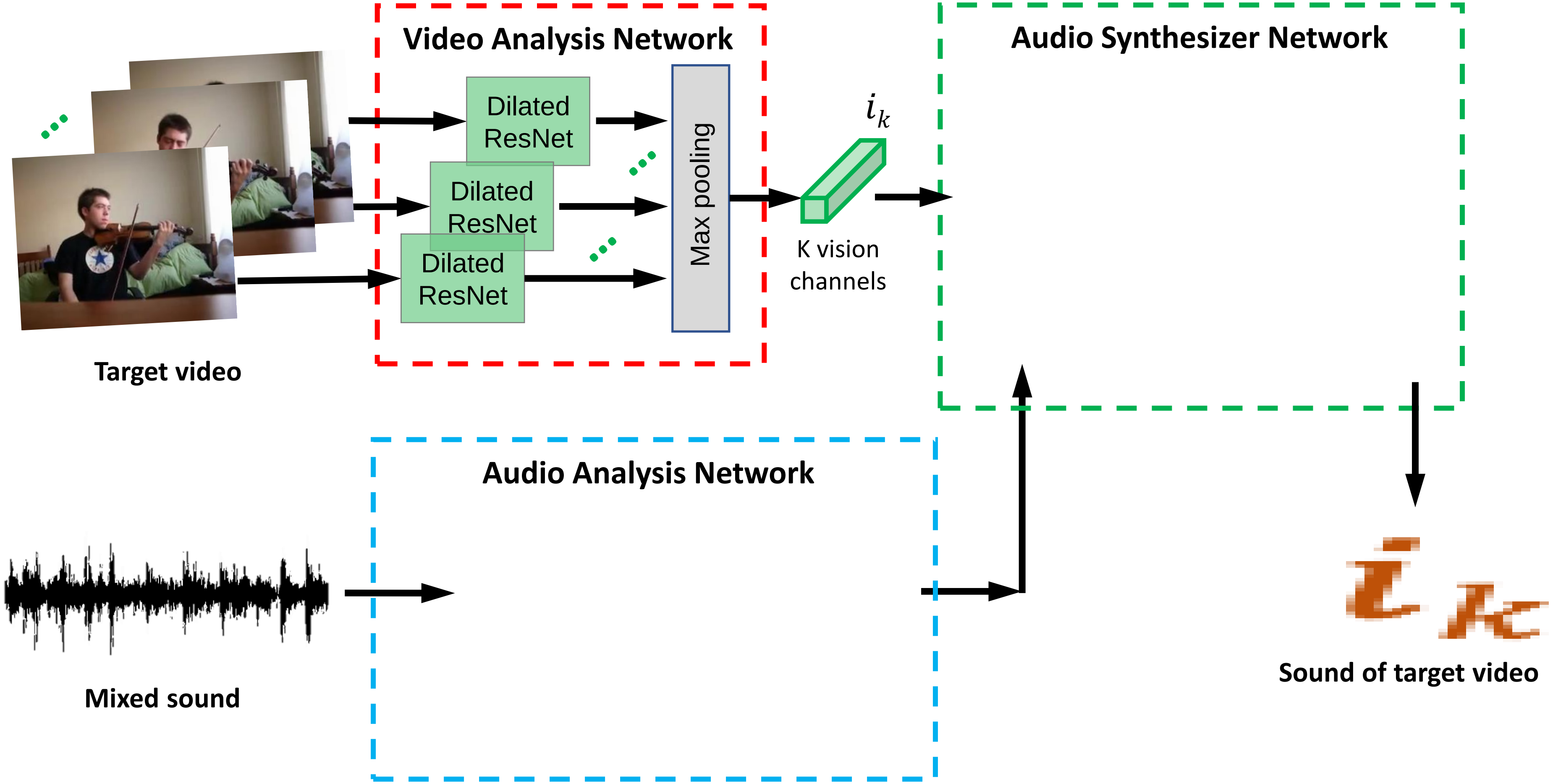
Mixed sound

Sound of target video

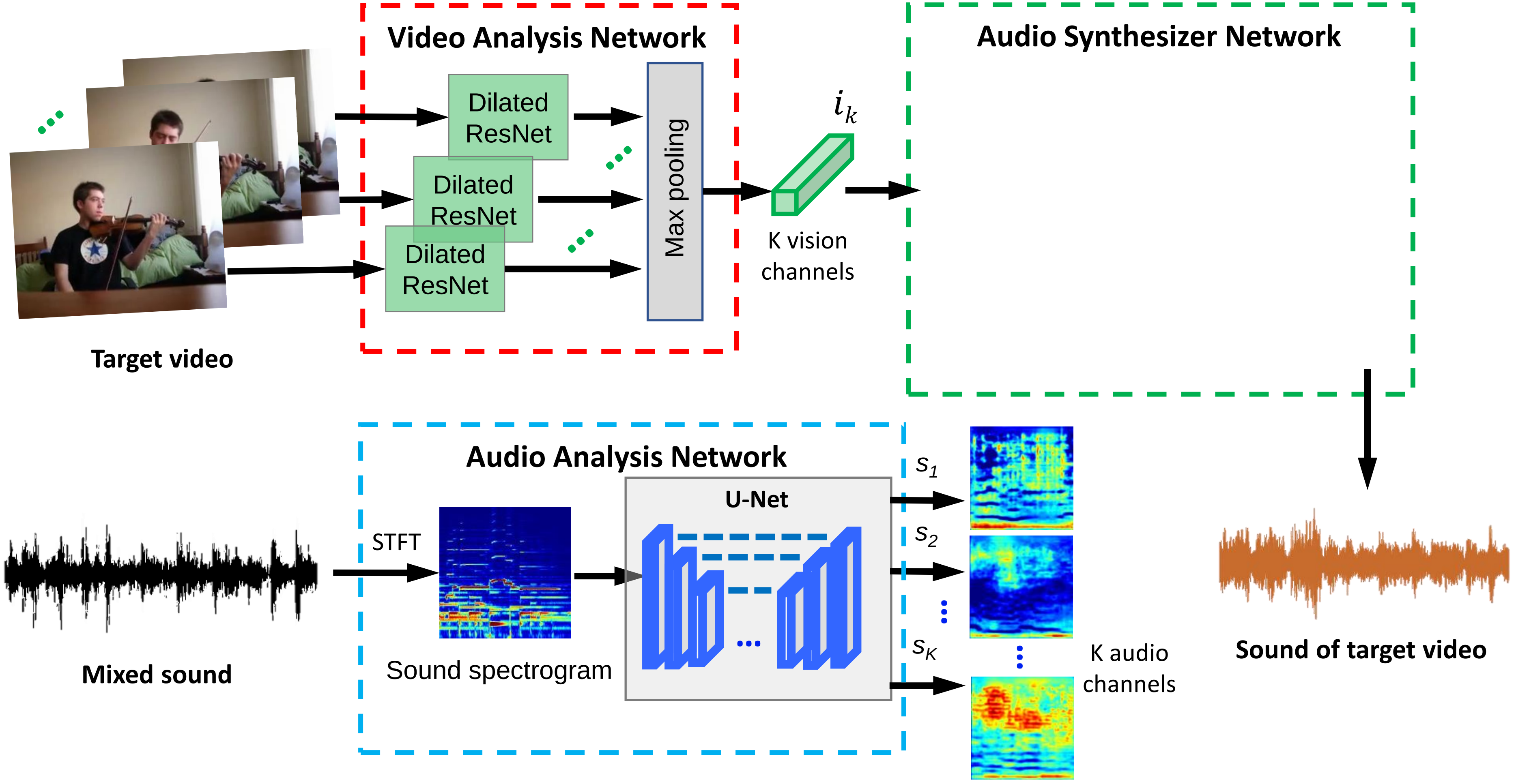
Mix-and-Separate Framework



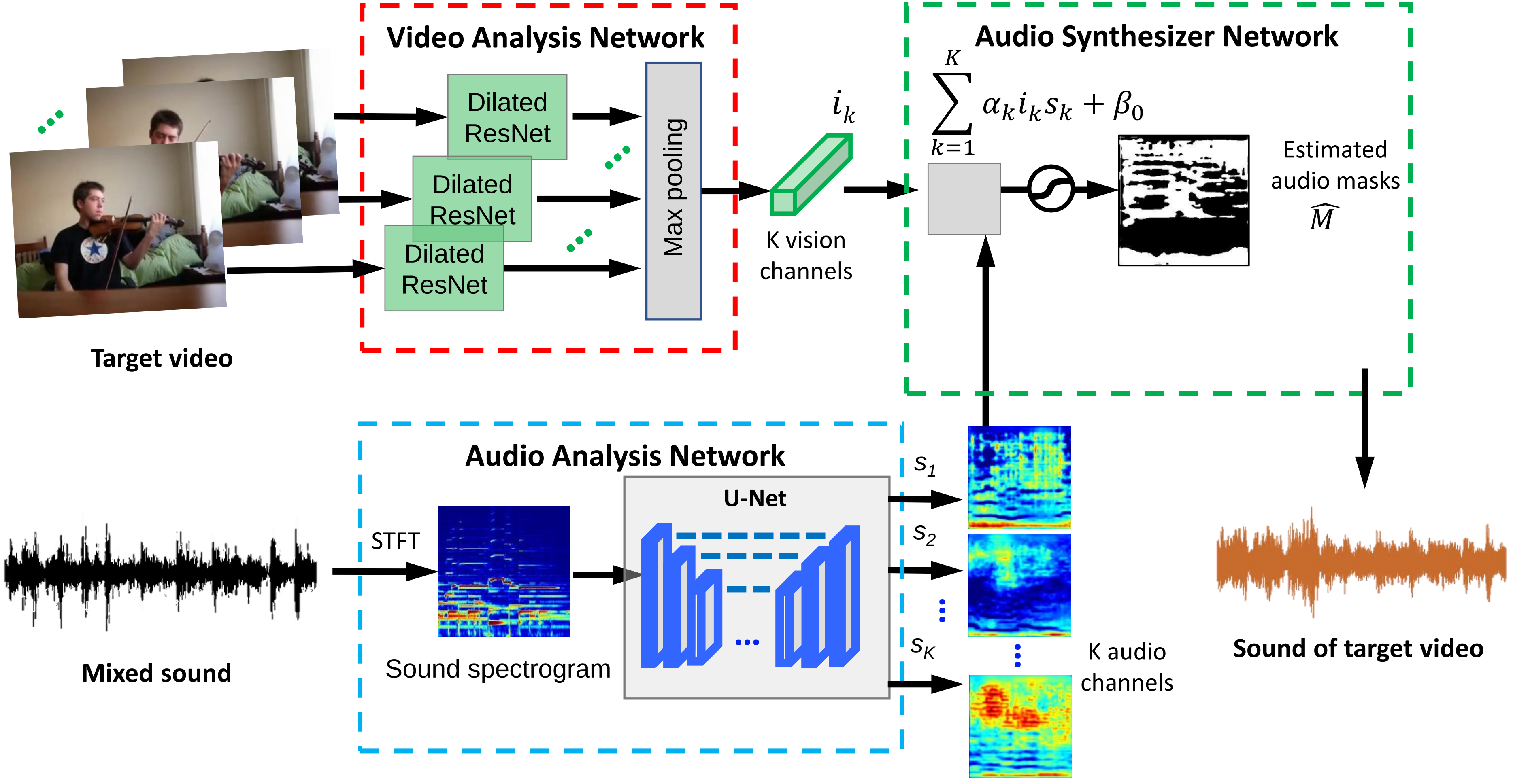
Mix-and-Separate Framework



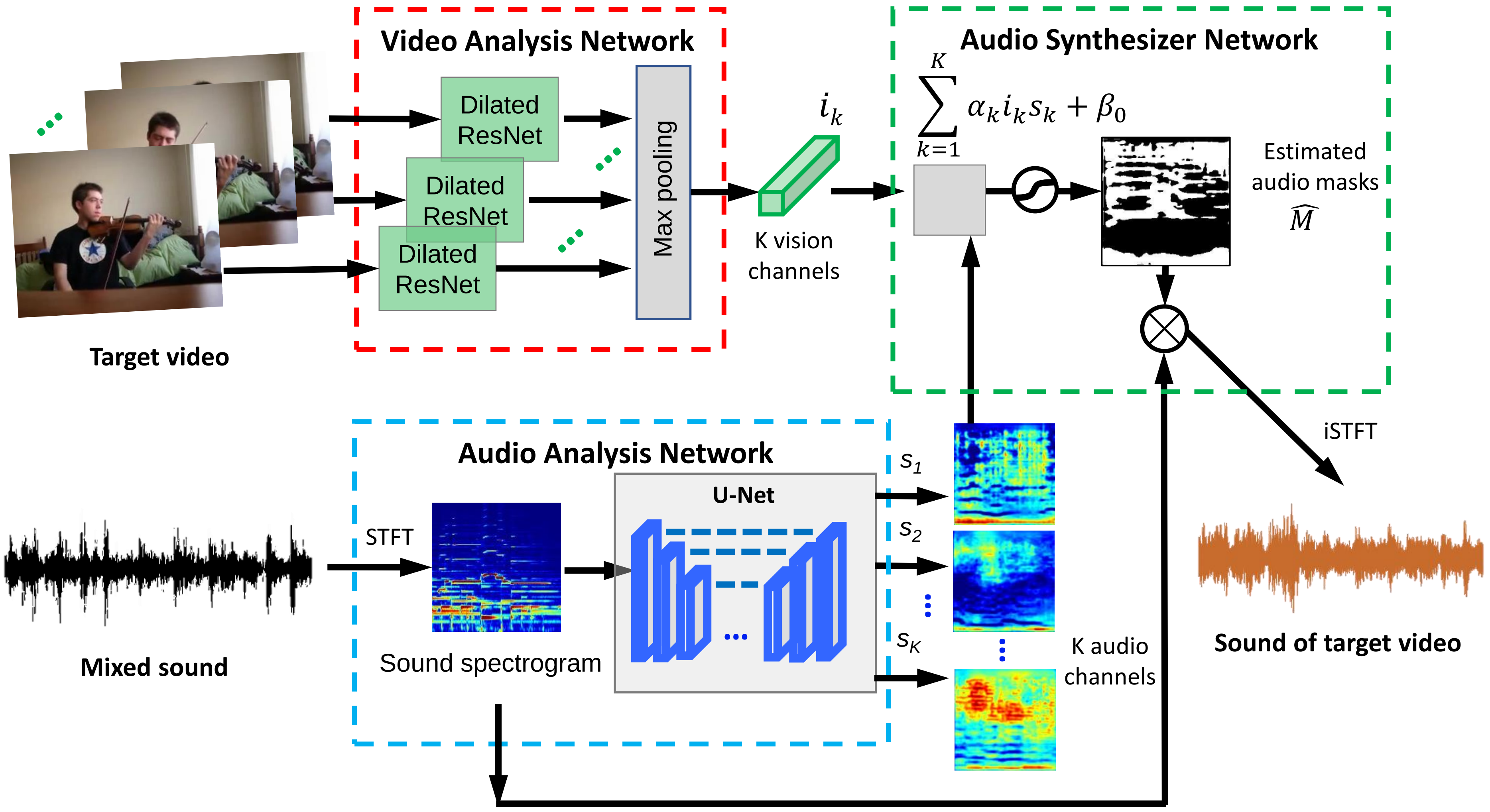
Mix-and-Separate Framework



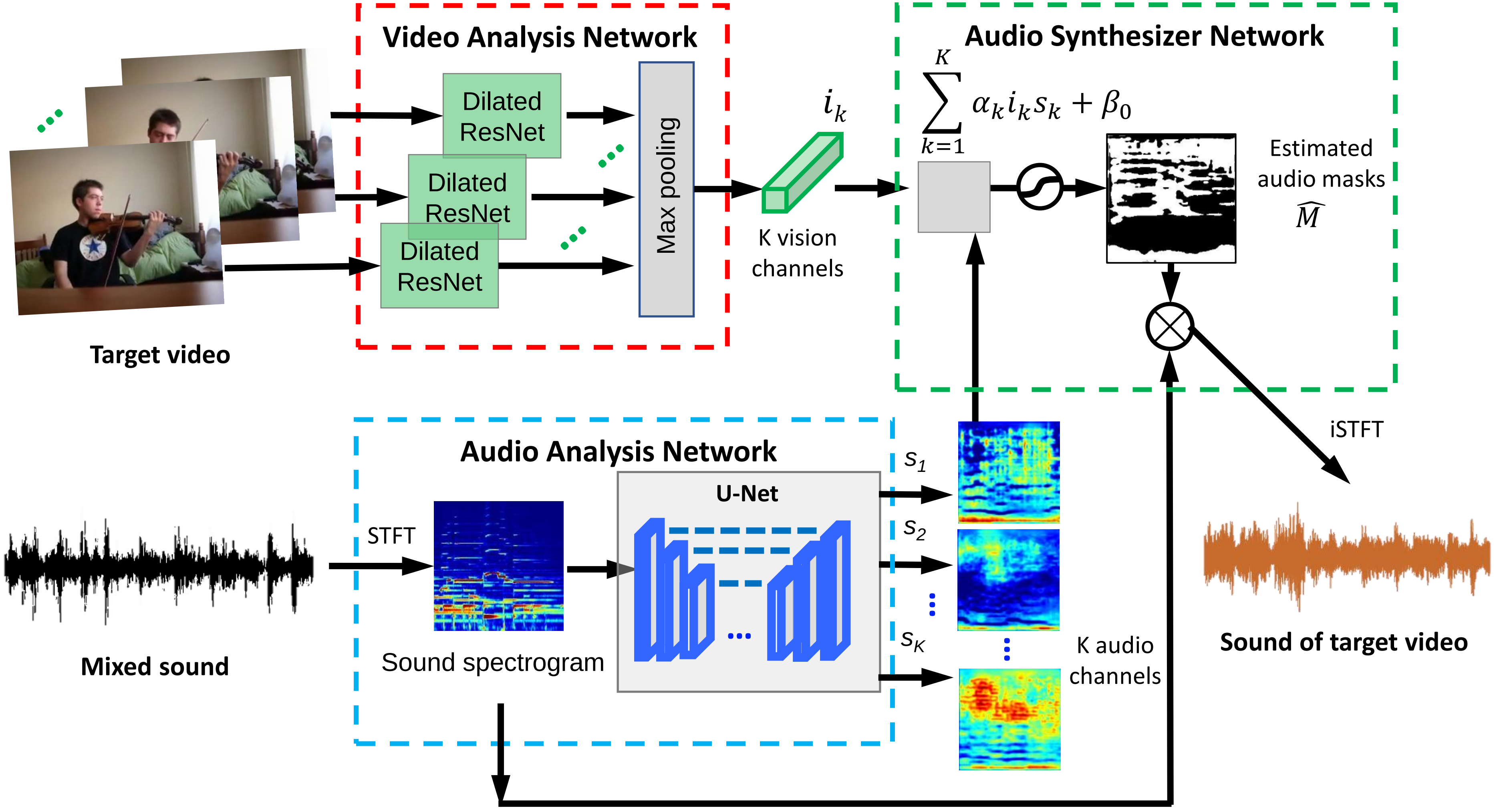
Mix-and-Separate Framework



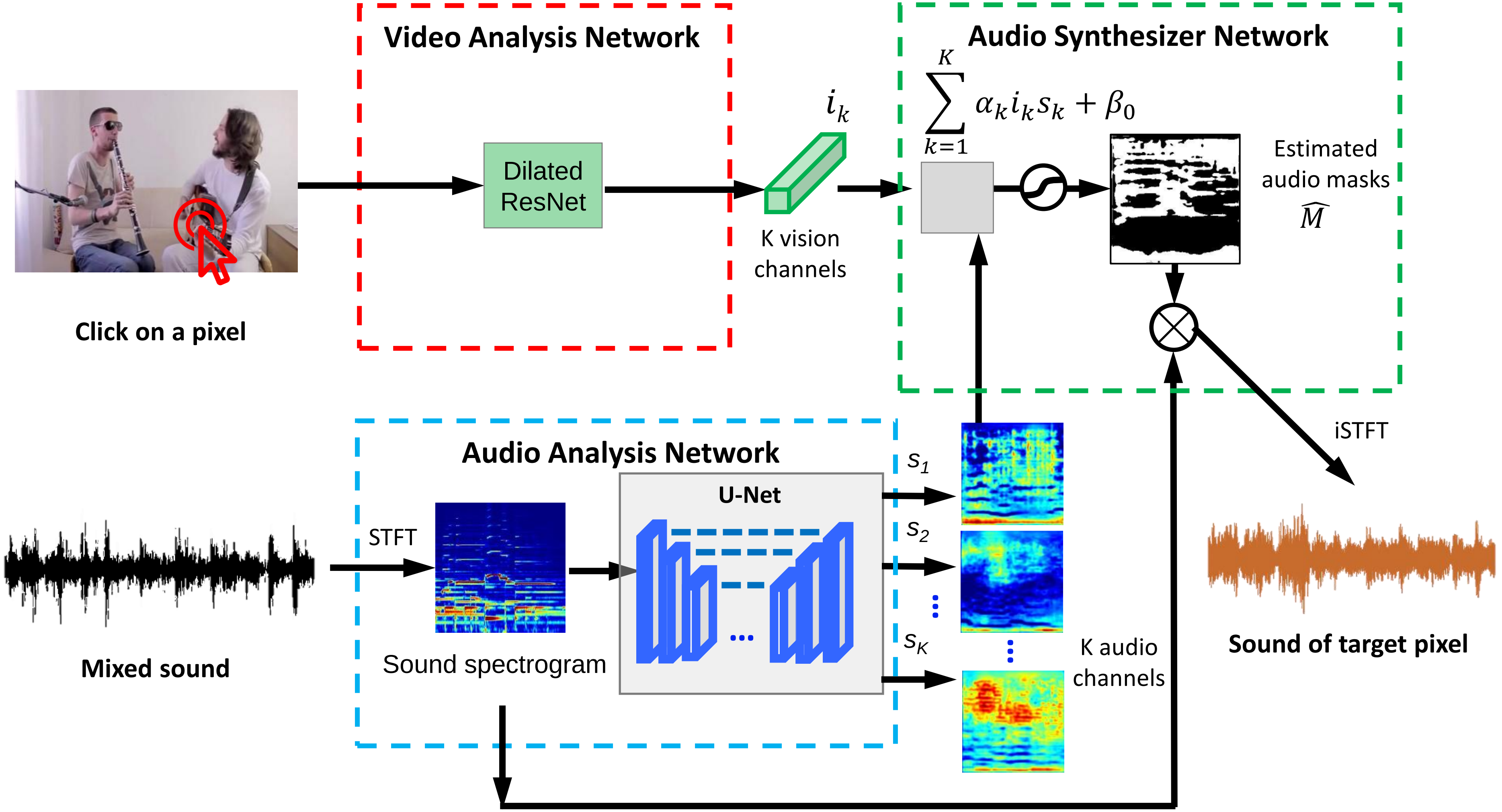
Mix-and-Separate Framework



Test Time


Video Analysis Network
Dilated ResNet
Dilated ResNet
Dilated ResNet
Max pooling
Target video
i_k
K vision channels
Audio Synthesizer Network
$\sum_{k=1}^{K} \alpha_k i_k s_k + \beta_0$
Estimated audio masks
$\widehat{M}$
iSTFT
Audio Analysis Network
STFT
U-Net
Sound spectrogram
Mixed sound
s_1
s_2
s_K
K audio channels
Sound of target video

