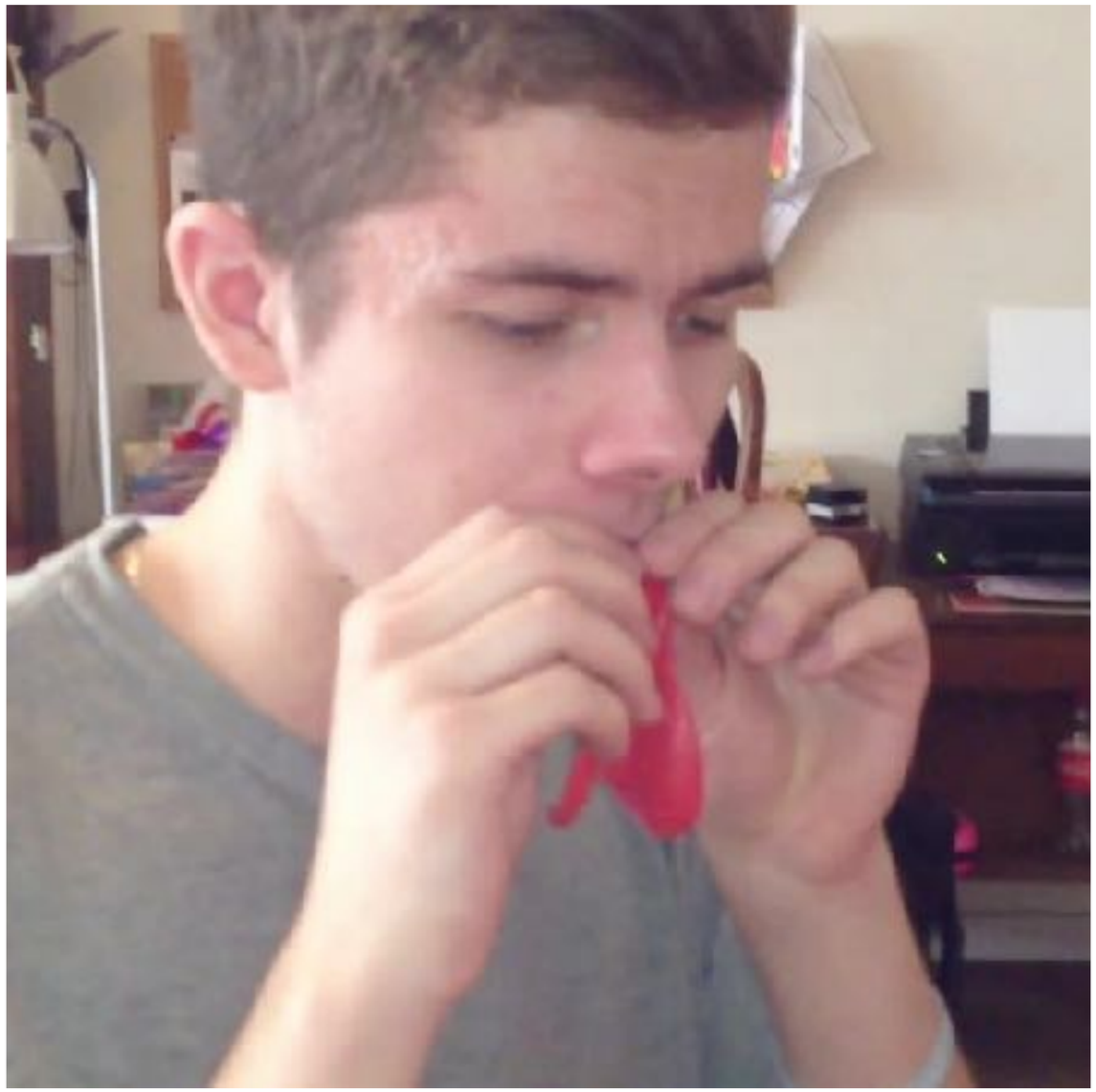
Learning from Unlabeled Video

Carl Vondrick

Google Research

What color is that pixel?