Test Time: using Pixel Feature instead



Original Video

The sound of clicked object…

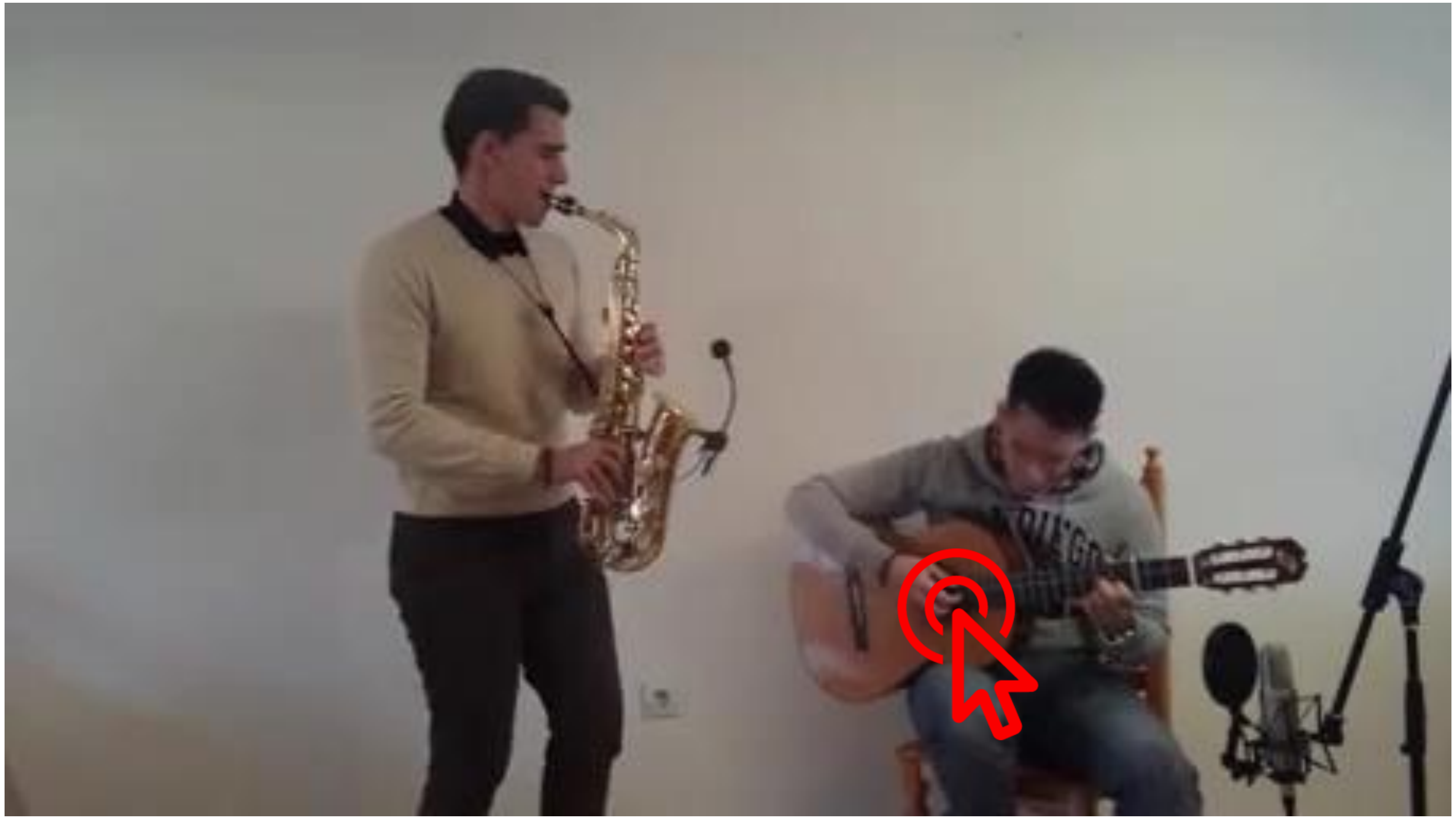
The sound of clicked object…

The sound of clicked object…

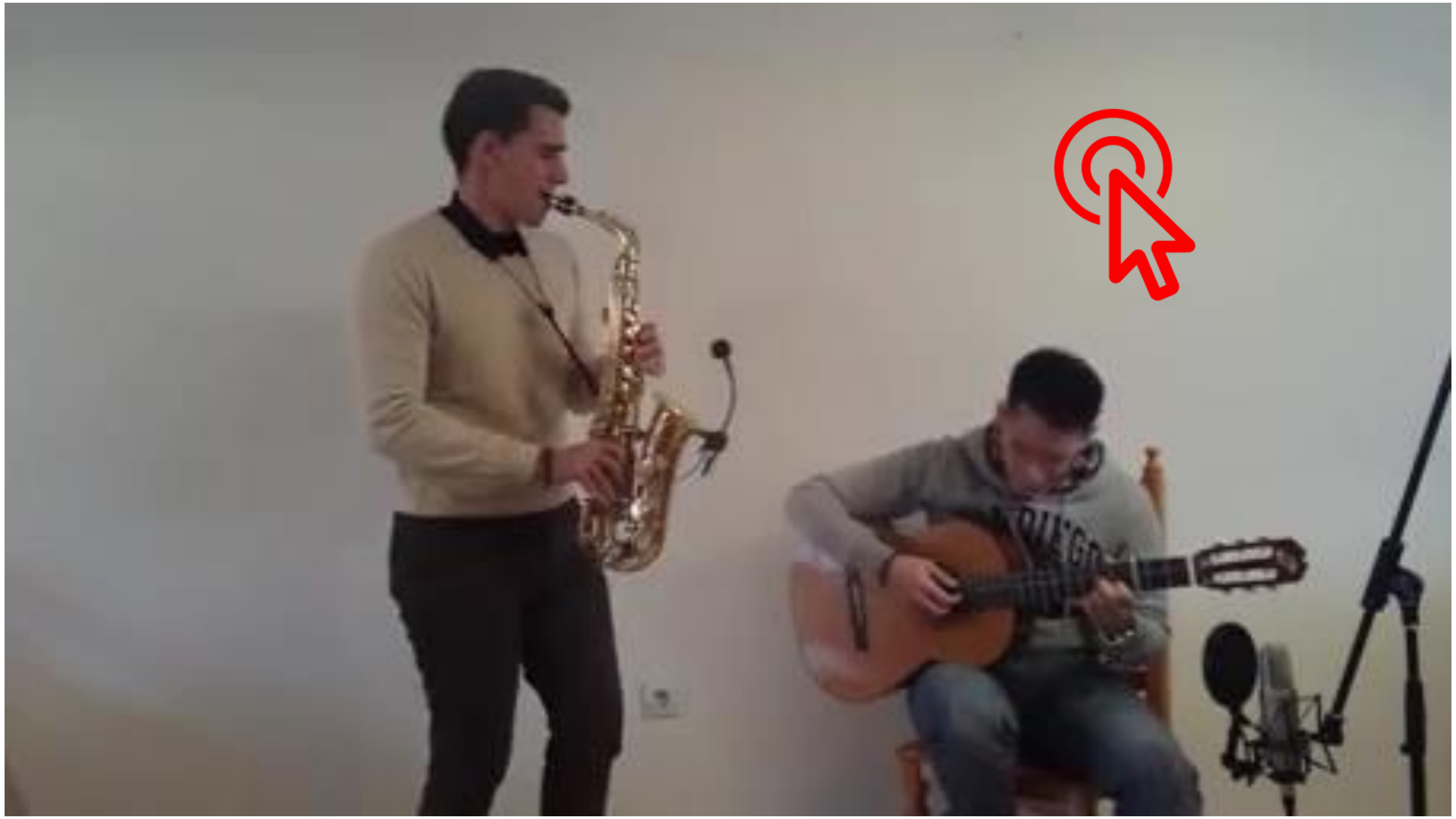
Original Video

The sound of clicked object…

The sound of clicked object…

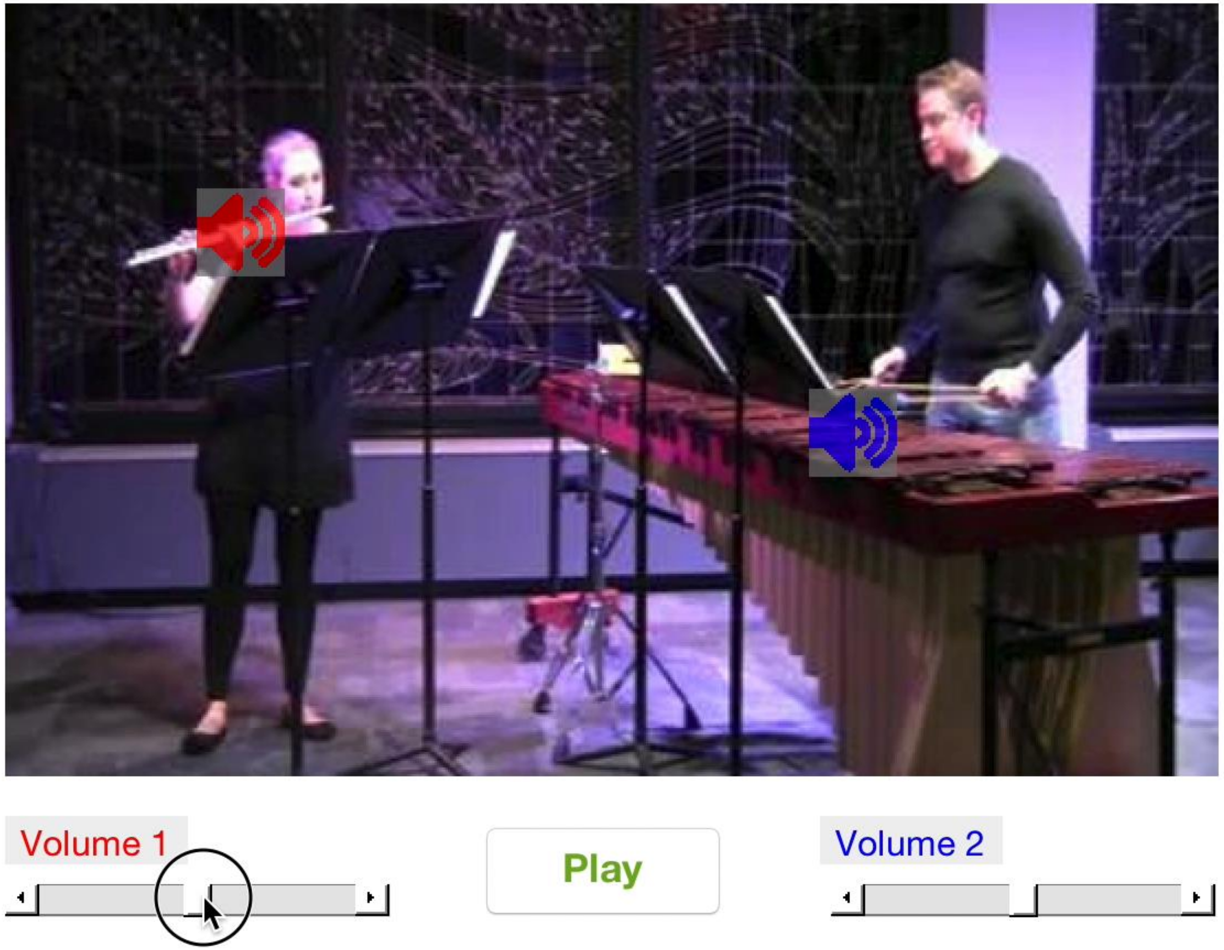


The sound of clicked object…



Application: Music Remix

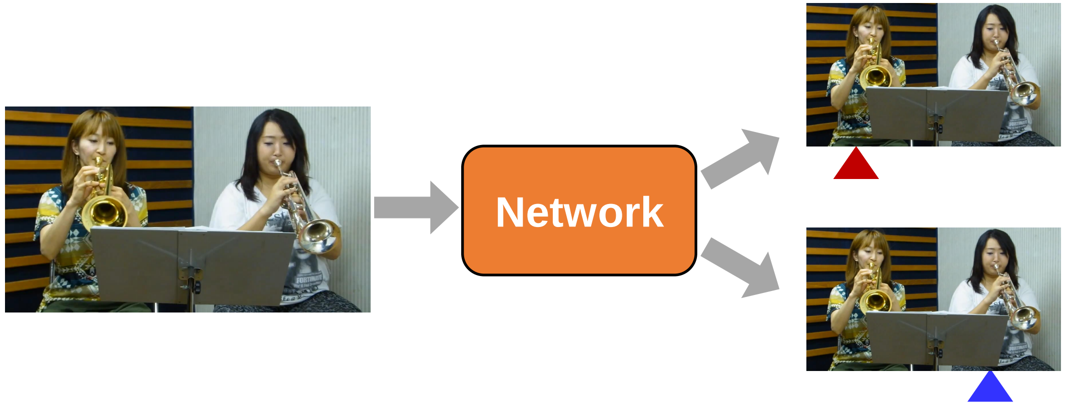
Application: Music Remix



Limitation

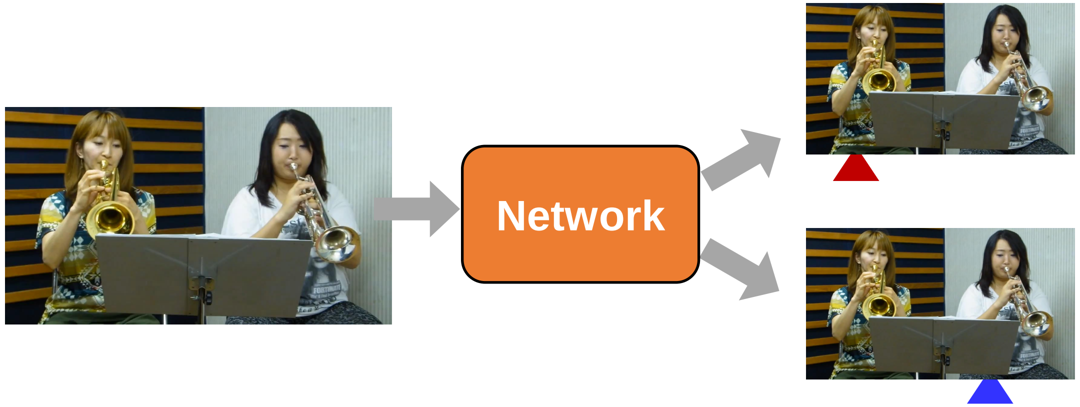
❑ Most existing methods use **raw pixel** or **optical flow** as input.

❑ Problem: limited to **separate multiple instruments of the same types**.



Our Ideas

❑ Problem: limited to **separate multiple instruments of the same types**.

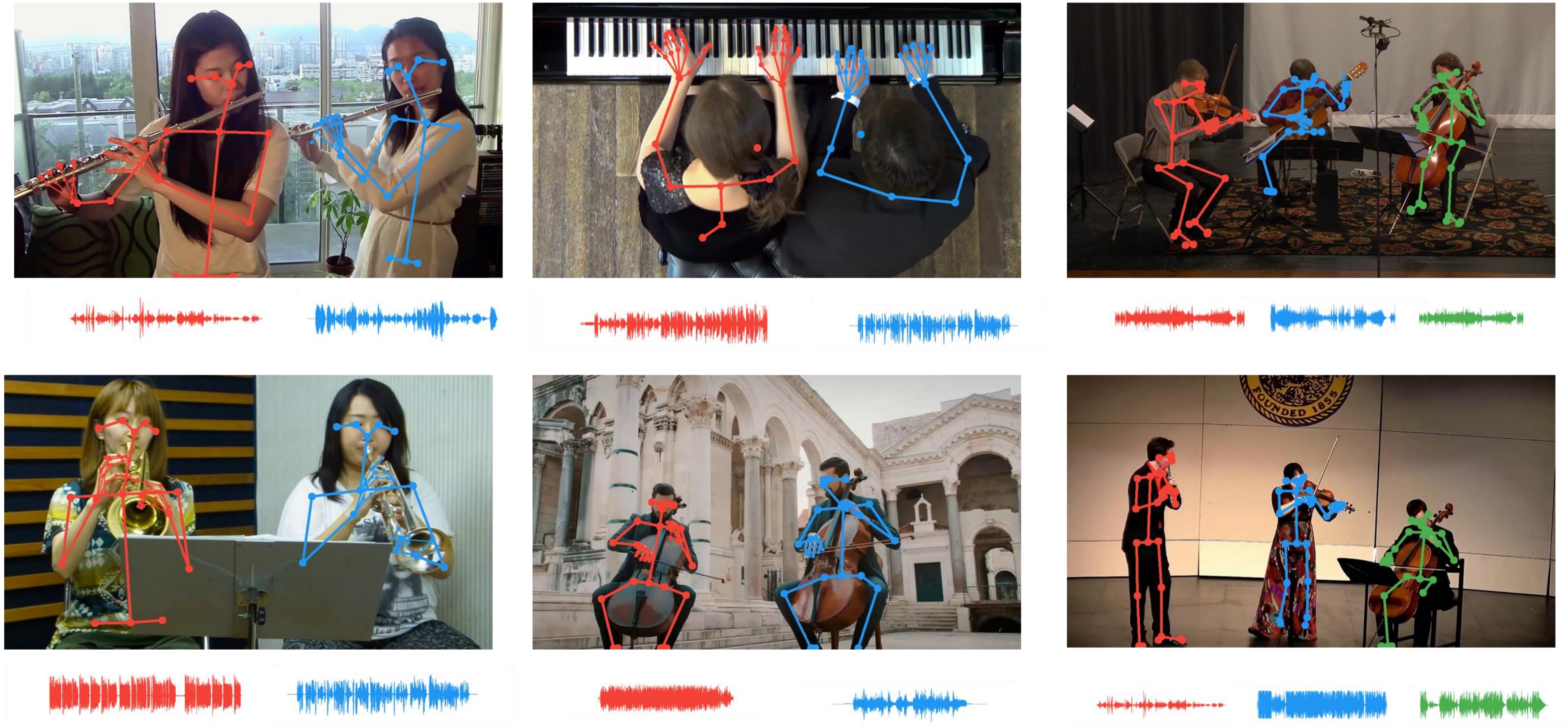


❑ We propose to Identifying a melody by studying a musician's body language using ``**Music Gesture**".

Gan et al. "Music Gesture for Visual Sound Separation." CVPR 2020.

Music Gesture

- Keypoint-based structured representations



Visual sound separation results

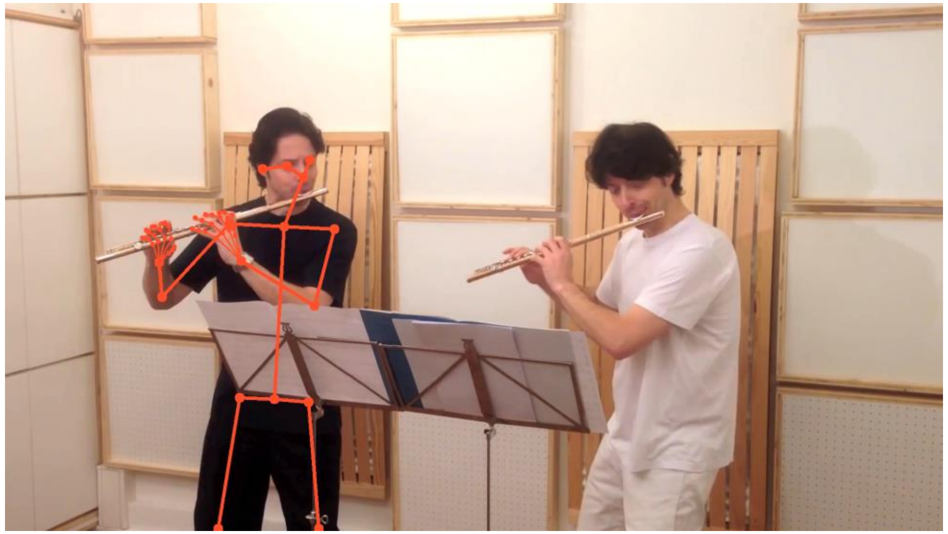
Sound of Motion

Mixed sound

Separated sound1

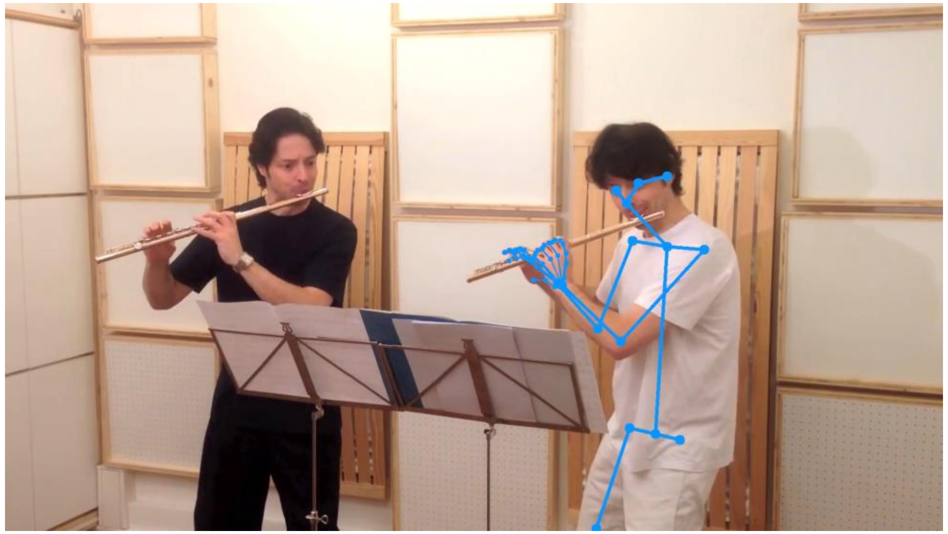
Separated sound2

Music Gesture



Separated sound1

Mixed sound



Separated sound2

Sound of Motion

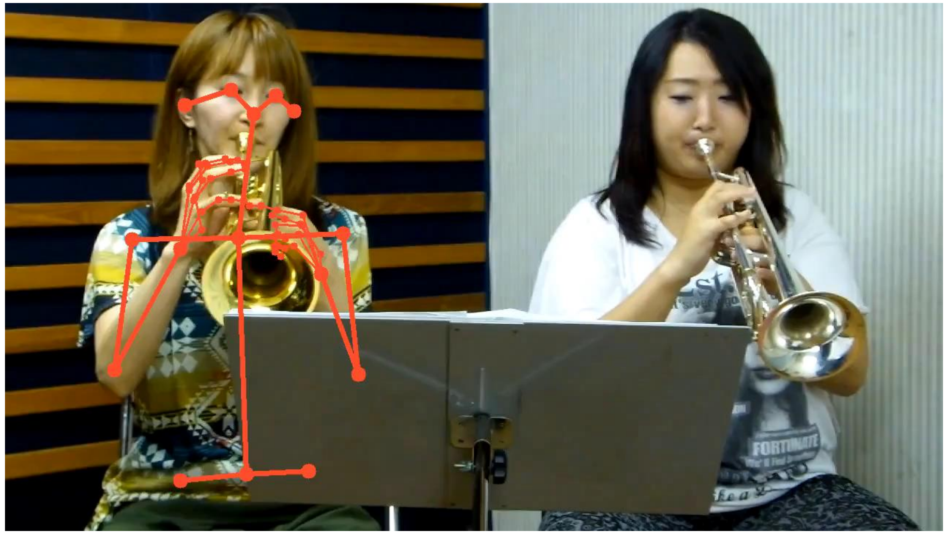
Separated sound1

Mixed sound

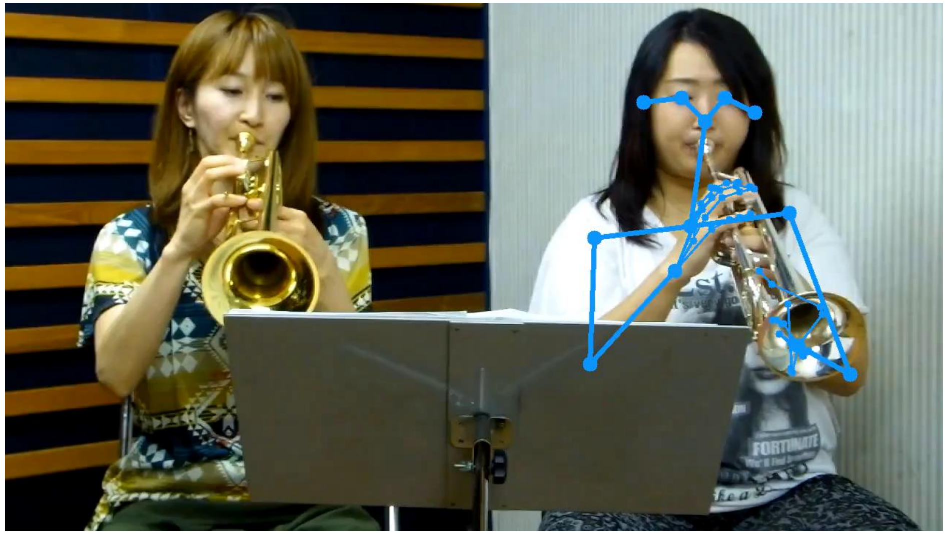
Separated sound2

Music Gesture

Mixed sound



Separated sound1



Separated sound2

Sound of Motion

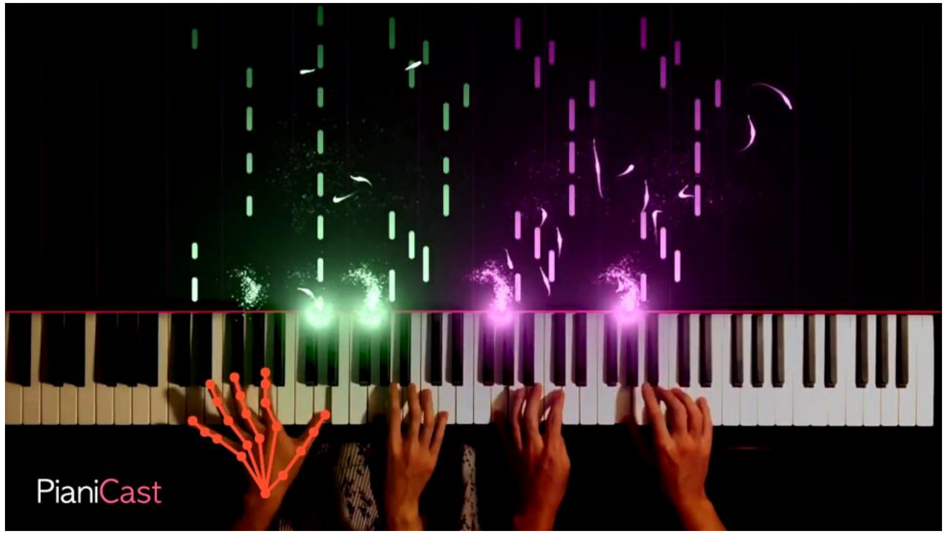
Mixed sound

Separated sound1

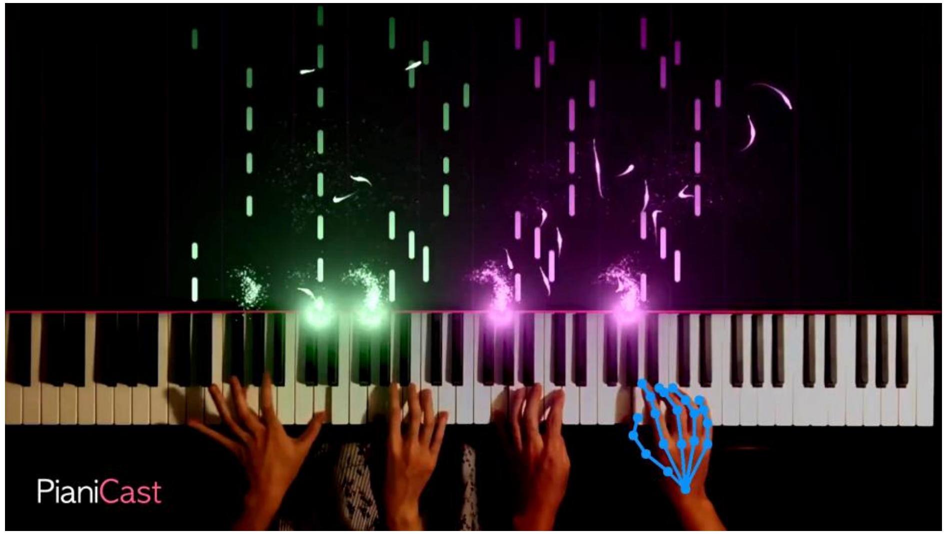
Separated sound2

Music Gesture

Mixed sound



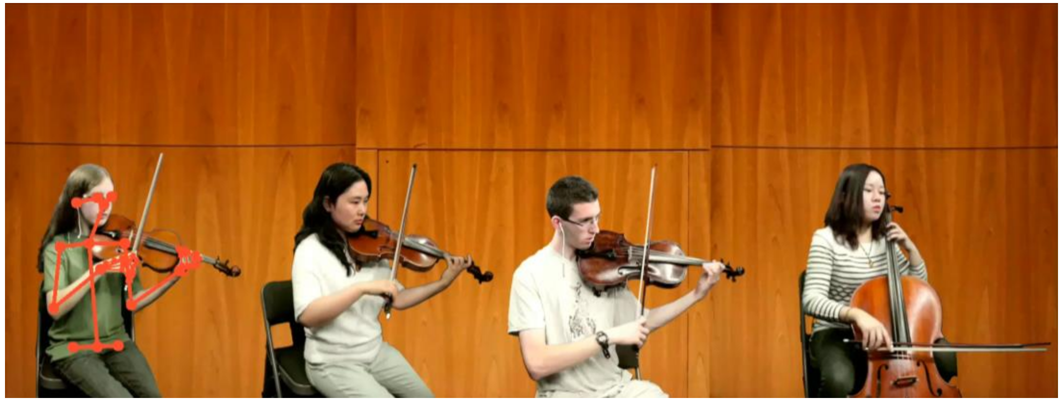
Separated sound1



Separated sound2

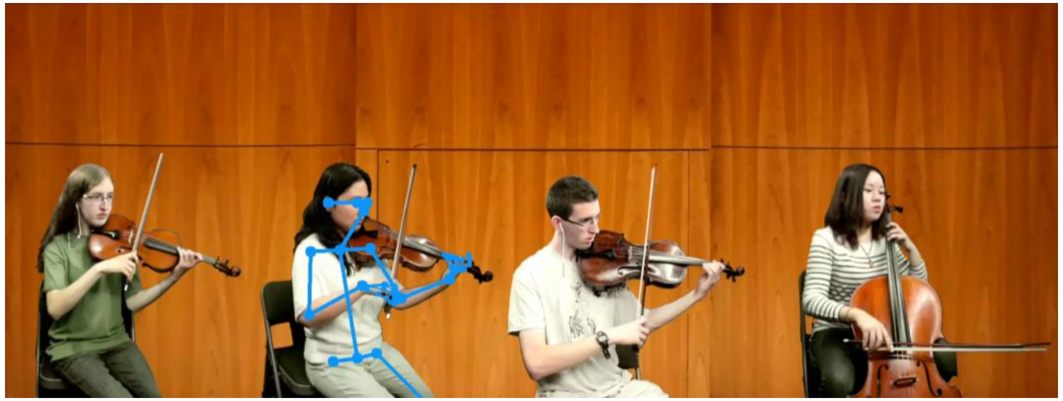
Multiple instruments

Music Gesture



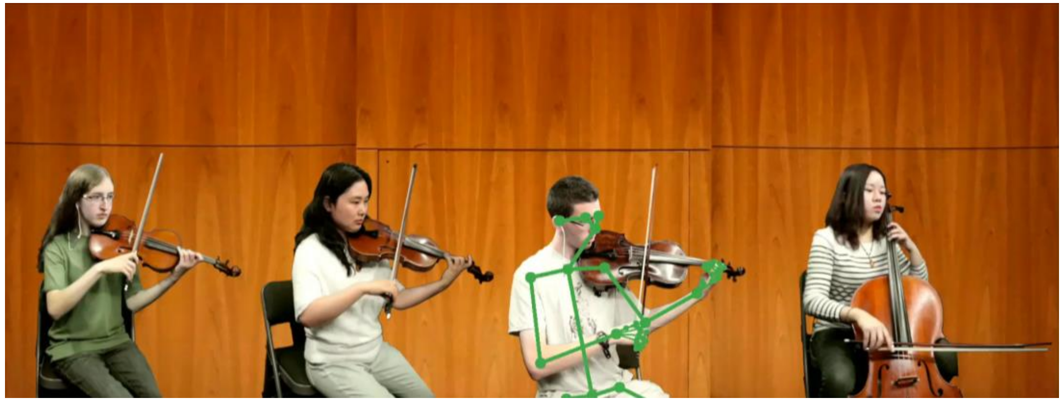
Separated sound1

Mixed sound



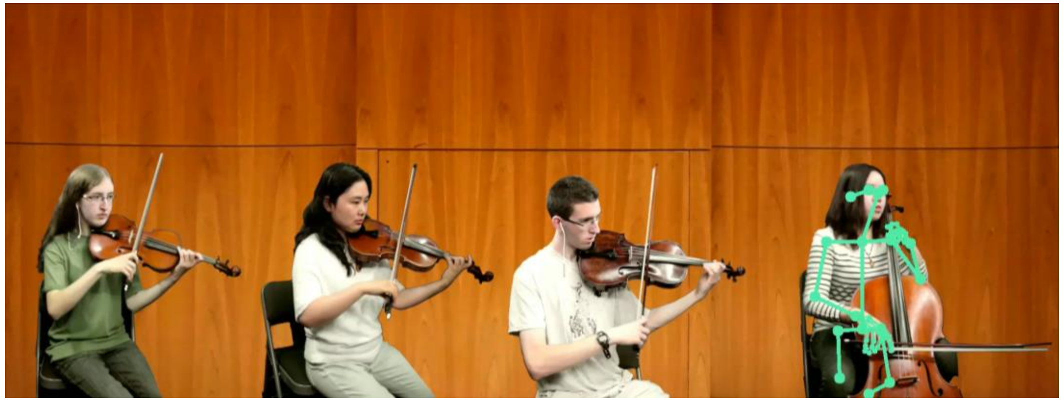
Separated sound2

Music Gesture



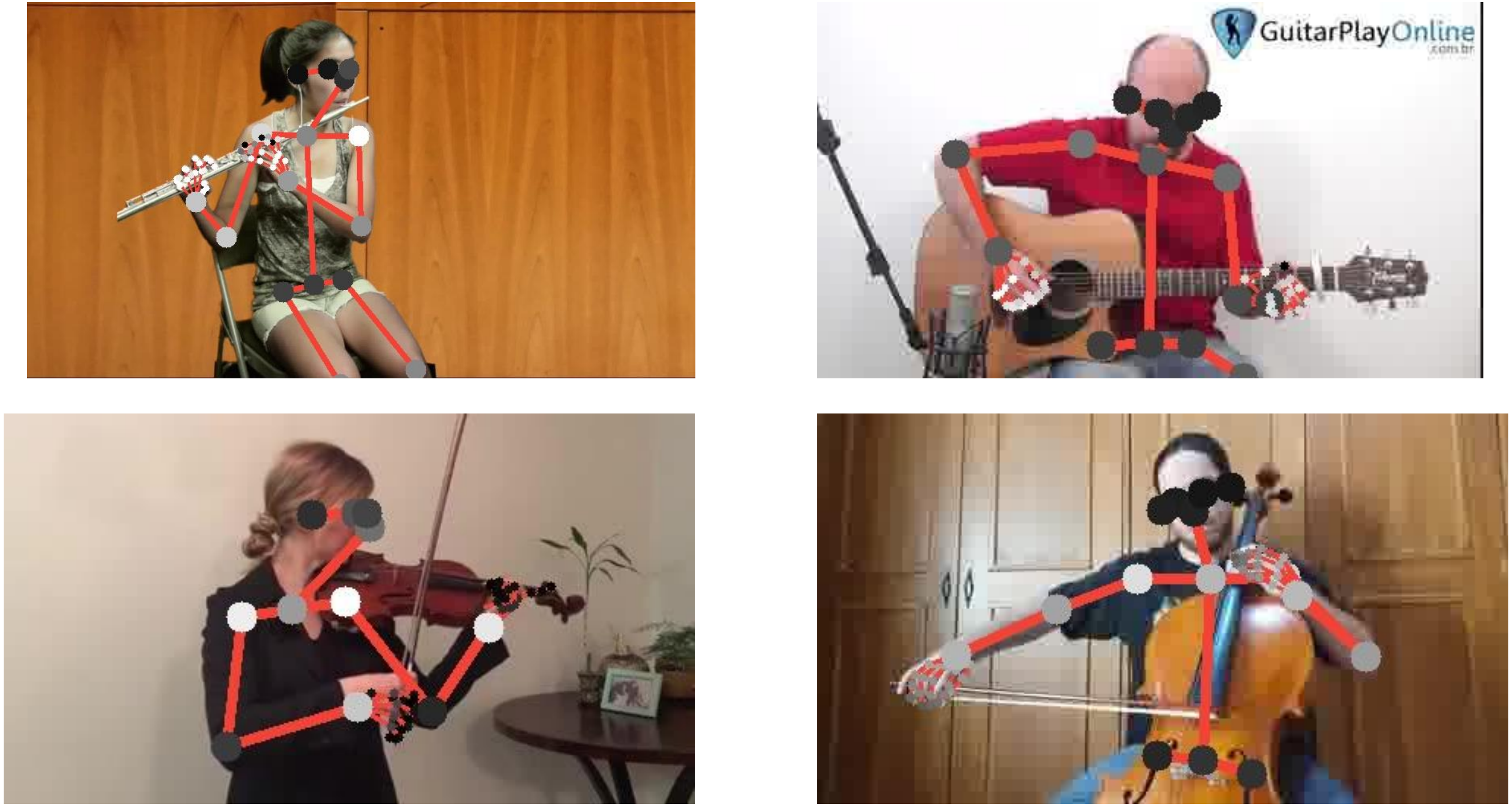
Separated sound3

Mixed sound



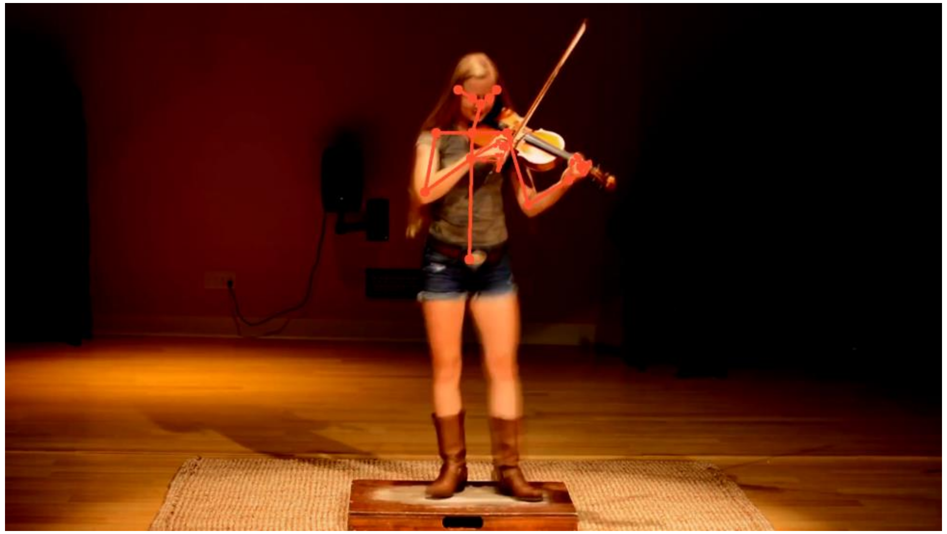
Separated sound4

Attention Map of Key points

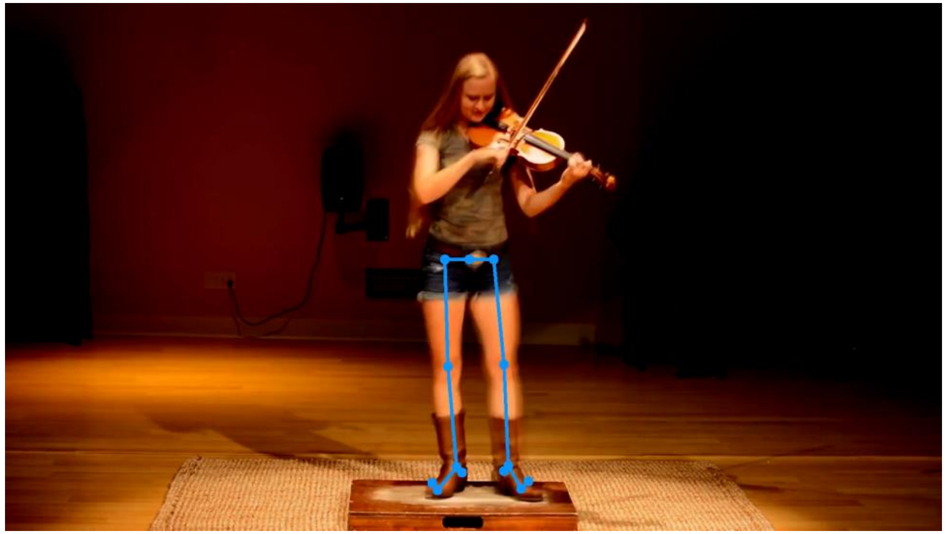


The sound of body parts

Mixed sound



Separated sound1



Separated sound2

Can we generate music from videos?

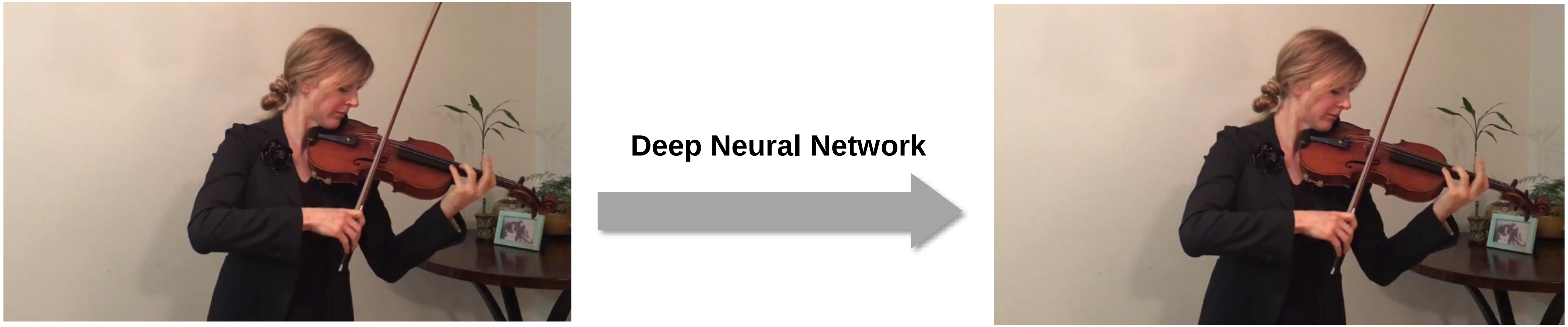
Given a silent music performance video…

Silent music performance video

Gan et al. “Foley Music: Learning to Generate Music from Videos.” ECCV 2020.

Can we generate music from videos?