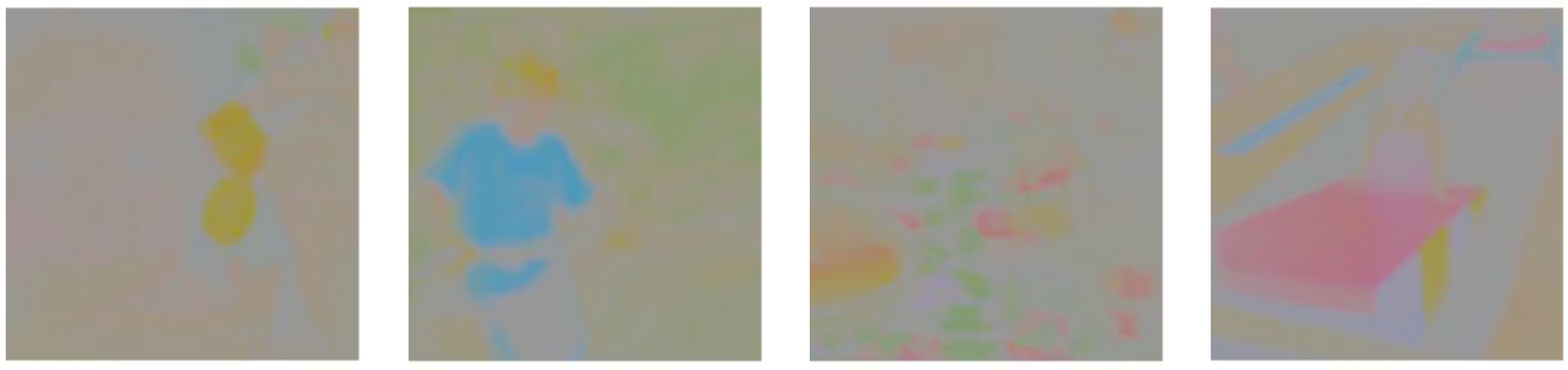


Time

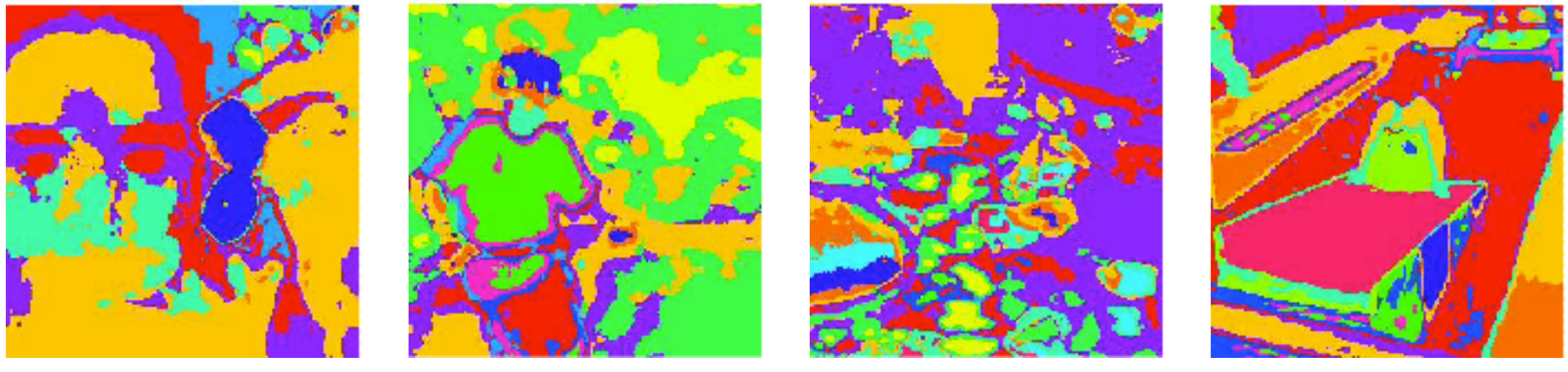
Temporal Coherence of Color

RGB

Color Channels

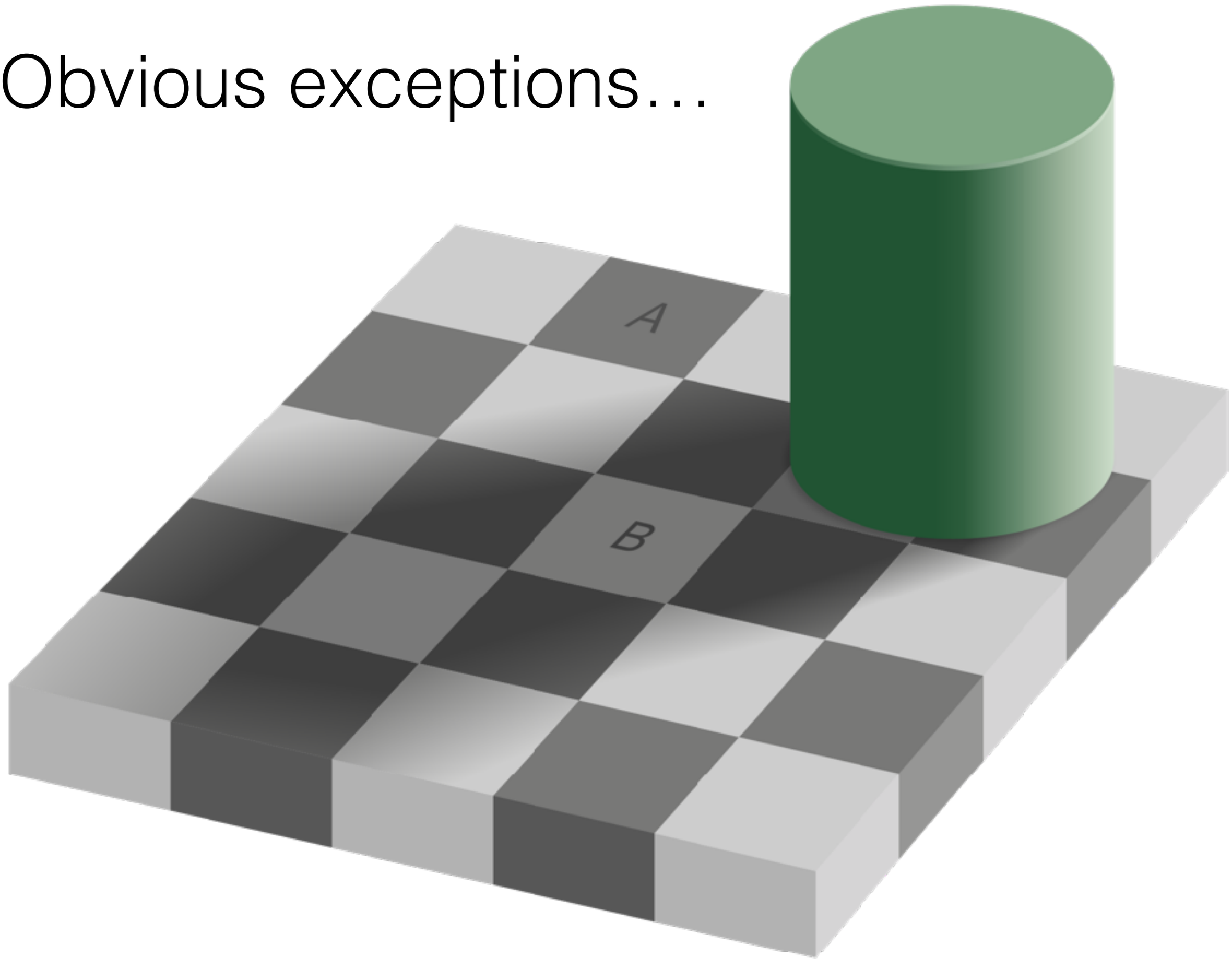


Quantized Color

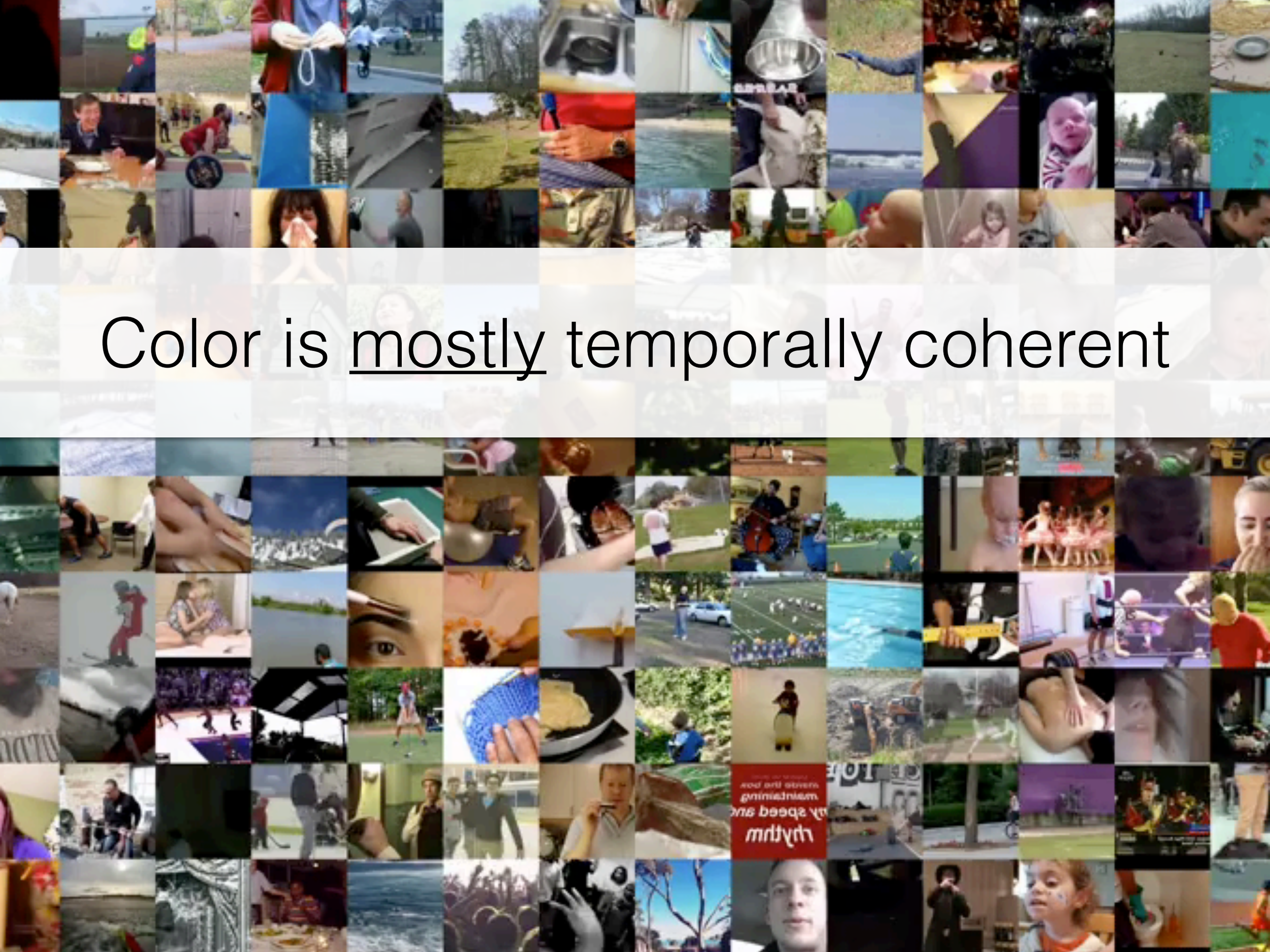


Obvious exceptions…

Obvious exceptions…



Obvious exceptions…



Color is mostly temporally coherent

Self-supervised Tracking

Reference Frame

Gray-scale Video

Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

Tracking underpins motion tasks

Action recognition

Physical reasoning

Future prediction

Summarization

Interaction, imitation

…

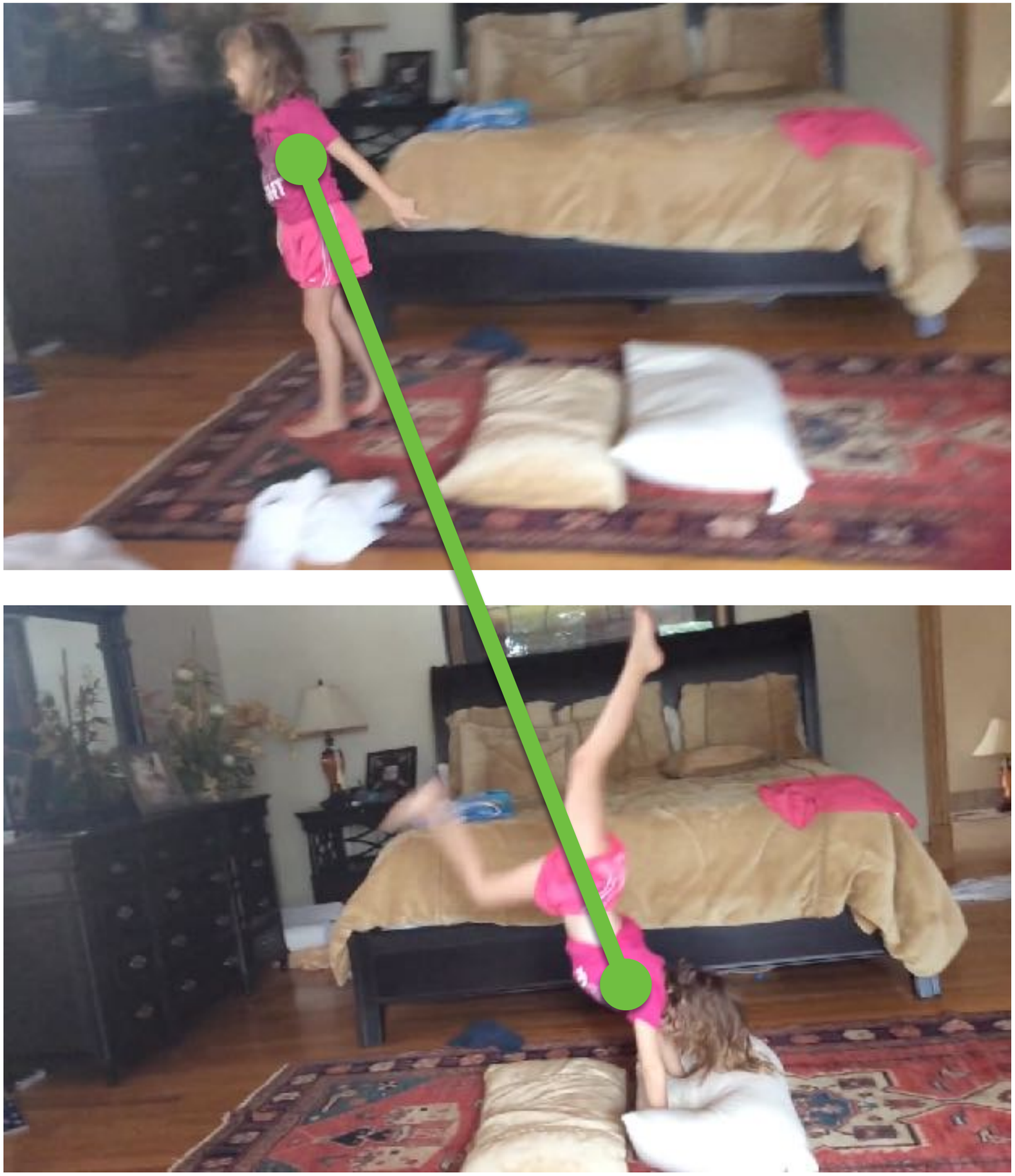
Colorization isn’t out of the blue

- **Ryan Dahl**. Automatic Colorization.
- **Aditya Deshpande, Jason Rock and David Forsyth.** Learning Large-Scale Automatic Image Colorization.
- **Zezhou Cheng, Qingxiong Yang, and Bin Sheng.** Deep Colorization.
- **Richard Zhang, Phillip Isola, Alexei A. Efros.** Colorful Image Colorization
- **Gustav Larsson, Michael Maire, and Gregory Shakhnarovich**. Learning Representations for Automatic Colorization.
- **Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa.** Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification.

What color is this?

Where to copy color?

Want to be safe!



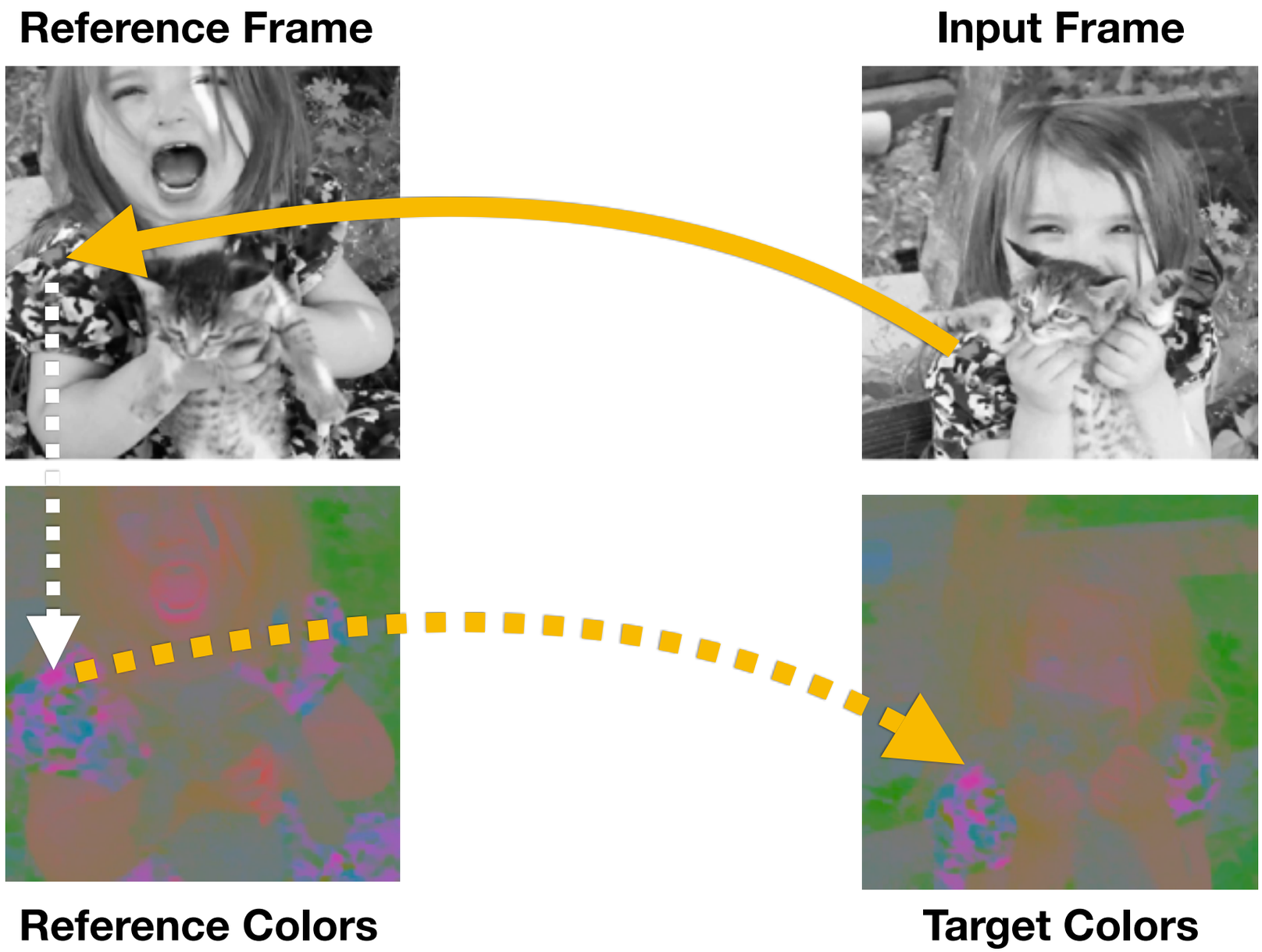
Where to copy color?

Color can be robust to occlusion

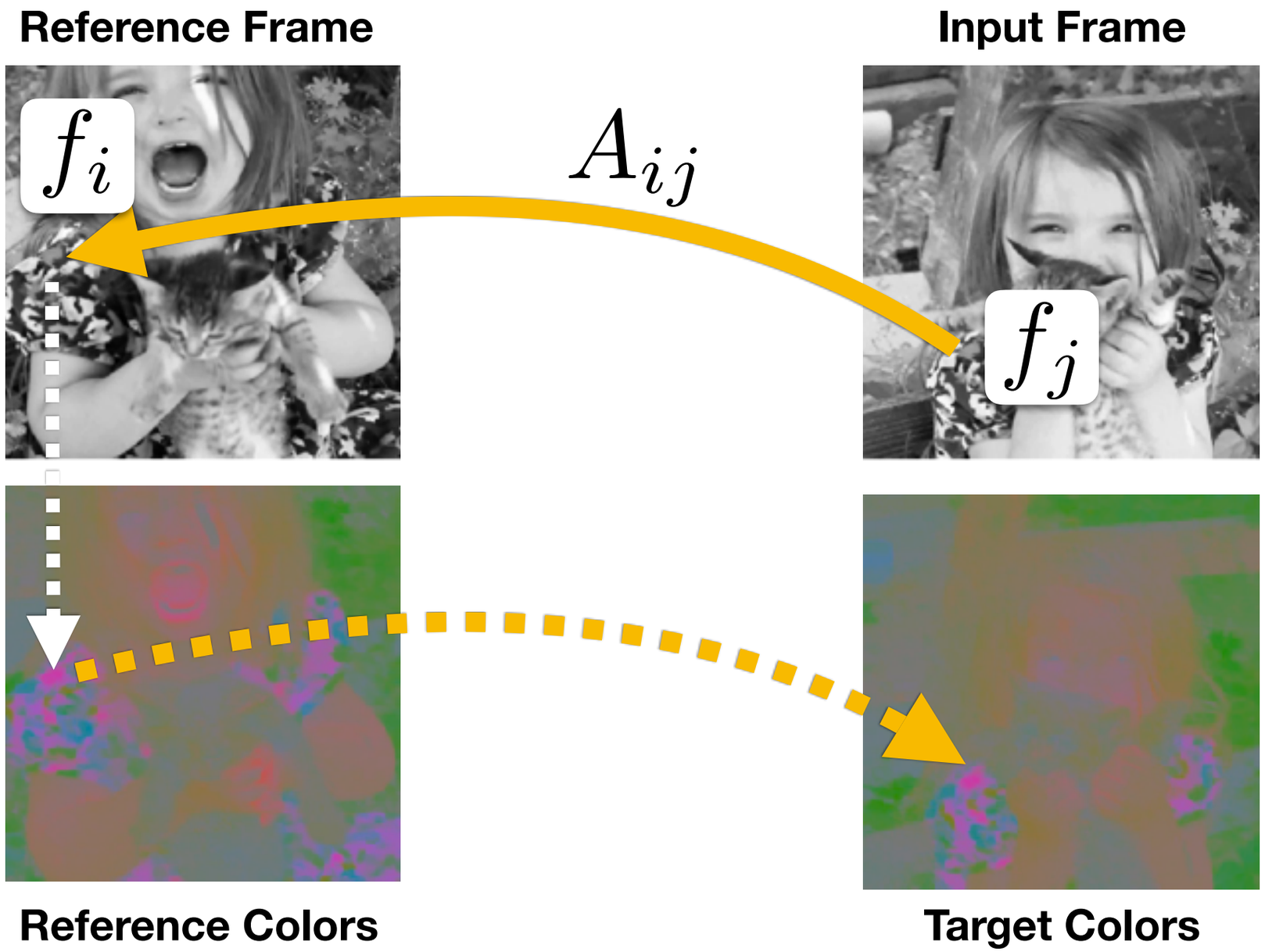
Input Frame

Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

Colorize by Pointing

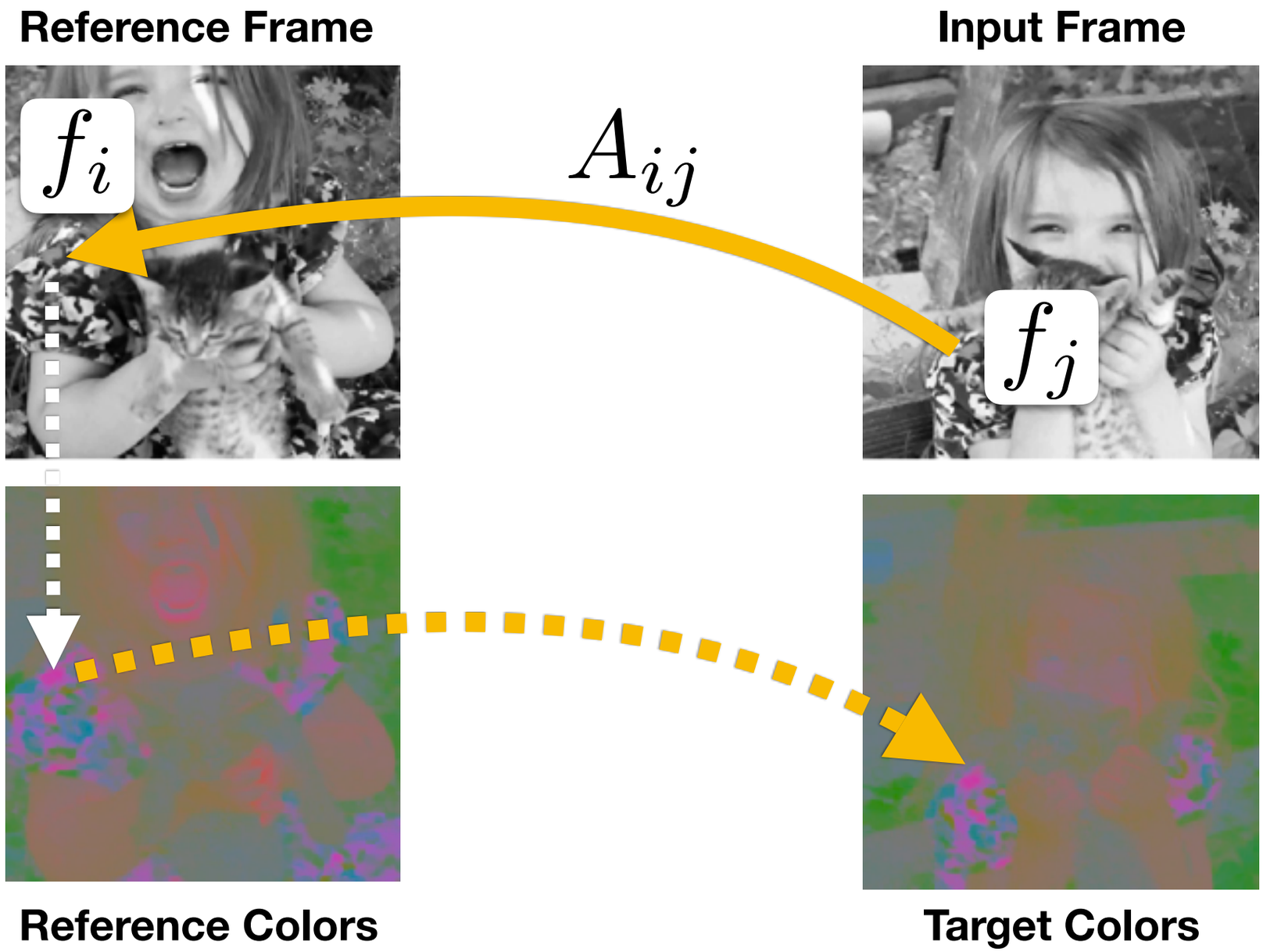


Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.