…we aim to generate plausible music.



Deep Neural Network

Silent music performance video

Performance with generated sound

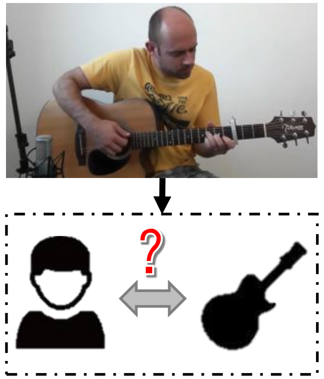
Gan et al. “Foley Music: Learning to Generate Music from Videos.” ECCV 2020.

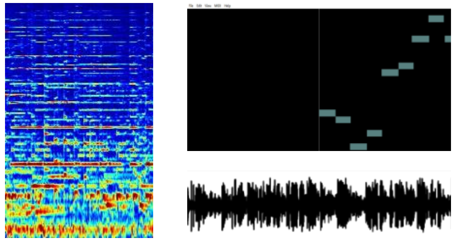
Challenges

□ Hard to learn visual-audio mappings from unlabeled video

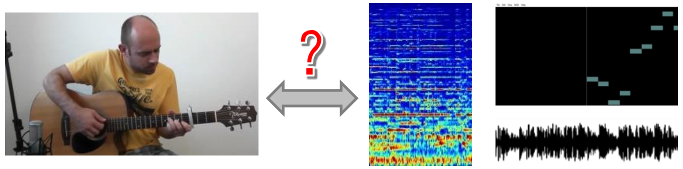
□ Three things matter：

- ◆ Visual perception module → interactions between instrument and player
- ◆ Audio representation → musical rules, easy to predict from visual signals
- ◆ Visual-audio model → association between two modalities





Choose ?



Challenges

□ Hard to learn visual-audio mappings from unlabeled video

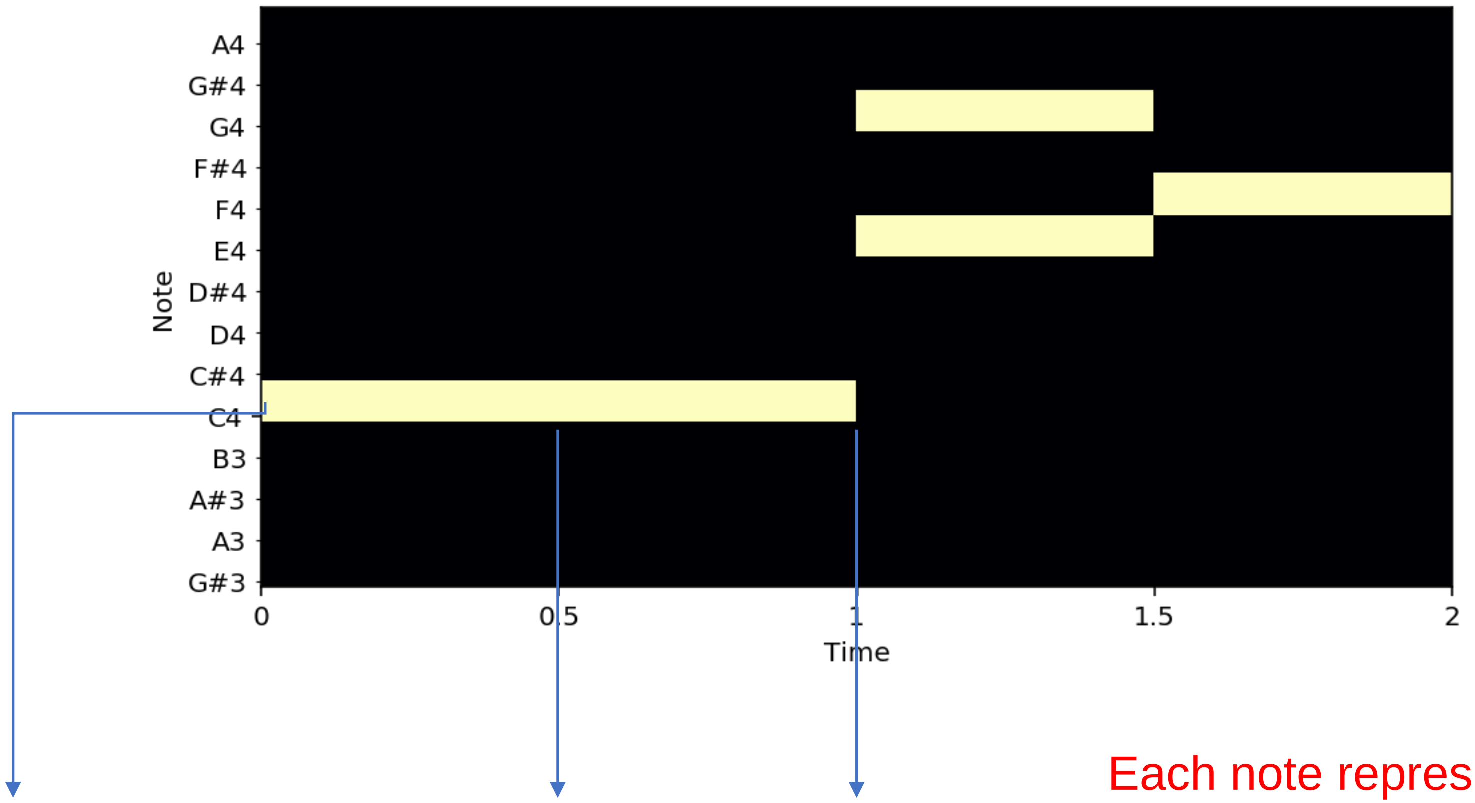
□ Three things matter：

- ◆ Visual perception module → interactions between instrument and player
- ◆ Audio representation → musical rules, easy to predict from visual signals
- ◆ Visual-audio model → association between two modalities

We use **body keypoints** to explicitly model the body and finger.

We use **Musical Instrument Digital Interface (MIDI)** to represent music.

MIDI Event Representations



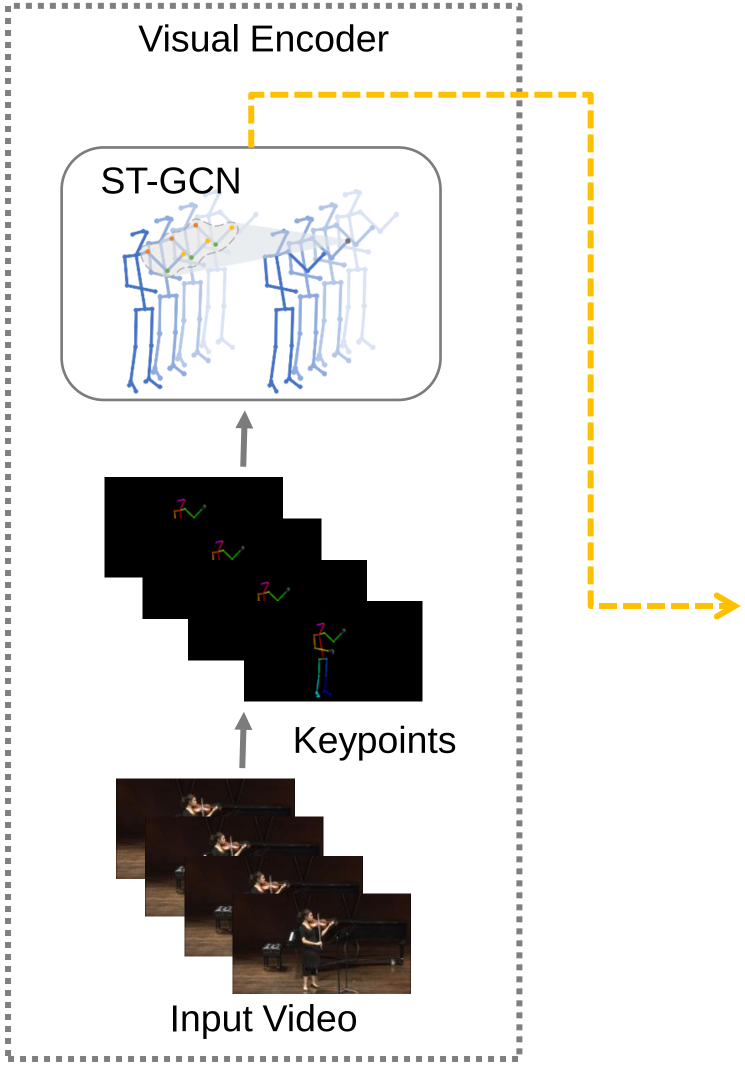
Velocity Event,	Note On Event	Time Shift Event	Note Off Event
201	39	237	127

Each note represented as a sequence of MIDI Events

- ❑ Note On Event$\in [0, 88]$, based on Pitch
- ❑ Note Off Event$\in [88, 176]$, based on Pitch
- ❑ Velocity Event$\in [176, 208]$, based on Velocity
- ❑ Time Shift Event$\in [208, 240]$, based on Duration

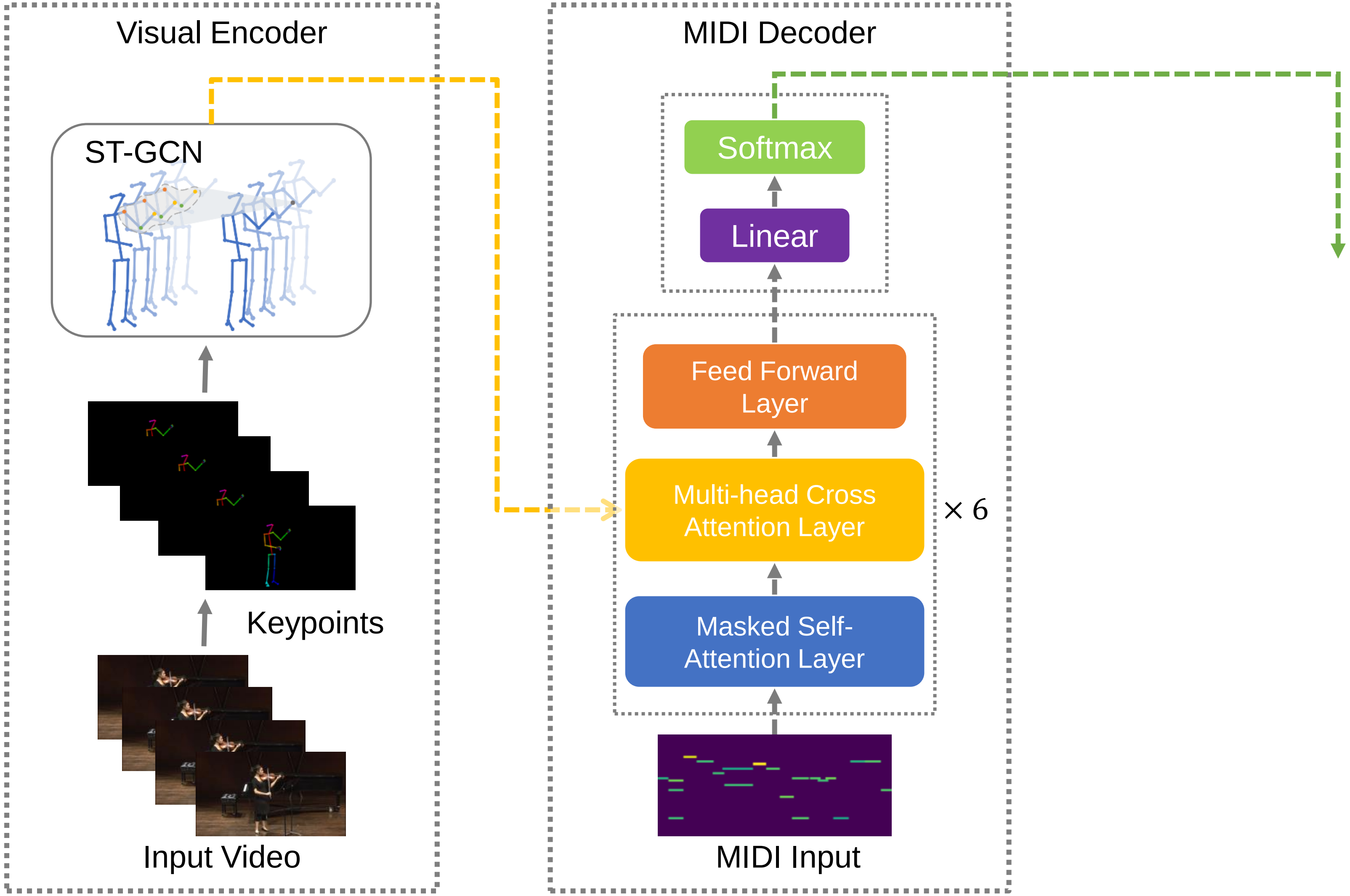
Our method

We first encode human keypoints by using graph CNN.



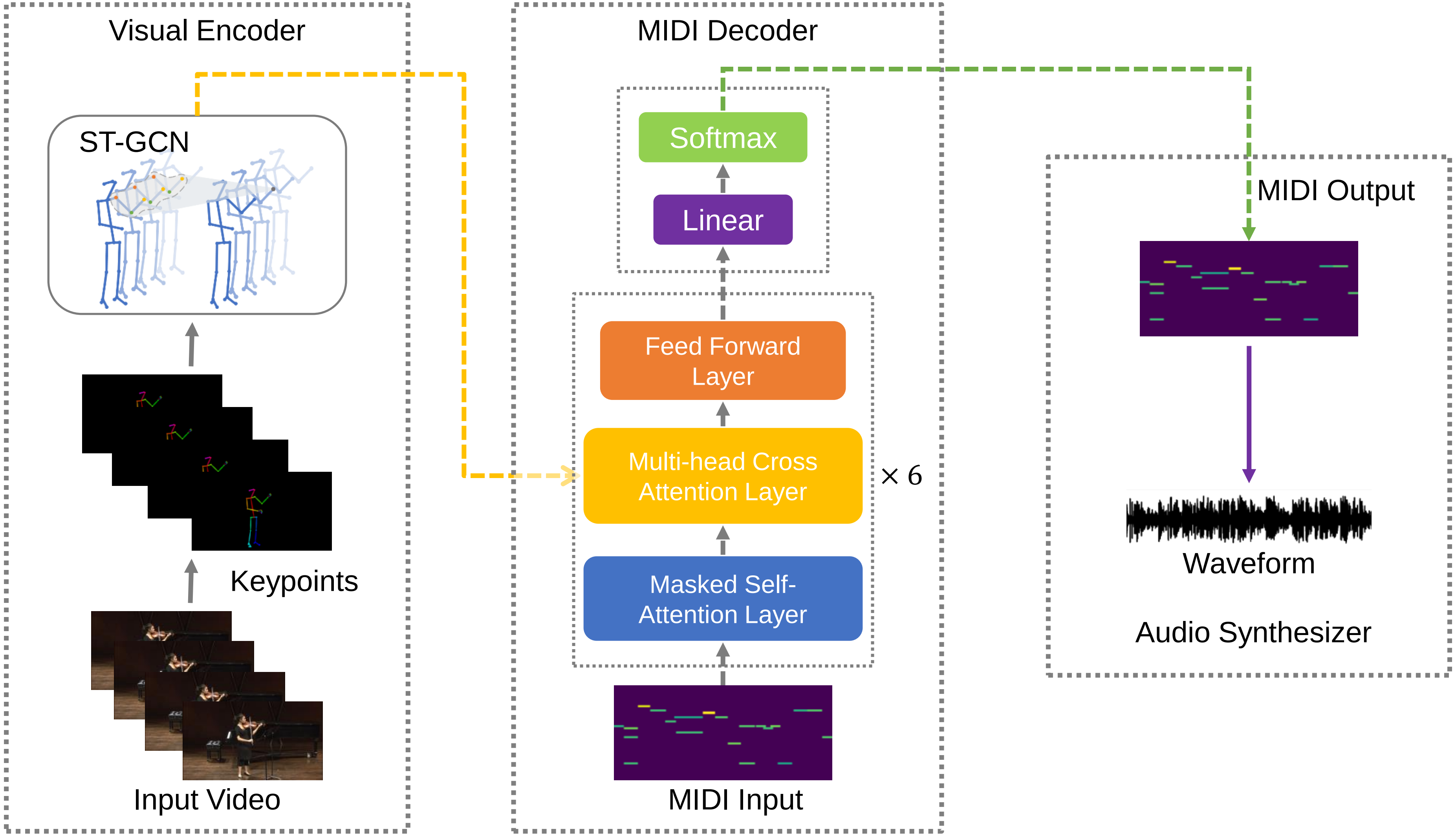
Our method

We then translate the keypoints features into MIDI using Transformers.



Our method

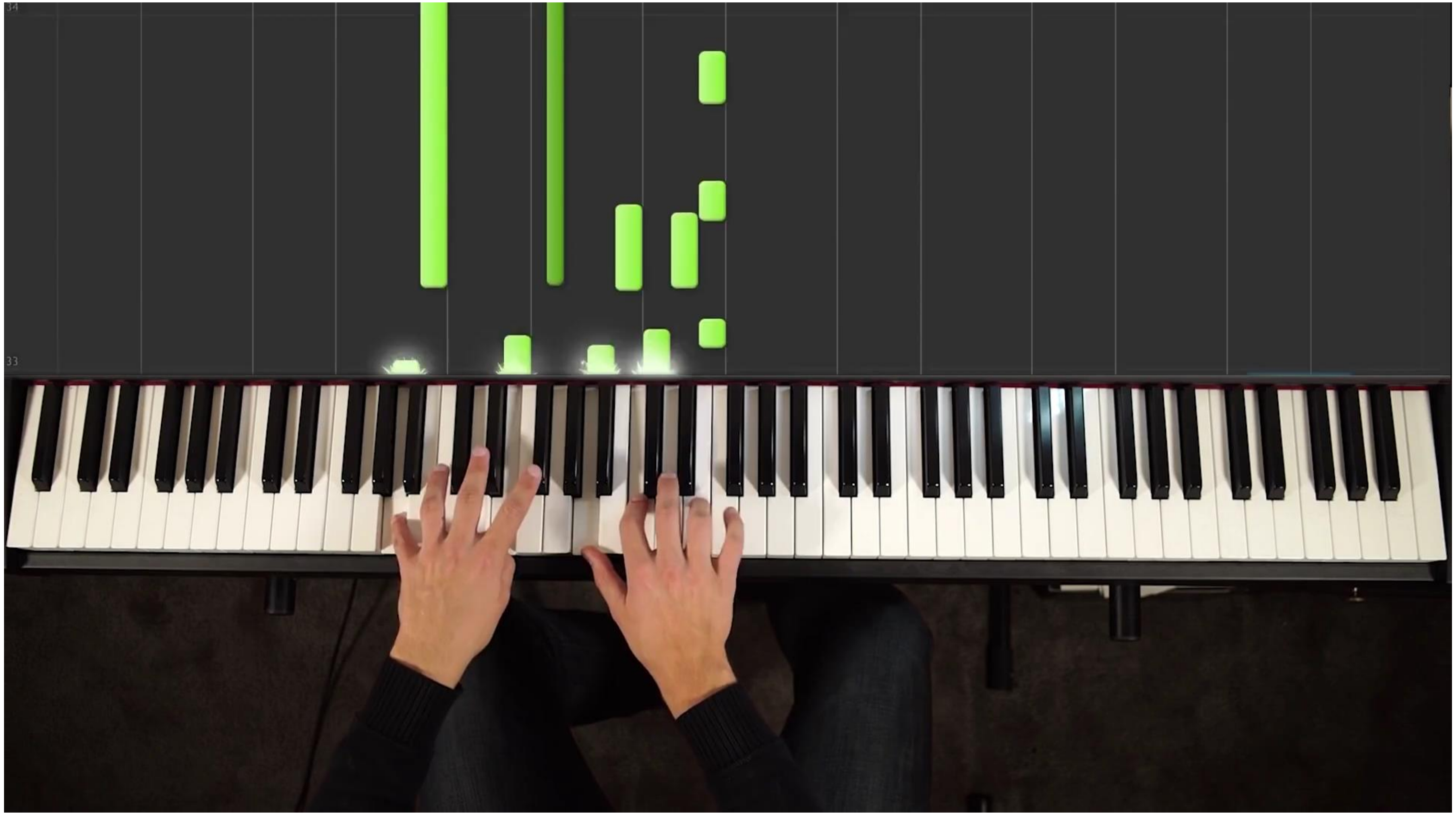
Finally, we synthesize audio from MIDI.



Music generation results

Ukulele

Piano



Guitar

Style editing results

Bass

Original prediction

Style editing

A major

F major

G major

Tuba

Original prediction

Style editing

A major

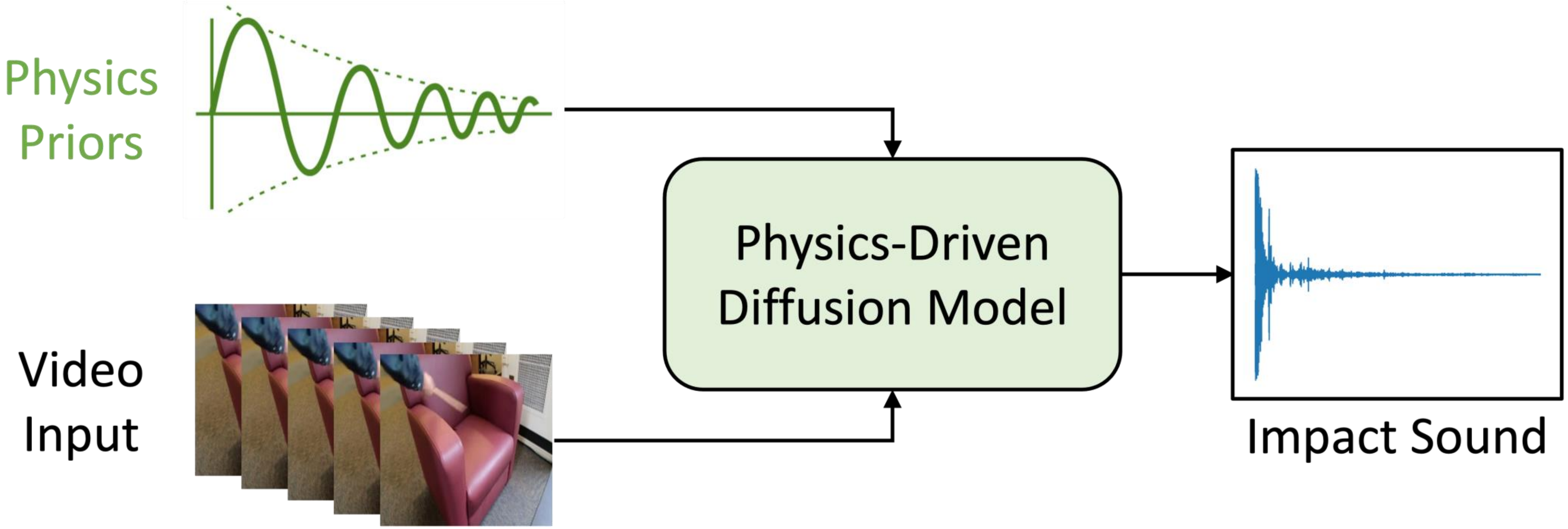
F major

G major

An impact sound of physical object interactions is critical to perception

Could we generate the impact sound from vision?