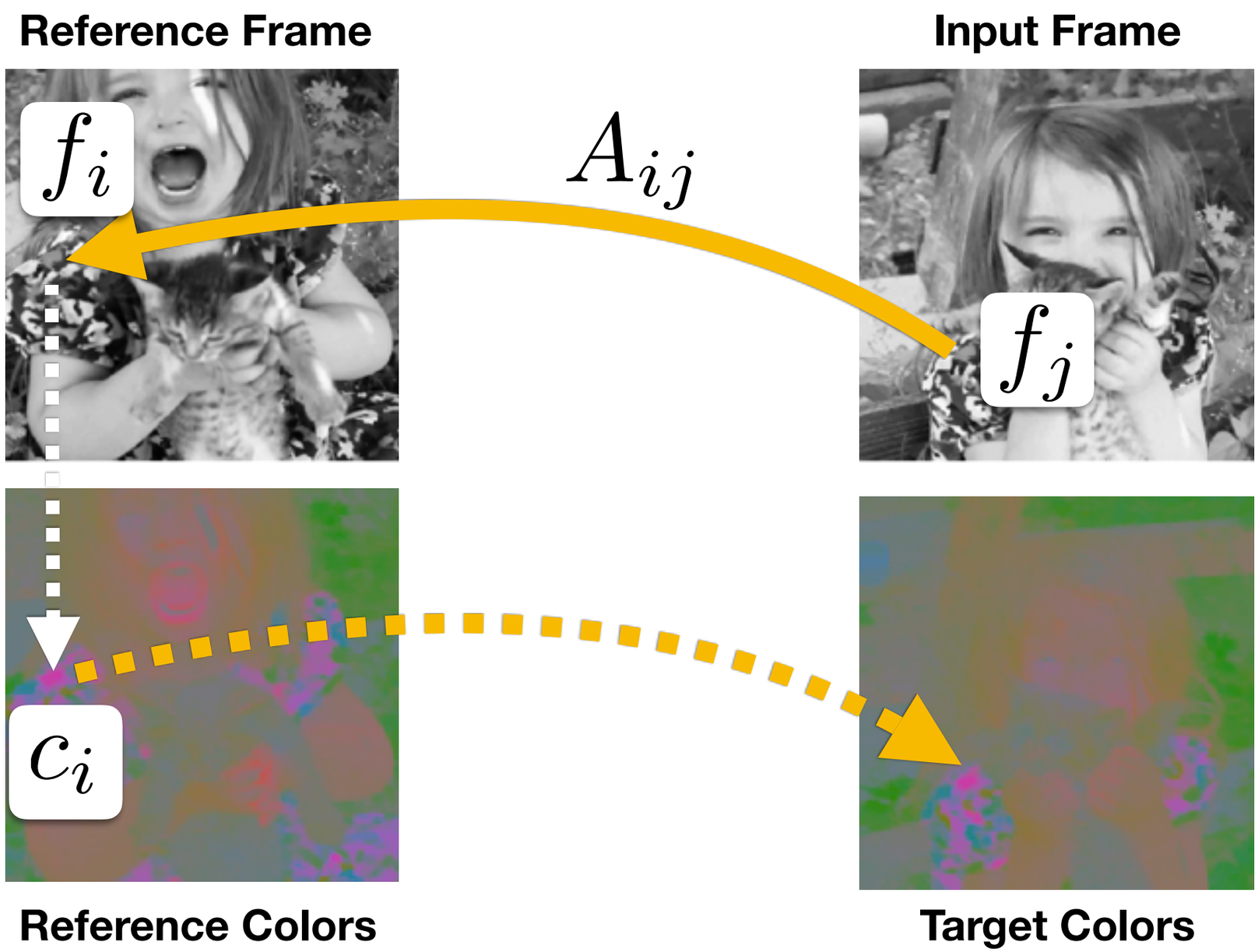
Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

$$A_{ij} = \frac{\exp\left(f_i^T f_j\right)}{\sum_k \exp\left(f_k^T f_j\right)}$$



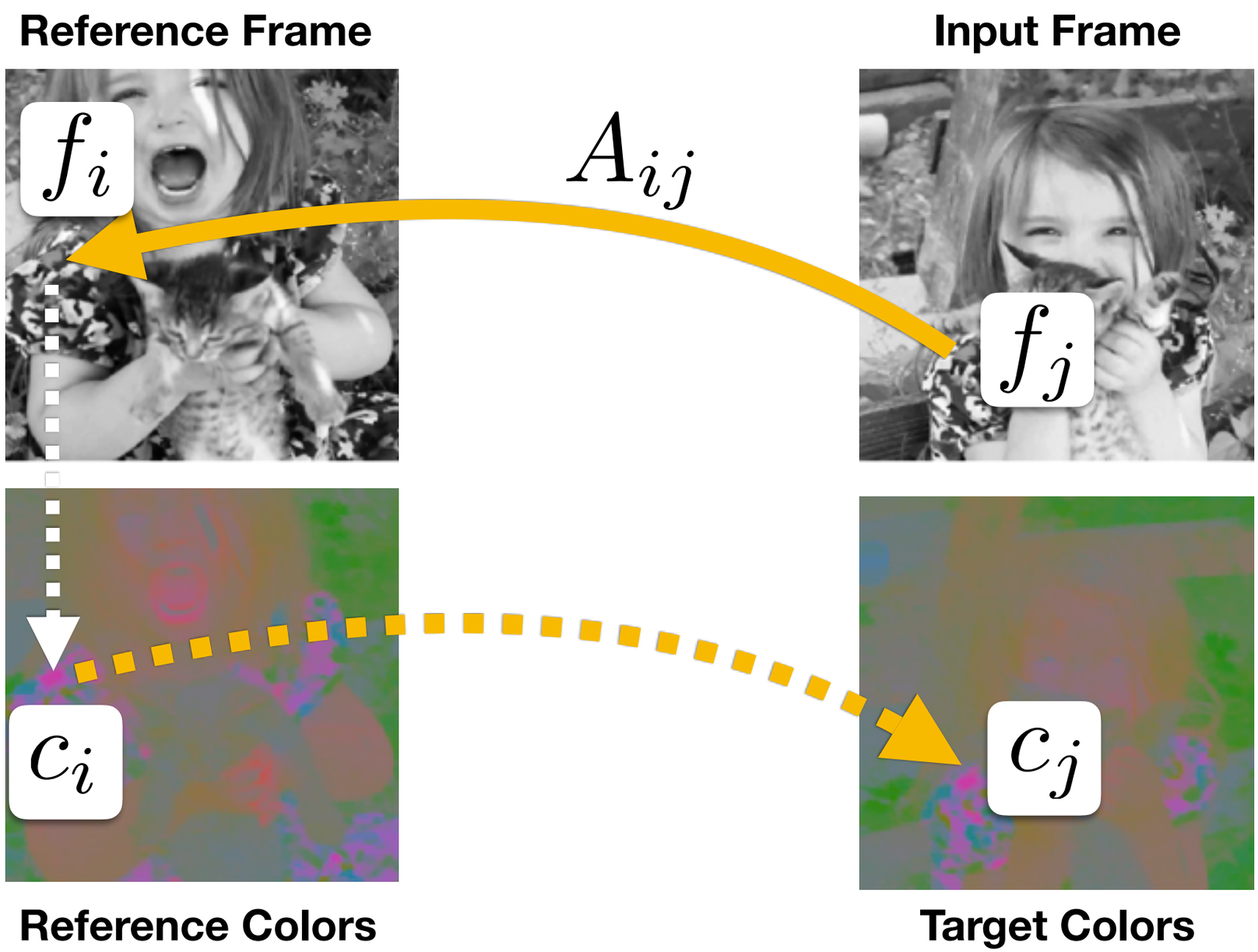
Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

$$\hat{c}_j = \sum_i A_{ij} c_i \qquad \text{where } A_{ij} = \frac{\exp\left(f_i^T f_j\right)}{\sum_k \exp\left(f_k^T f_j\right)}$$



Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

$$\min_f \ \mathcal{L}\left(c_j, \sum_i A_{ij} c_i\right) \ \text{where } A_{ij} = \frac{\exp\left(f_i^T f_j\right)}{\sum_k \exp\left(f_k^T f_j\right)}$$



Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

Video Colorization

Reference Frame

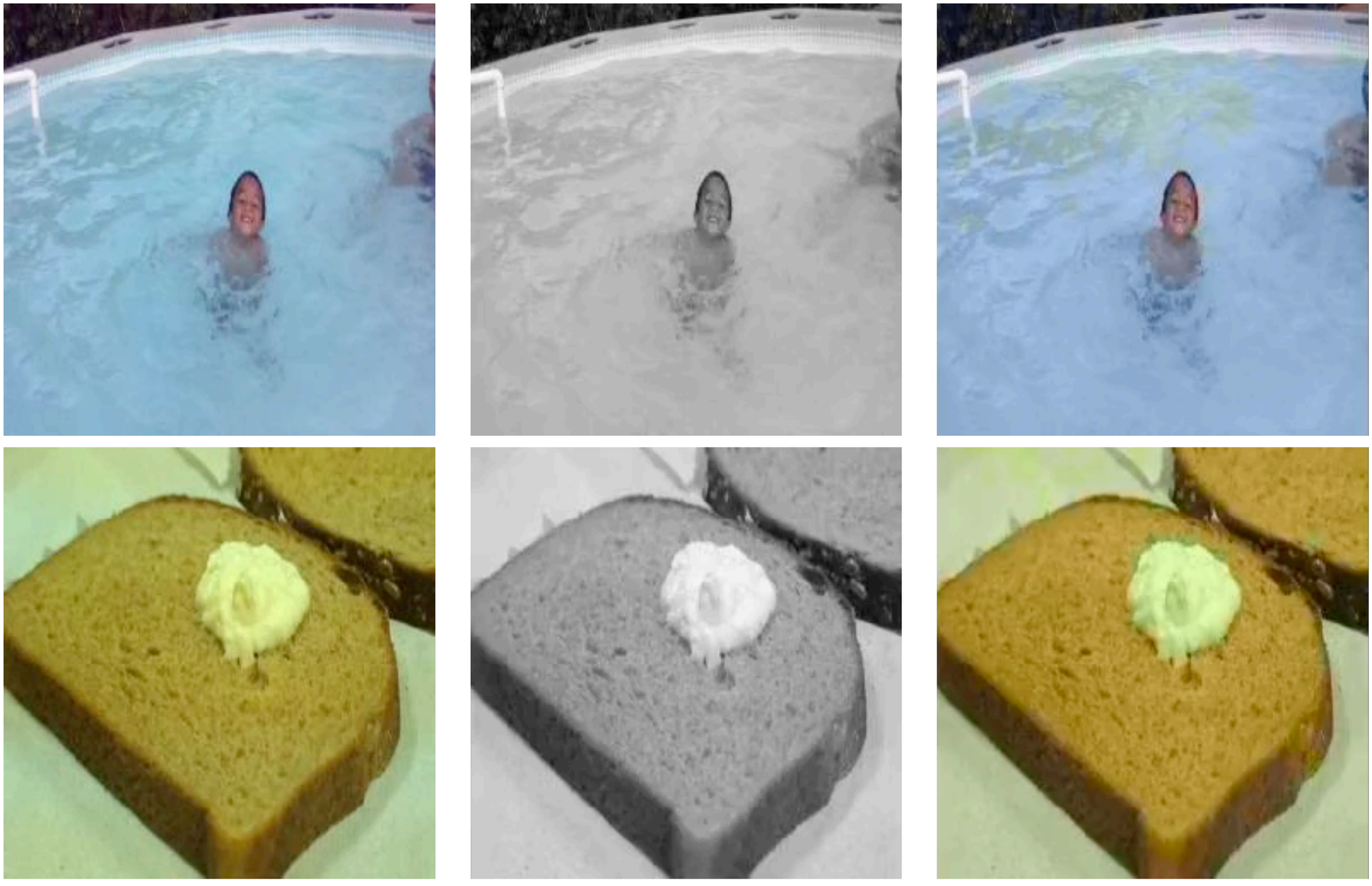
Gray-scale Video

Predicted Color

Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

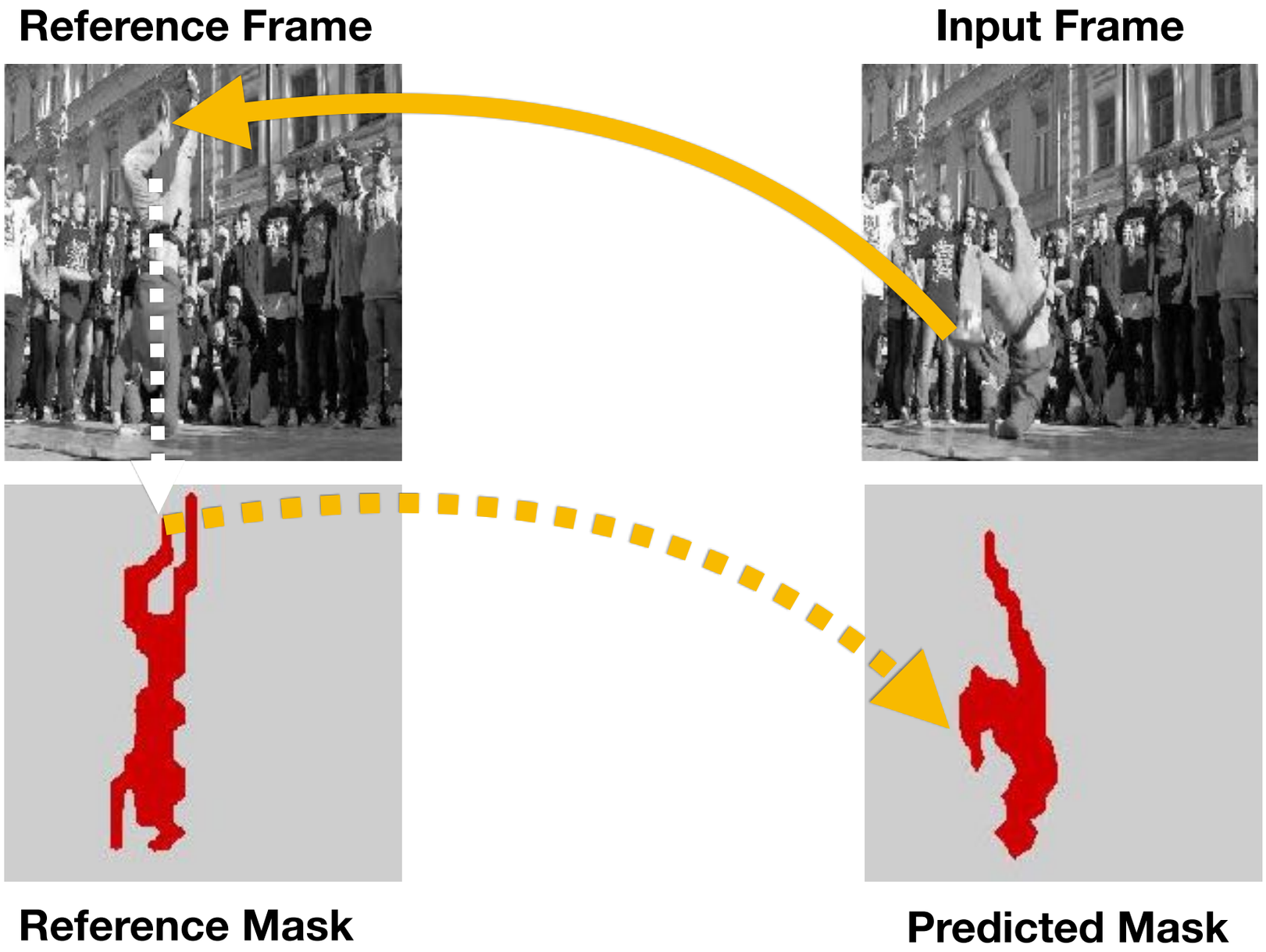
Video Colorization

Reference Frame | Gray-scale Video | Predicted Color



Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

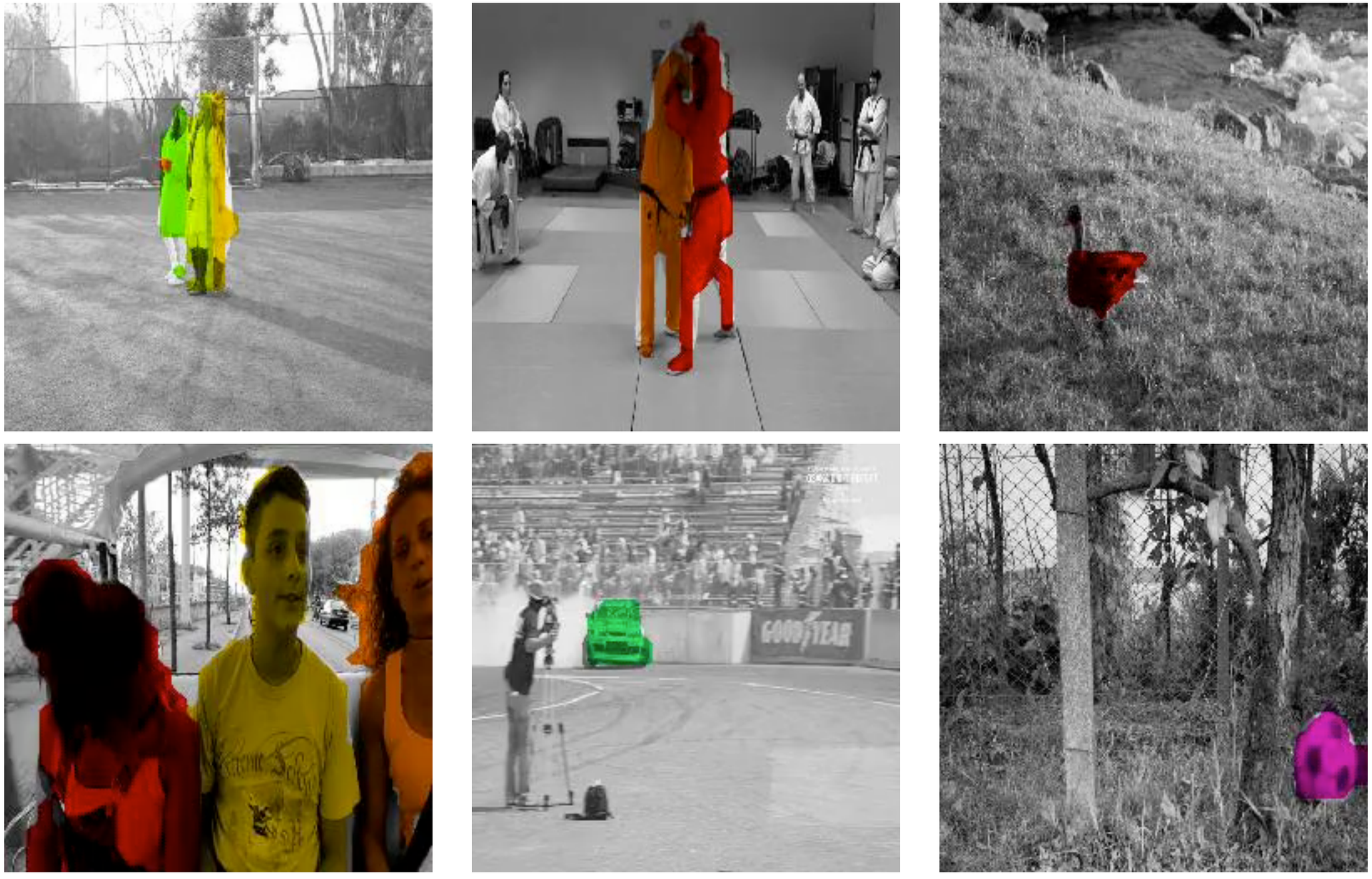
Tracking Emerges!



Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

Segment Tracking Results

Only the first frame is given. Colors indicate different instances.



Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

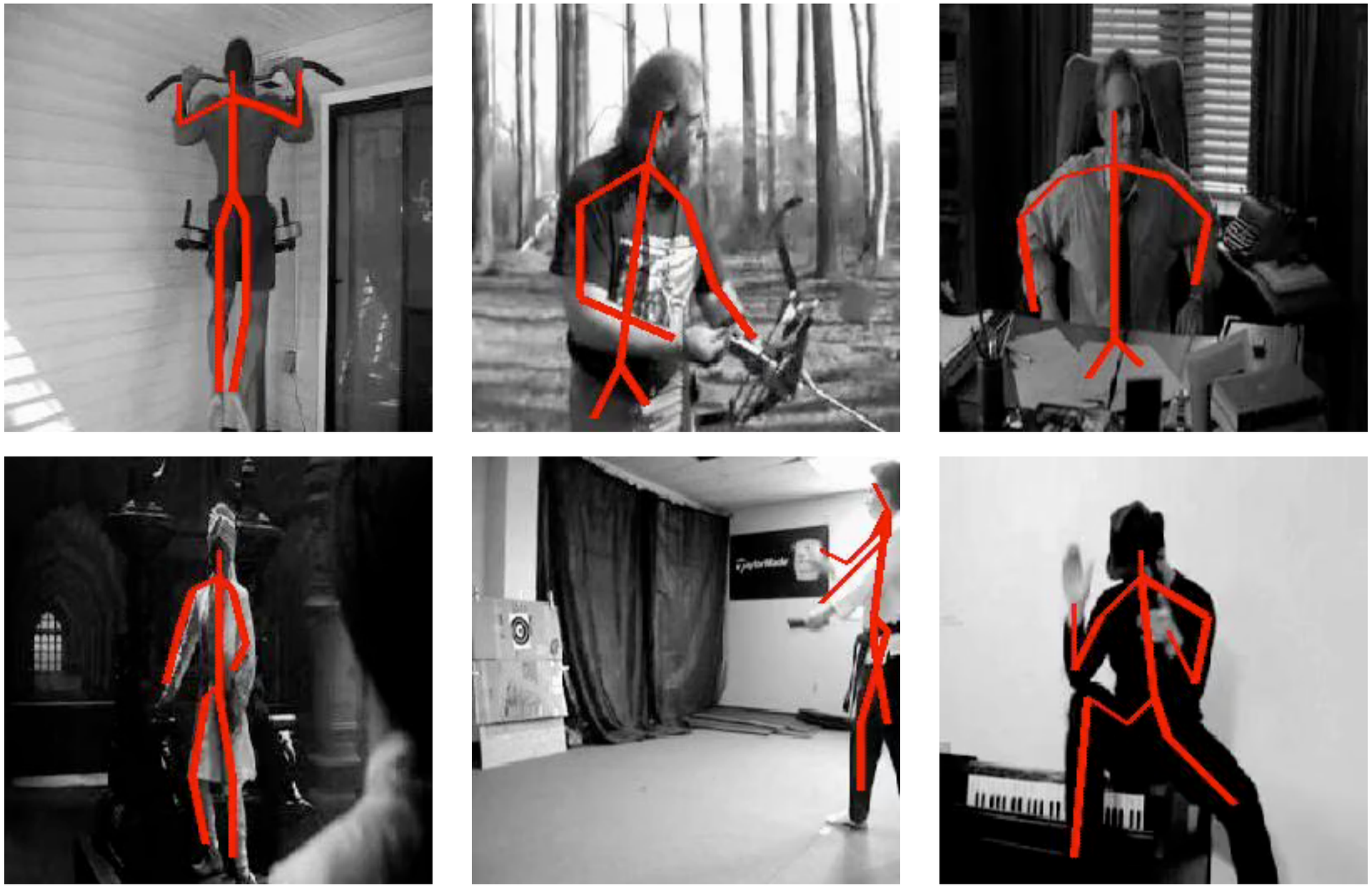
Segment Tracking Results

Only the first frame is given. Colors indicate different instances.

Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

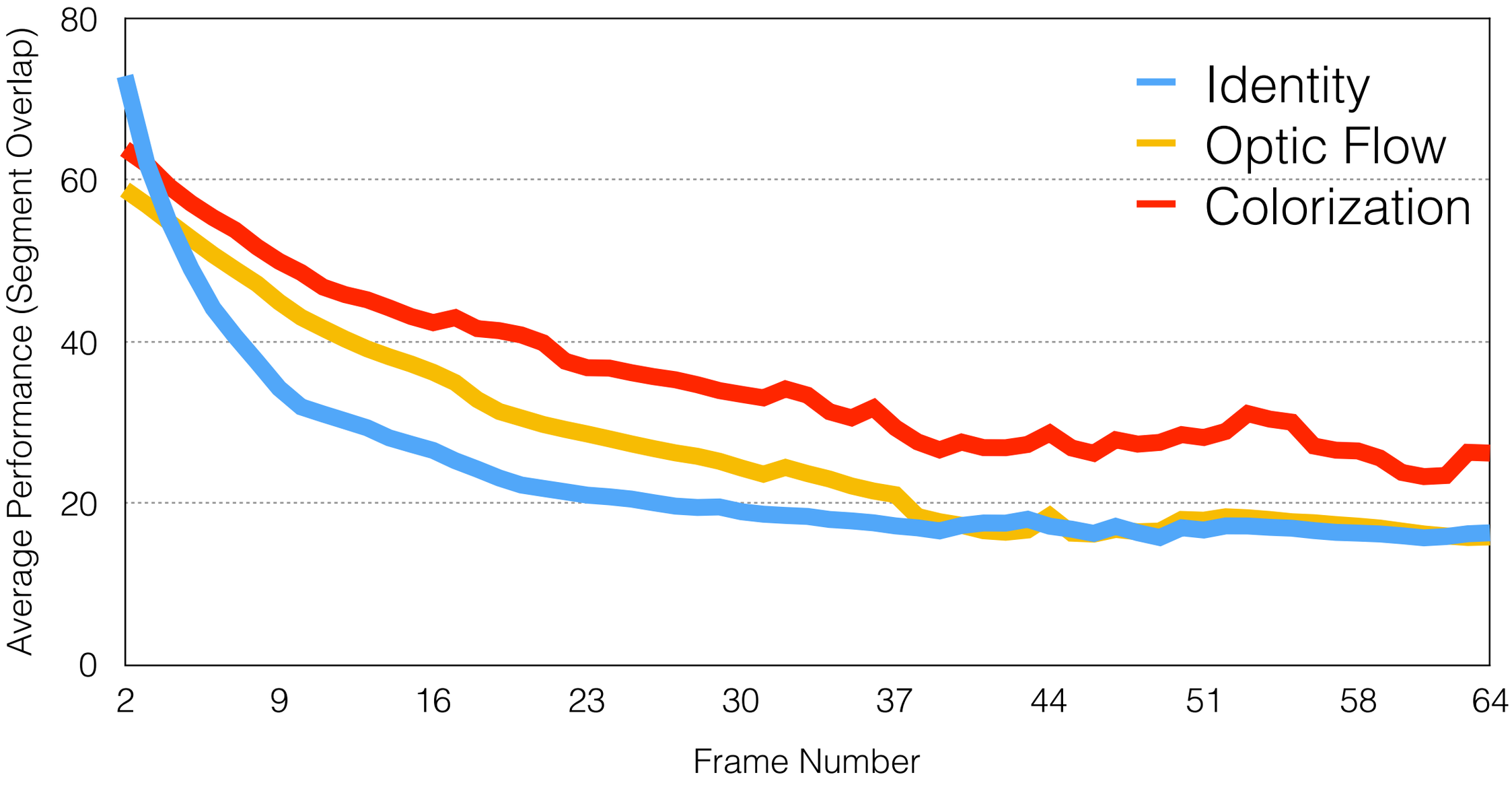
Pose Tracking Results

Only the skeleton in the first frame is given.



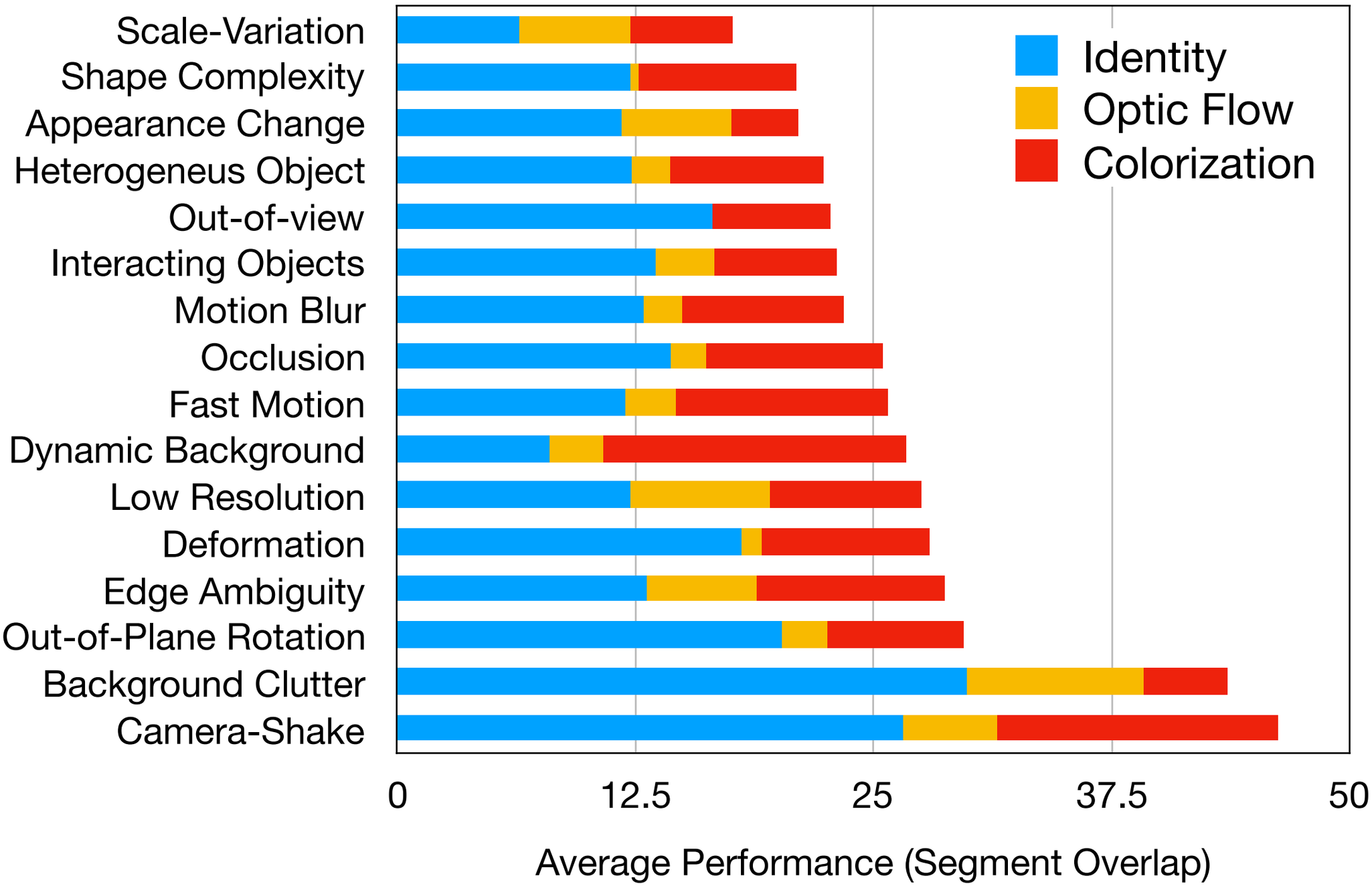
Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

Tracking Performance



Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

Tracking Performance



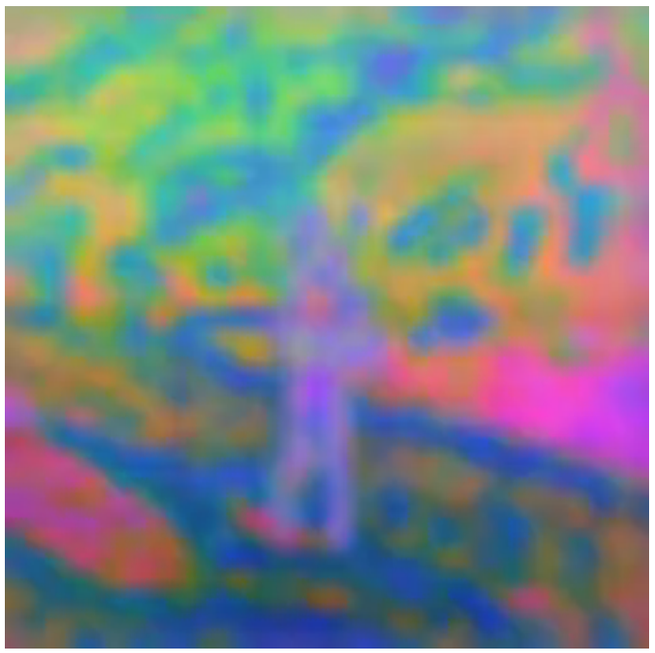
Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

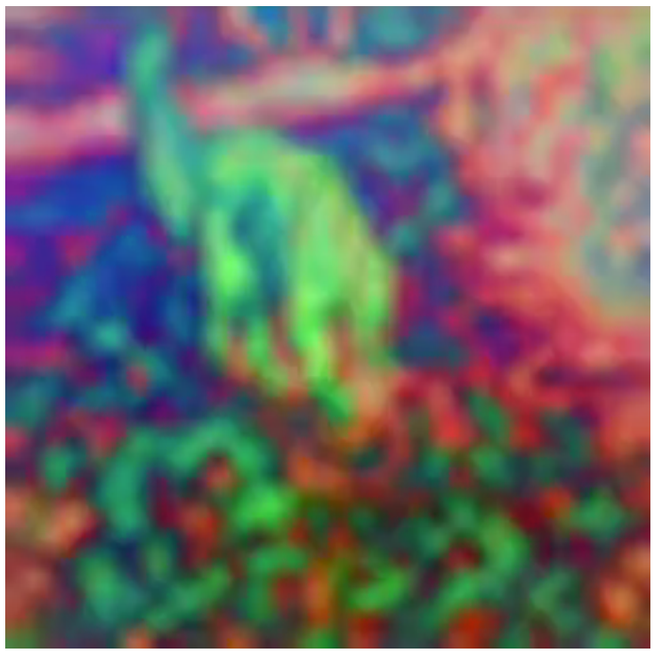
Visualizing Embeddings

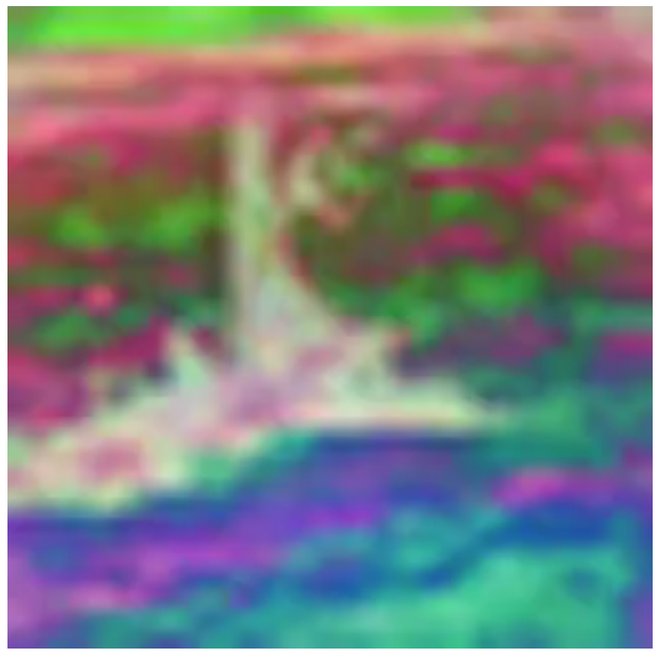
Project embedding to 3 dimensions and visualize as RGB