Our Framework



Our model is applicable to
a variety of videos and materials

The transparency of **physics priors** allows us to perform interesting **sound editing**

Physics Priors of **Glass** + Video Input

Original Result

Physics Priors of **Glass** + Video Input

Original Result

Transformed Result

Physics Priors of **Cloth** + Video Input

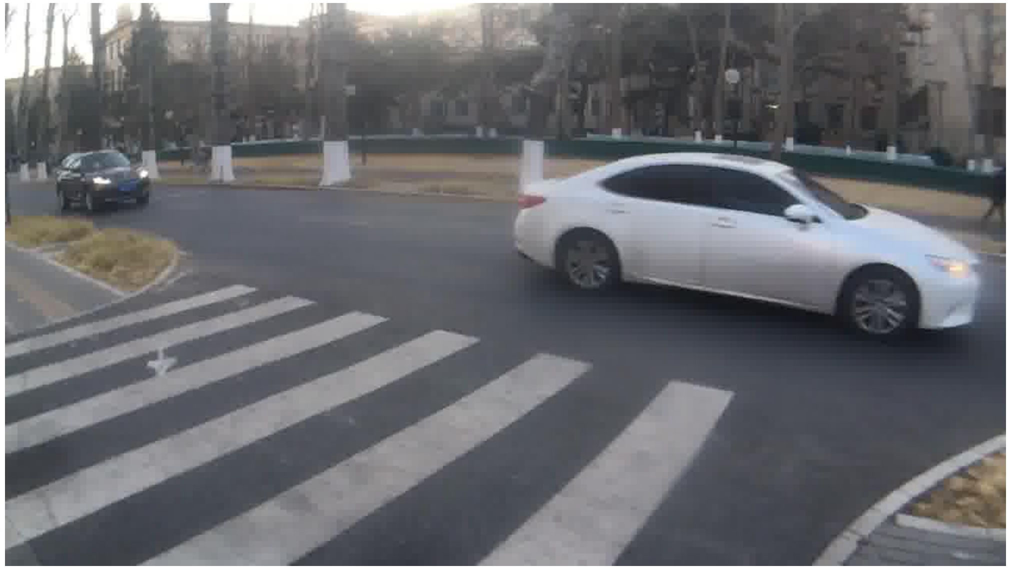
Original Result

Physics Priors of **Cloth** + Video Input

Original Result

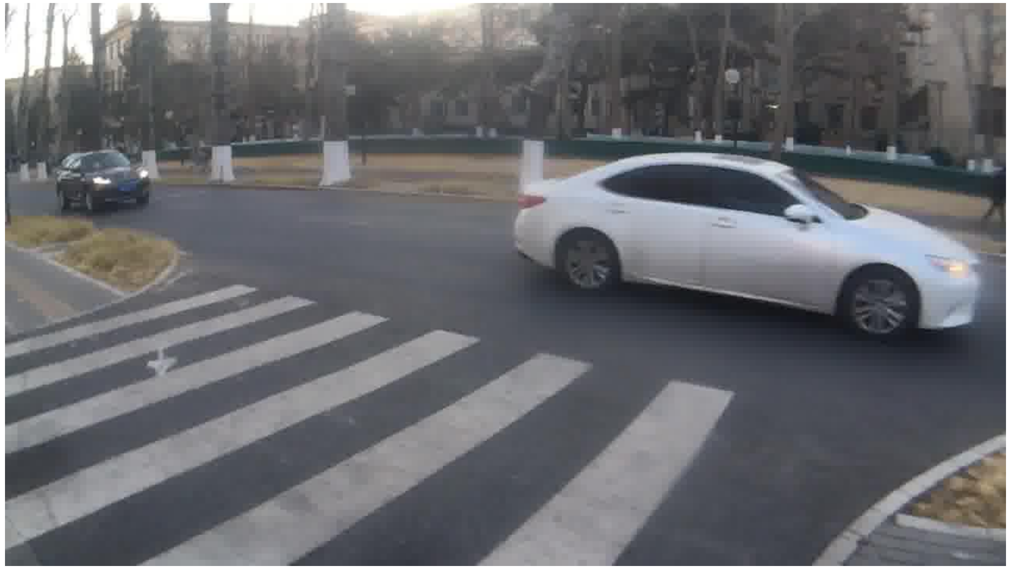
Transformed Result

Given an input video…

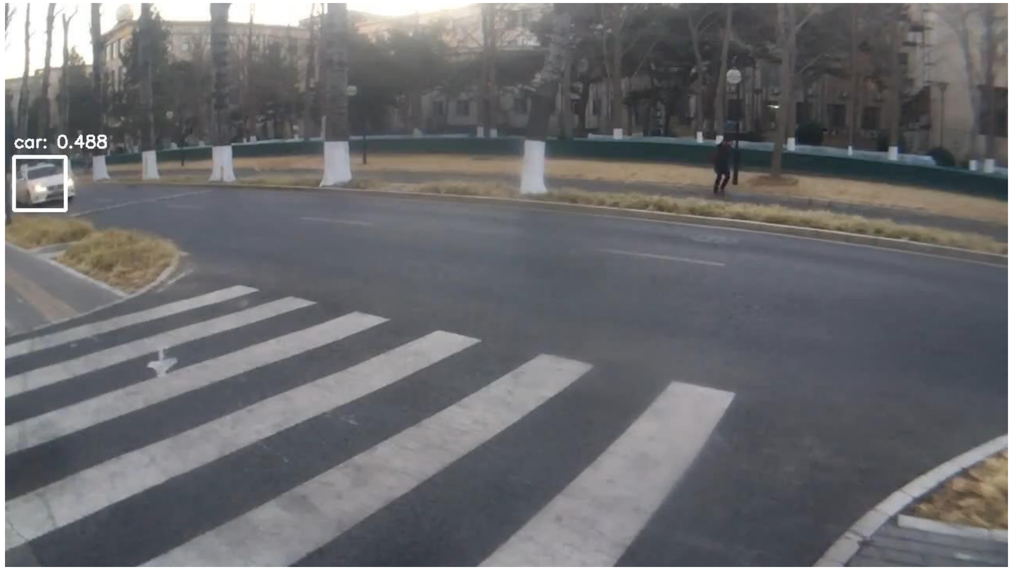


Self-supervised moving vehicle tracking with stereo sound. Gan et.at. ICCV 19

Given an input video…

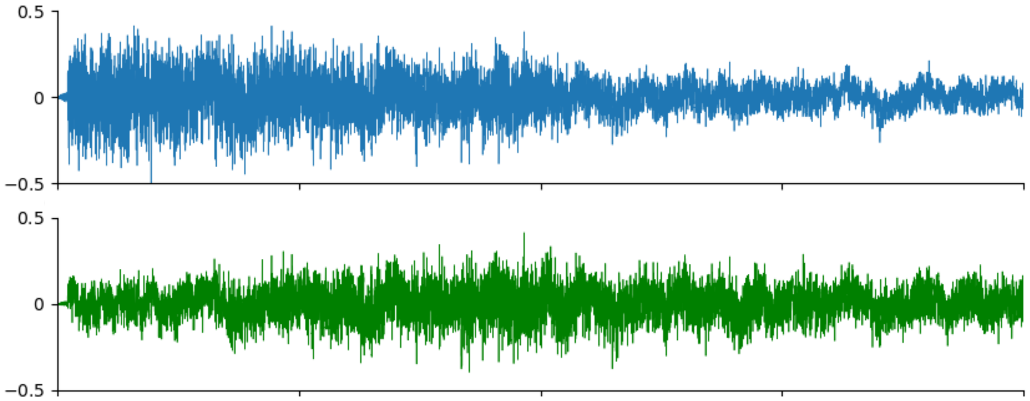


Visual detection network can track the vehicles using visual input.

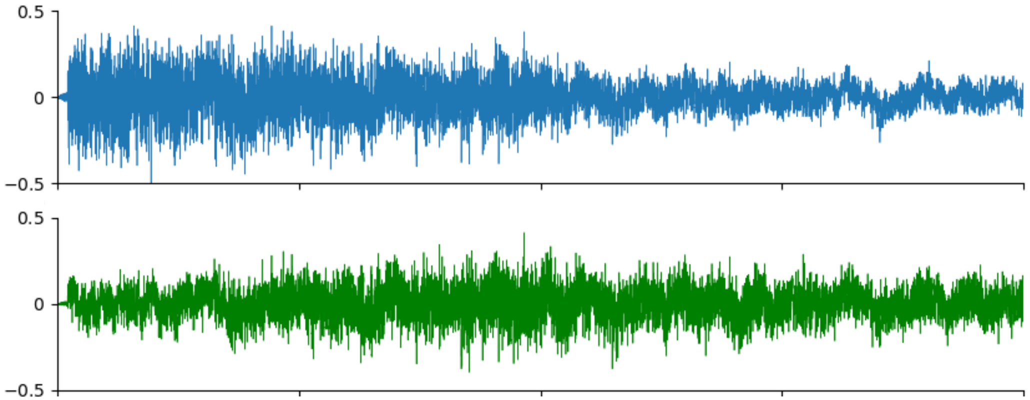


Visual tracking

What if given a piece of input stereo sound only?



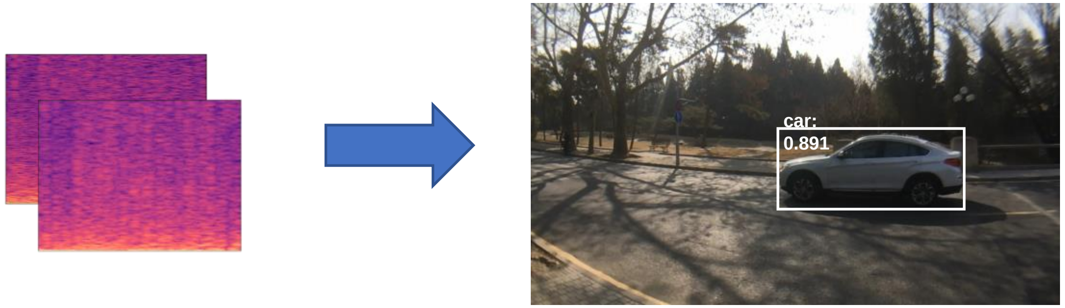
What if given a piece of input stereo sound only?



Where are the vehicles?

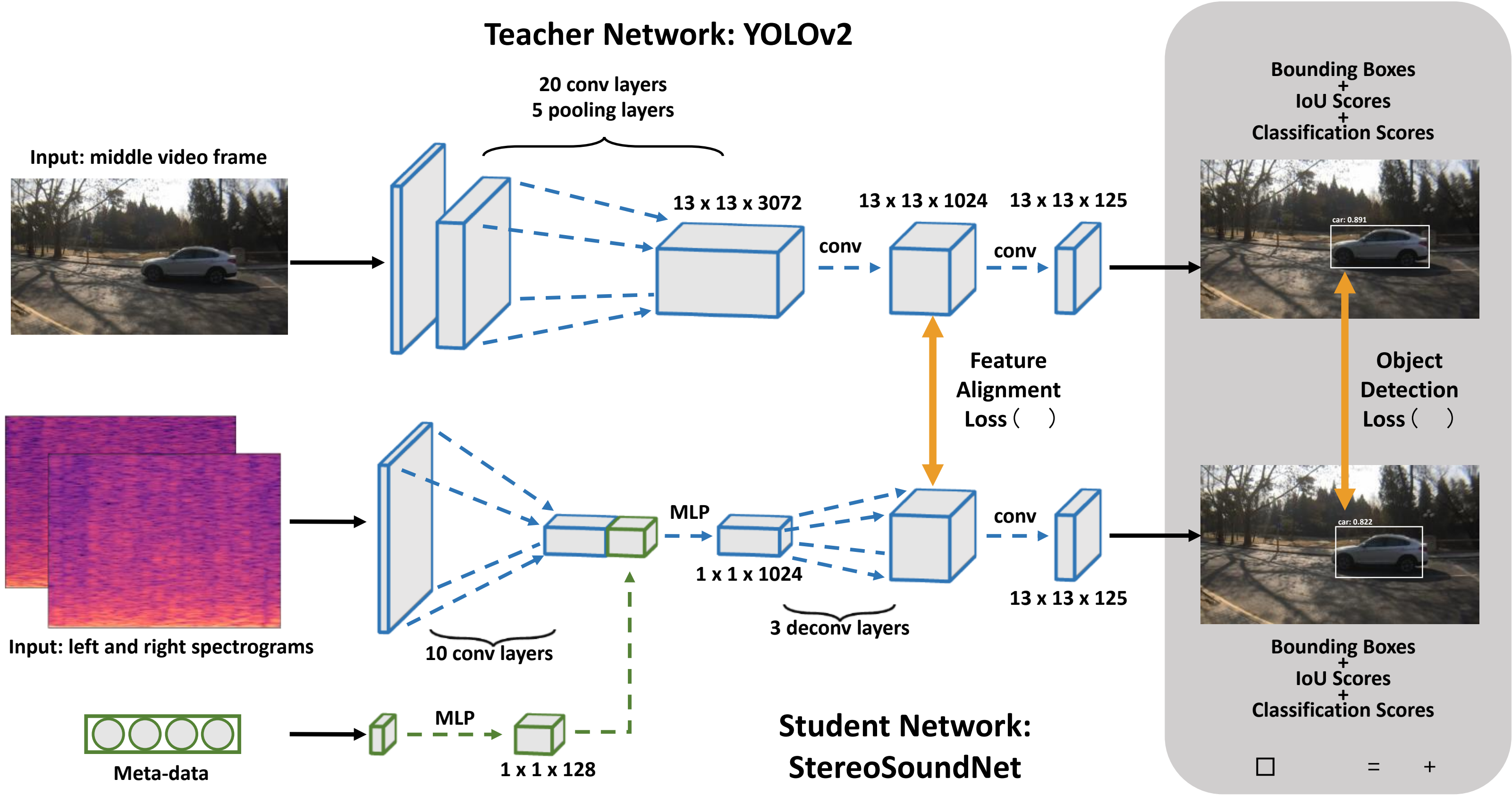
Applications

- Tracking under poor lighting scene
- Tracking under visual occlusion scene
- Energy-efficiency, privacy-preserving



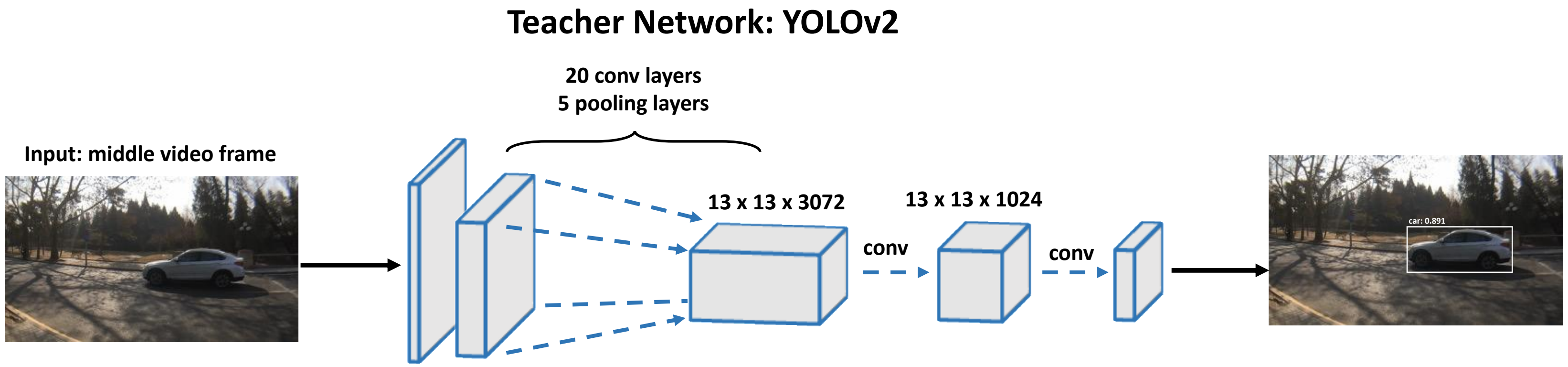
Our Methods

- Teacher-student alignment



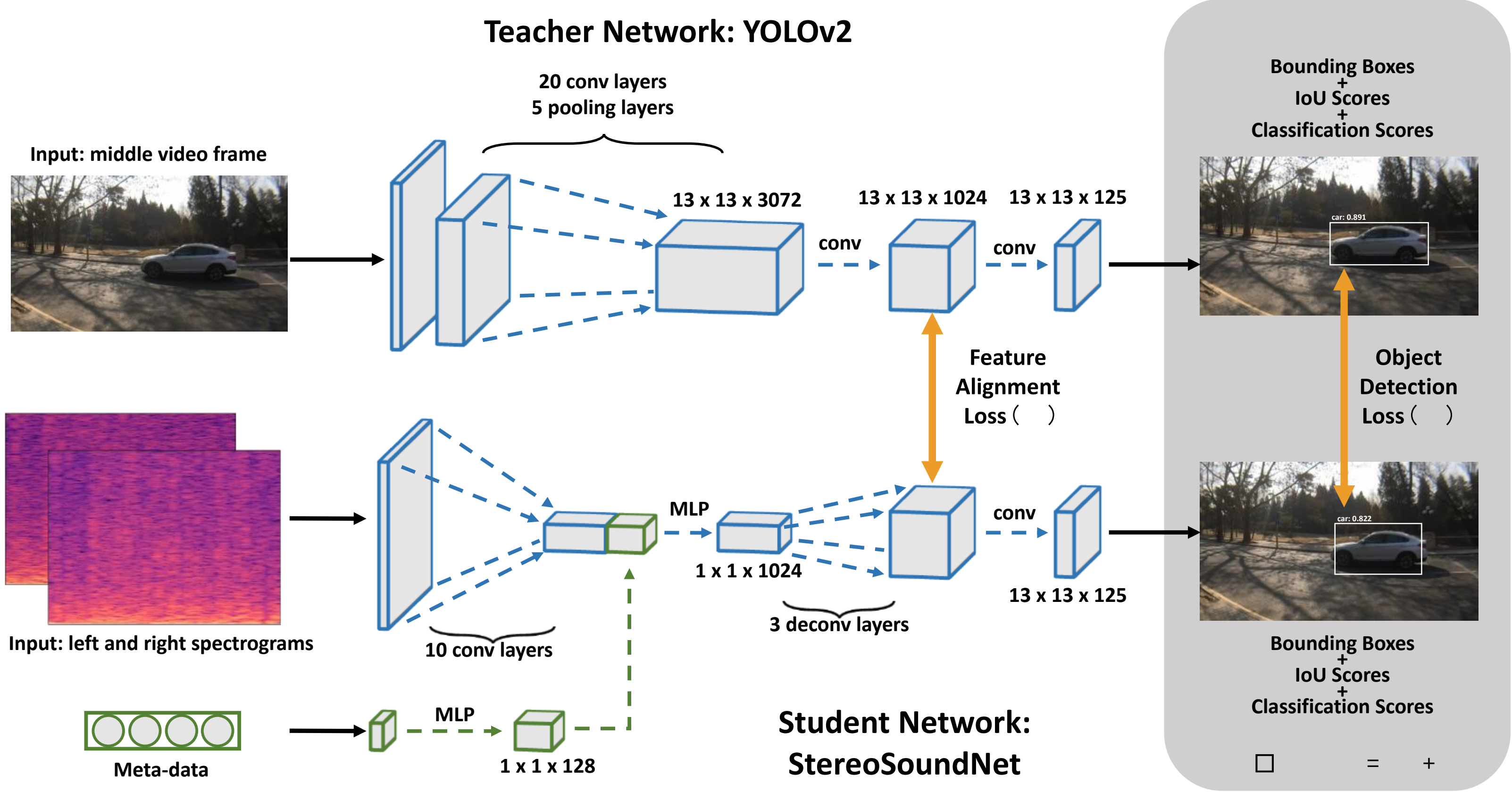
Our Methods

- We first get **pseudo localization labels** and **visual features** from a pre-trained YOLO



Our Methods

- We then train a sound branch using
  - Feature alignment
  - Object alignment

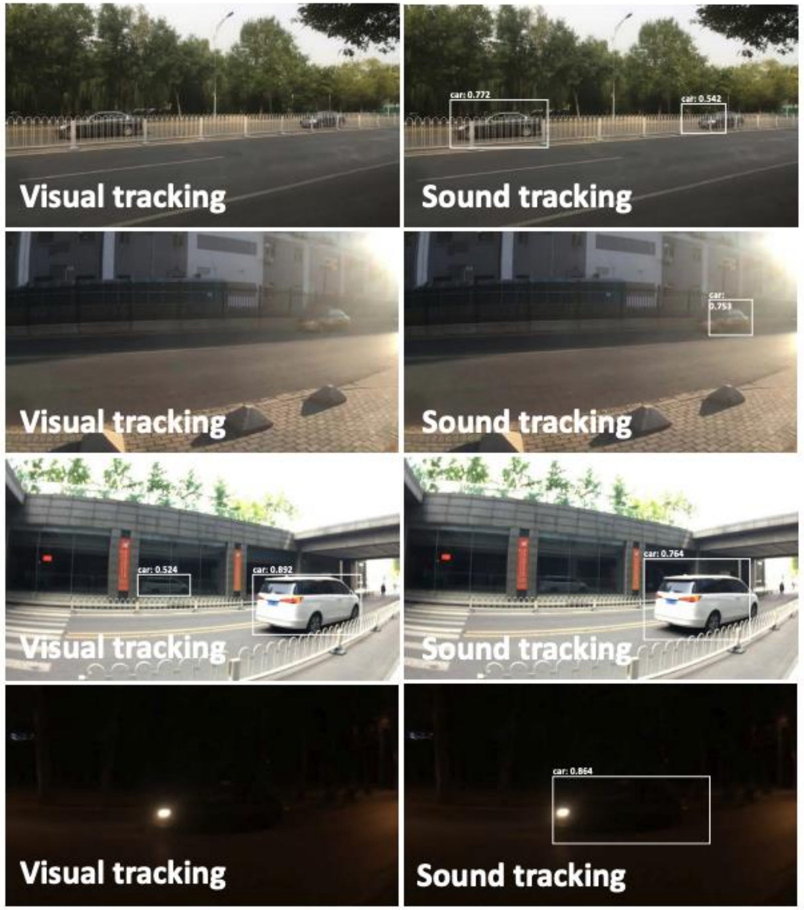


Datasets

- We have collected a dataset on diverse scenes

Results

- Visual tracking fails in many situations:
  - Occlusion
  - Backlighting
  - Reflection on the windows
  - Night scene

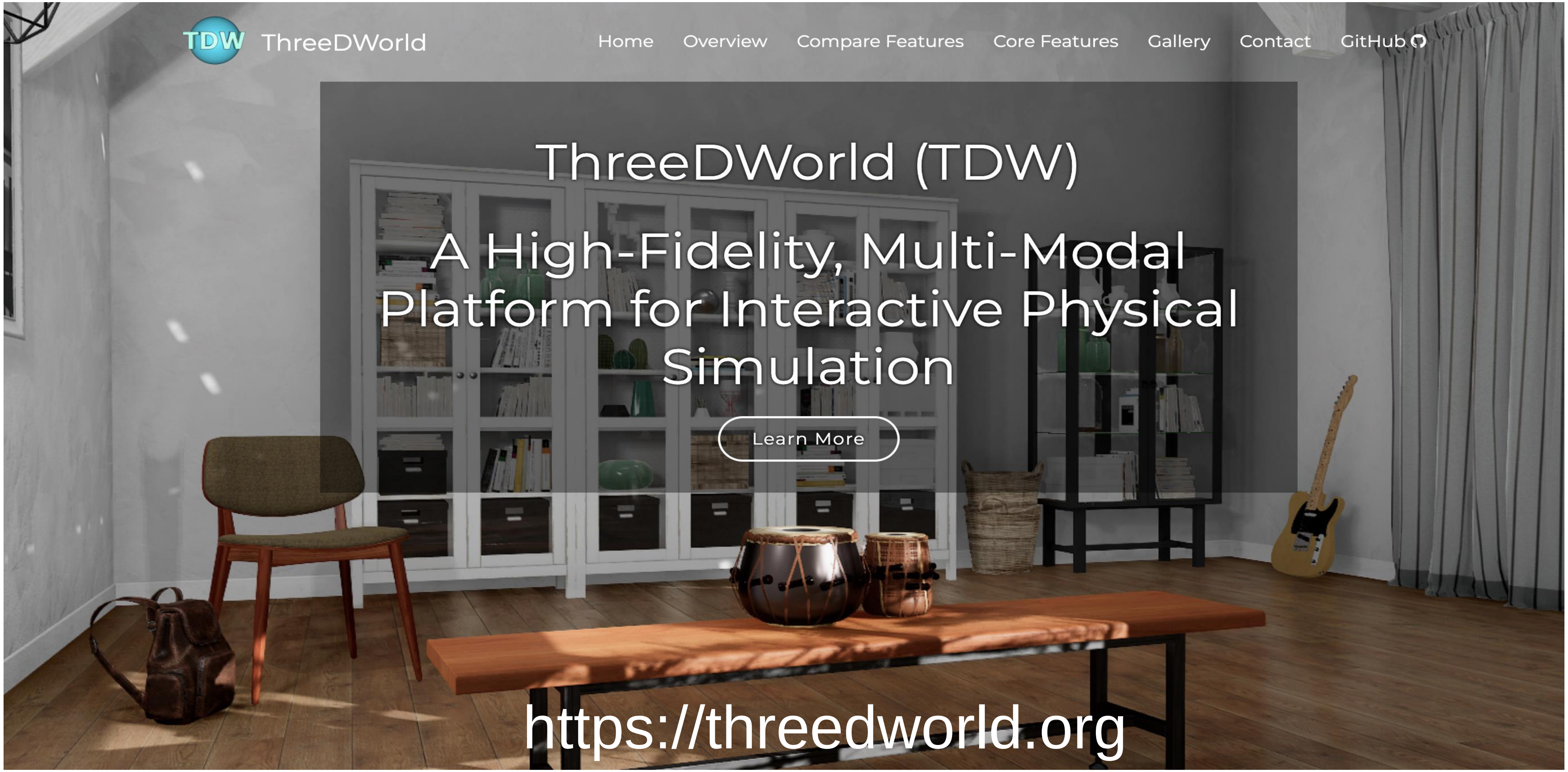


Results

Sound tracking

Tracking Under Poor Lighting

TheeDWorld: 3D Virtual World



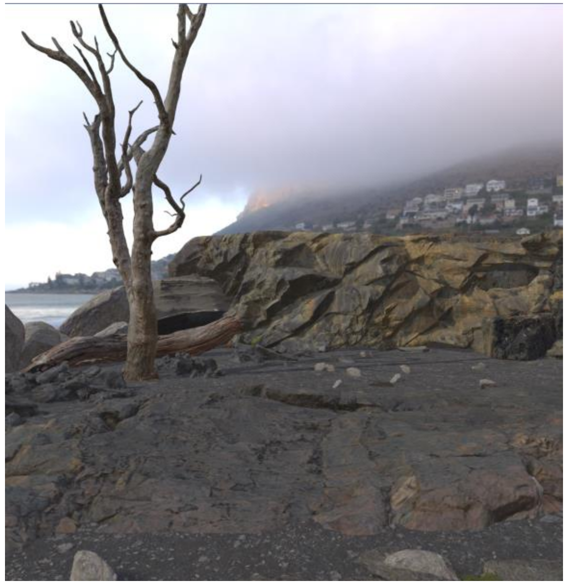
ThreeDWorld (TDW)

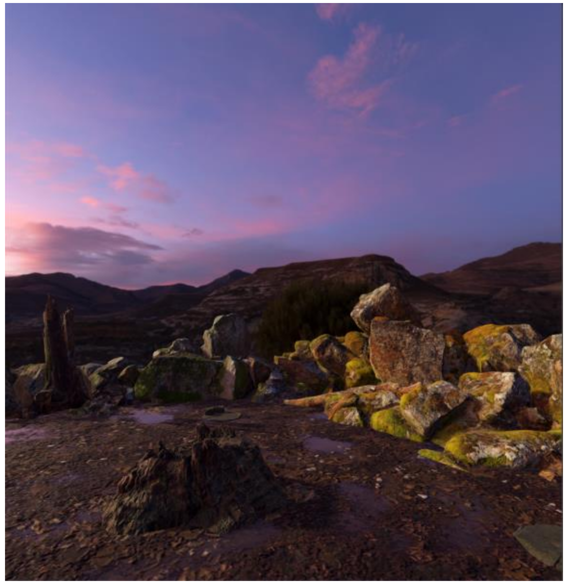
A High-Fidelity, Multi-Modal Platform for Interactive Physical Simulation

https://threedworld.org

Gan, Jeremy, et al, Threedworld: A platform for interactive multi-modal physical simulation. NeurIPS, 21

Vision: Photo-Realistic Rendering





Audio: Physics-Triggered Sound

Physics Simulation

Rigid-body

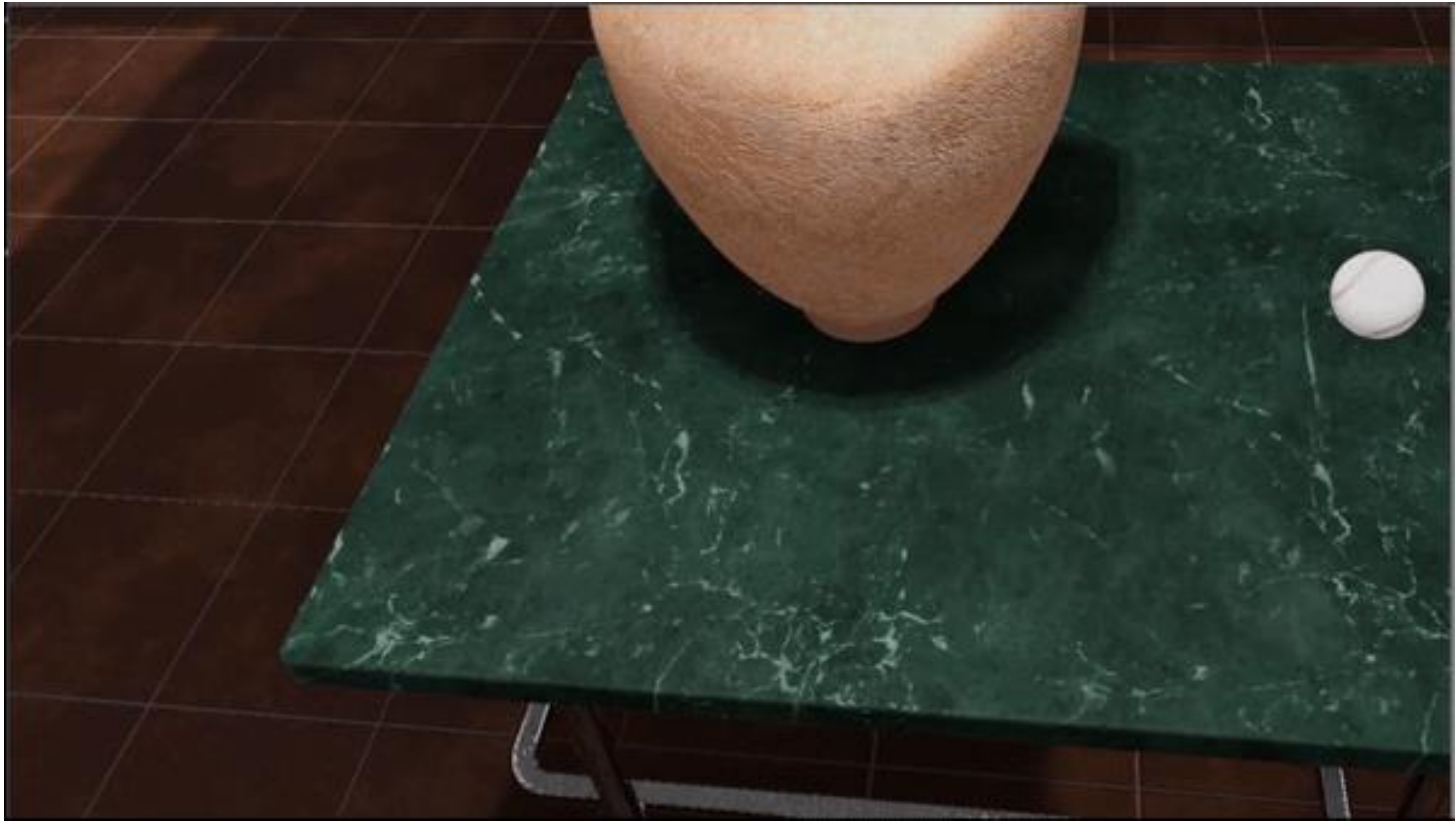
Soft-body

Agent Interaction

Robot

Humanoid Avatar

Object Interaction in VR



t

QA?