Original Video

Embedding Visualization







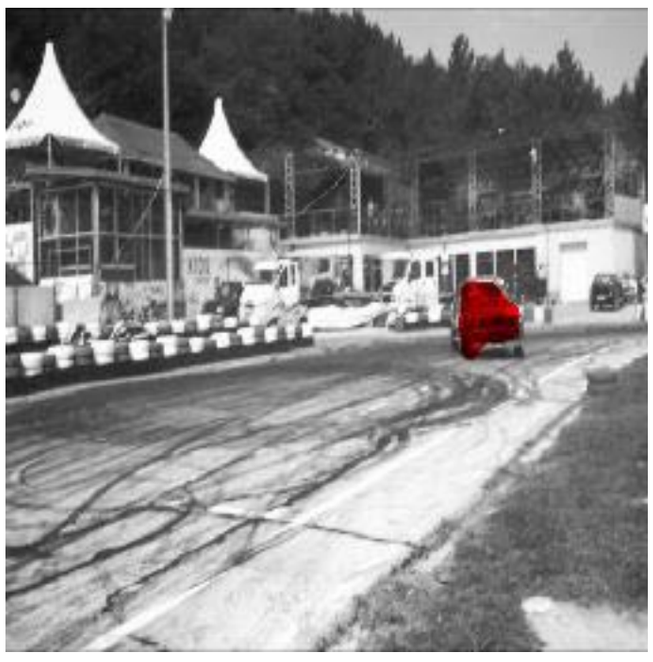
Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

When does it fail?

Reference Colors

Predicted Colors

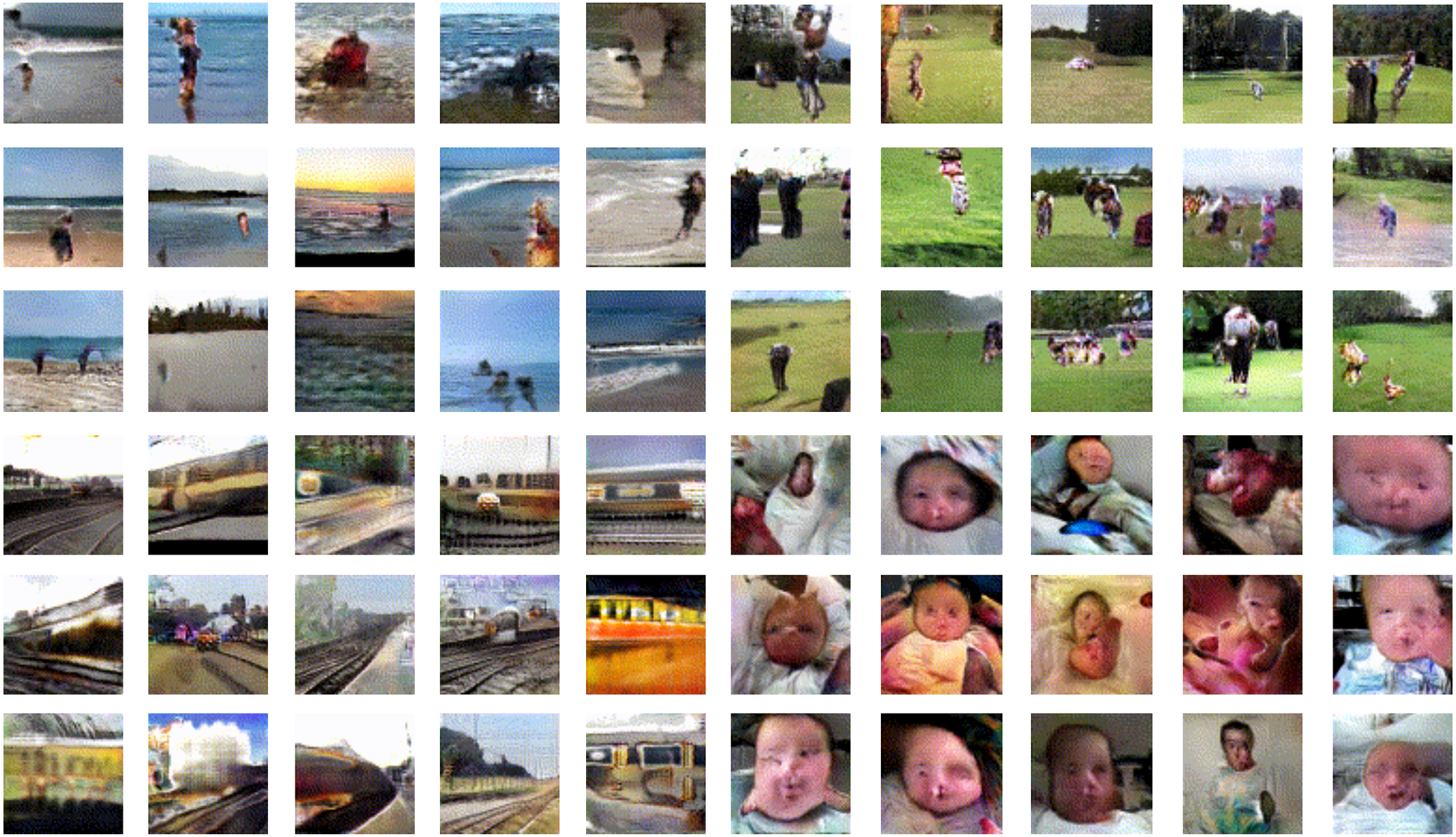
Reference Mask



Predicted Mask

Vondrick, Shrivastava, Fathi, Guadarrama, Murphy. In submission.

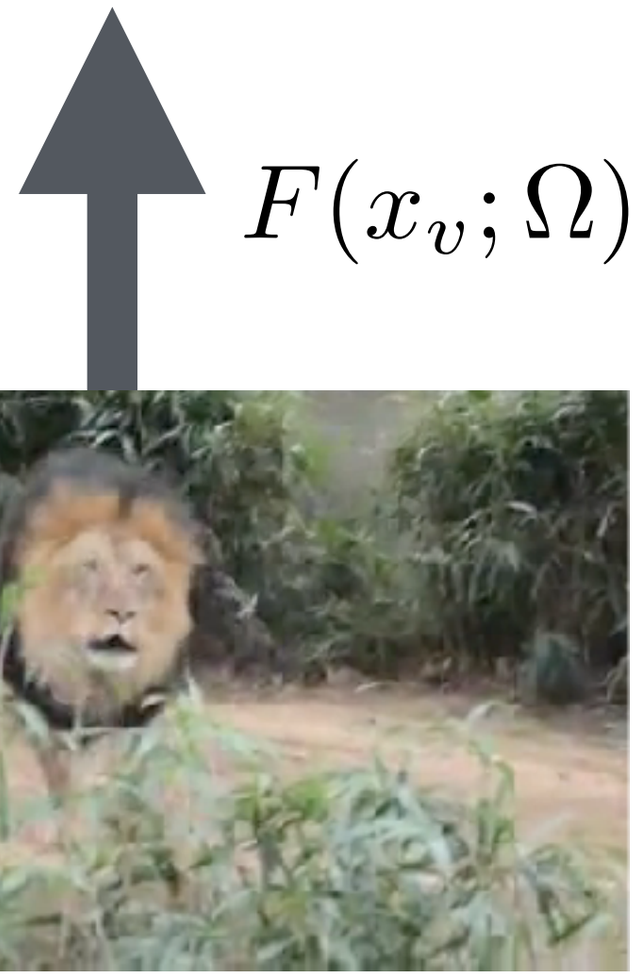
Generating Videos with GANs



Vondrick, Pirsiavash, Torralba. NIPS 2016

Learning from unlabeled video

Lion



Vision

Aytar, Vondrick, Torralba. NIPS 2016

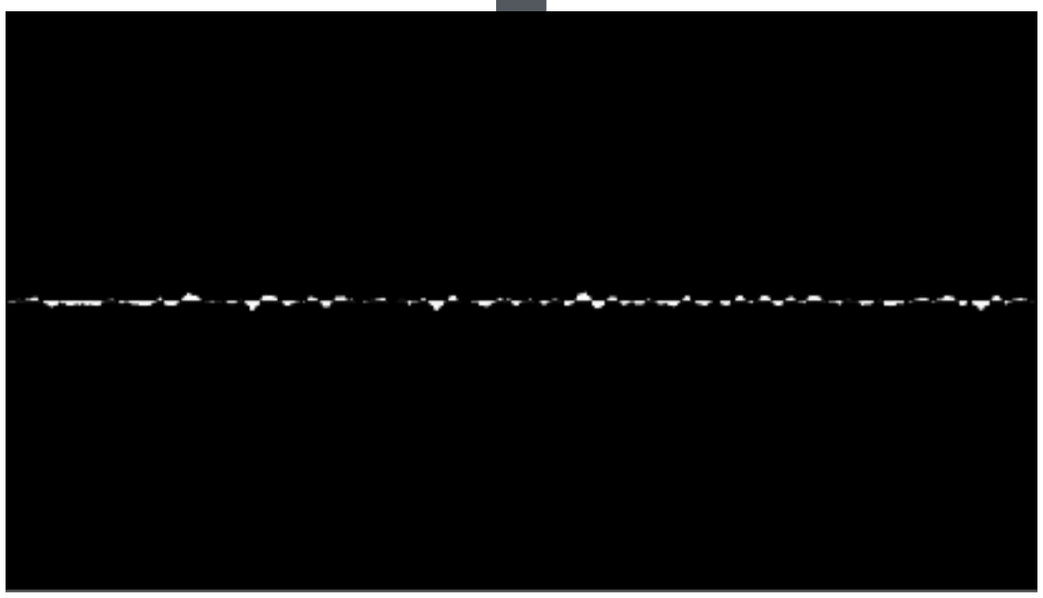
Learn to Hear by Seeing

$$\min_f \sum_i D_{\mathrm{KL}}\left(F(x_i)||f(x_i)\right)$$

Lion

$f(x_s;\omega)$

$F(x_v;\Omega)$



Sound

Vision

Aytar, Vondrick, Torralba. NIPS 2016

pasture: 18.99%

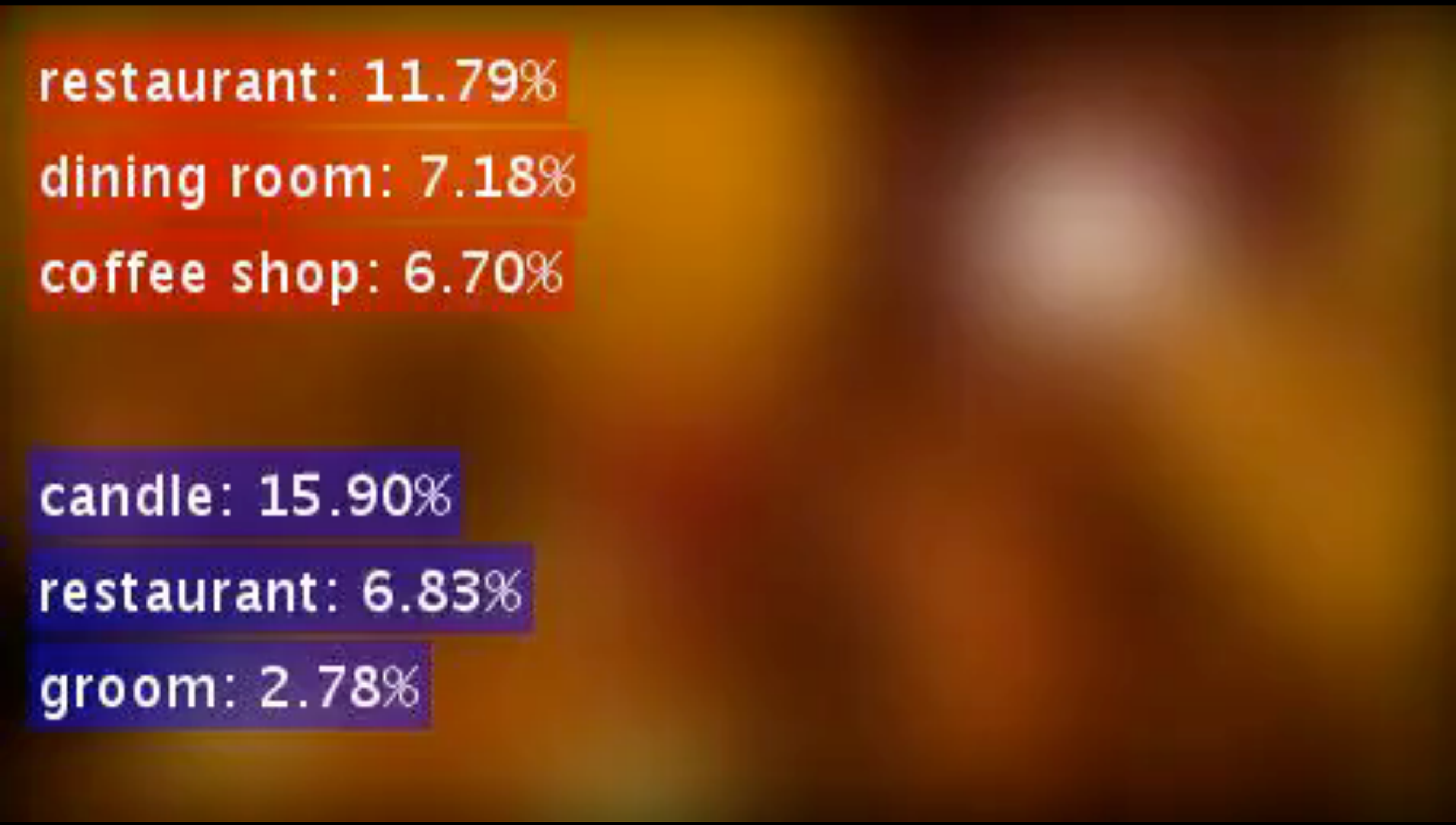
yard: 9.89%

lawn: 7.39%

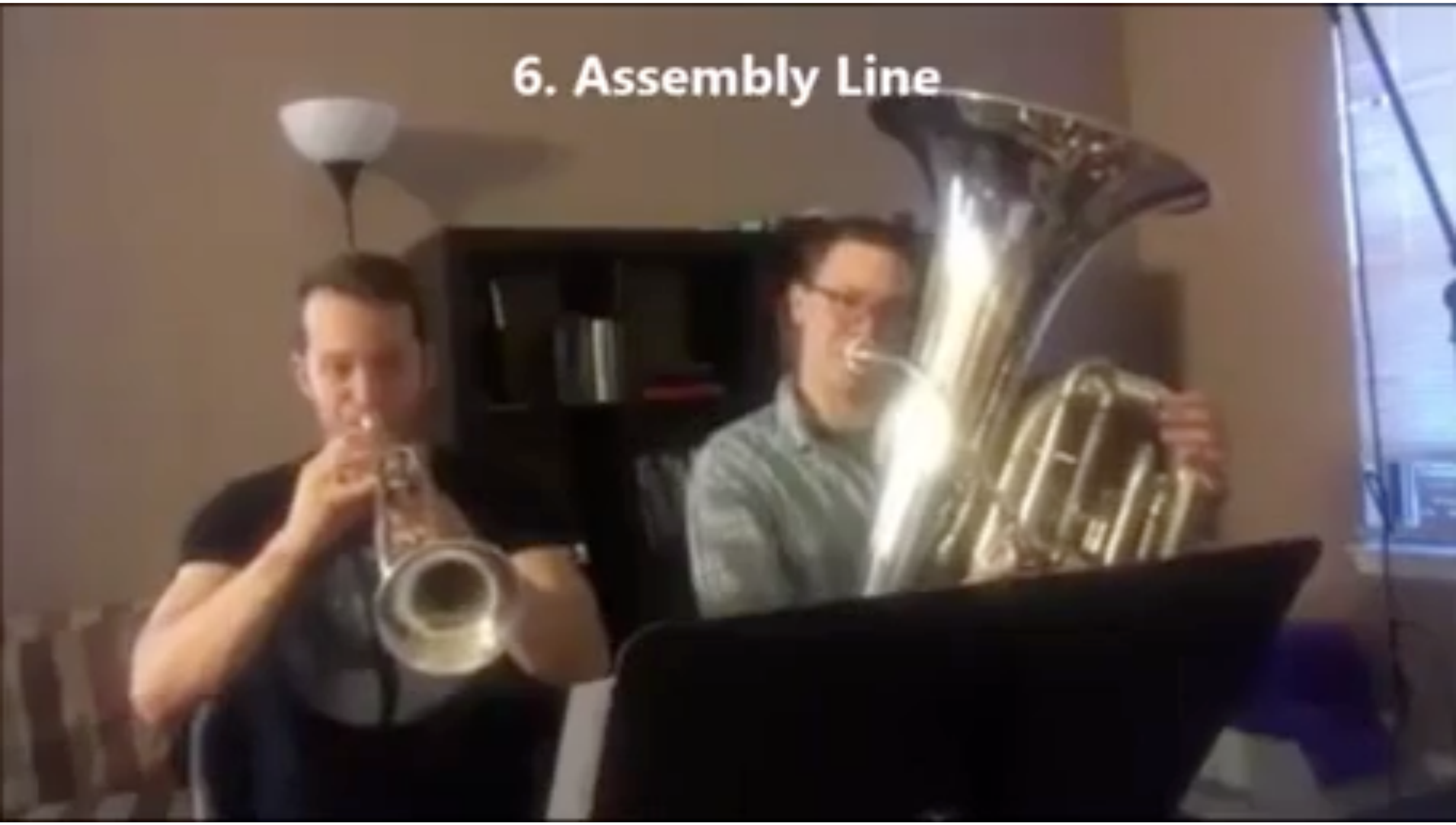
English springer: 8.65%

Welsh springer spaniel: 2.20%

Border collie: 1.65%

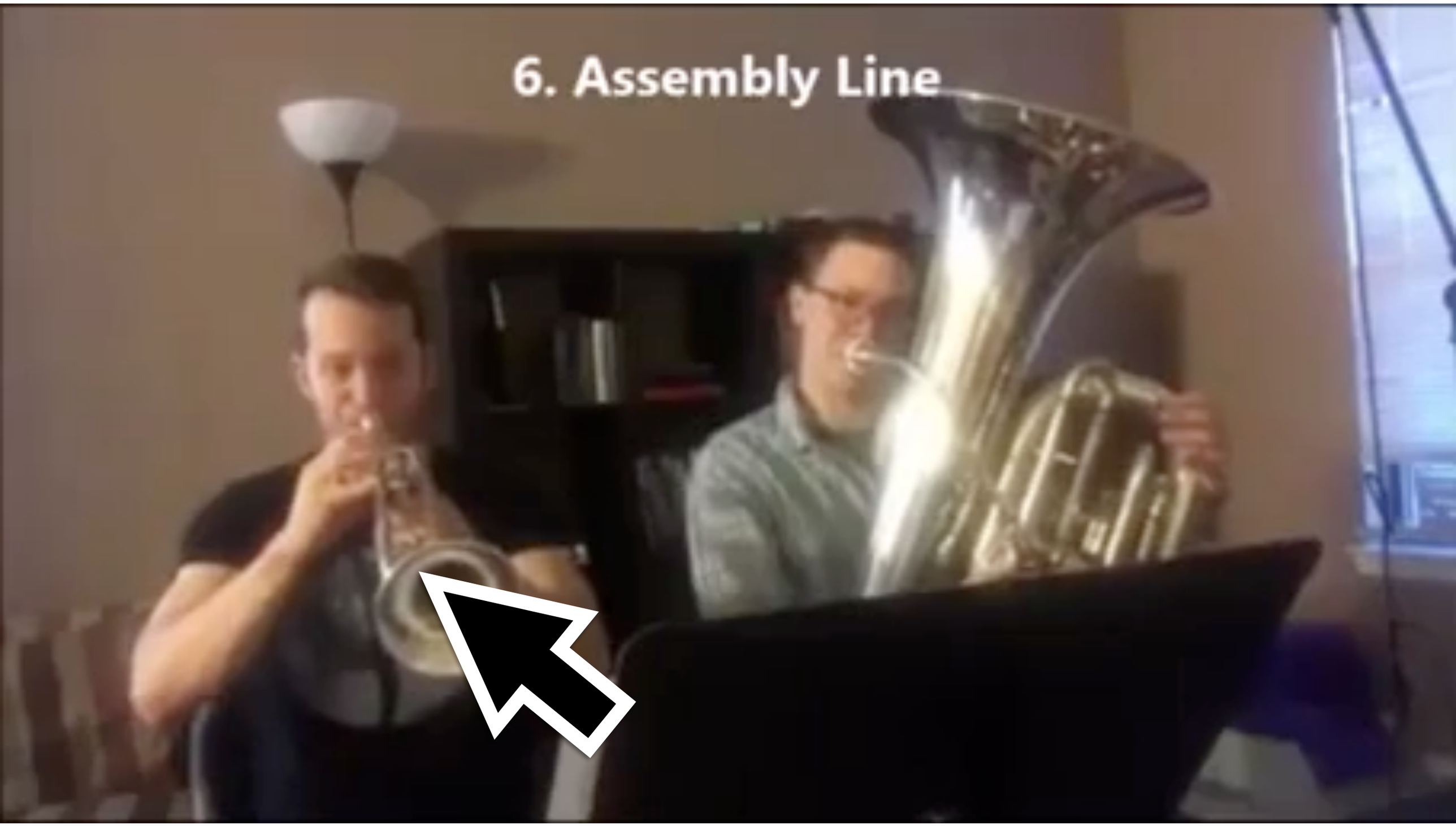

restaurant: 11.79%
dining room: 7.18%
coffee shop: 6.70%
candle: 15.90%
restaurant: 6.83%
groom: 2.78%

Which objects make which sounds?



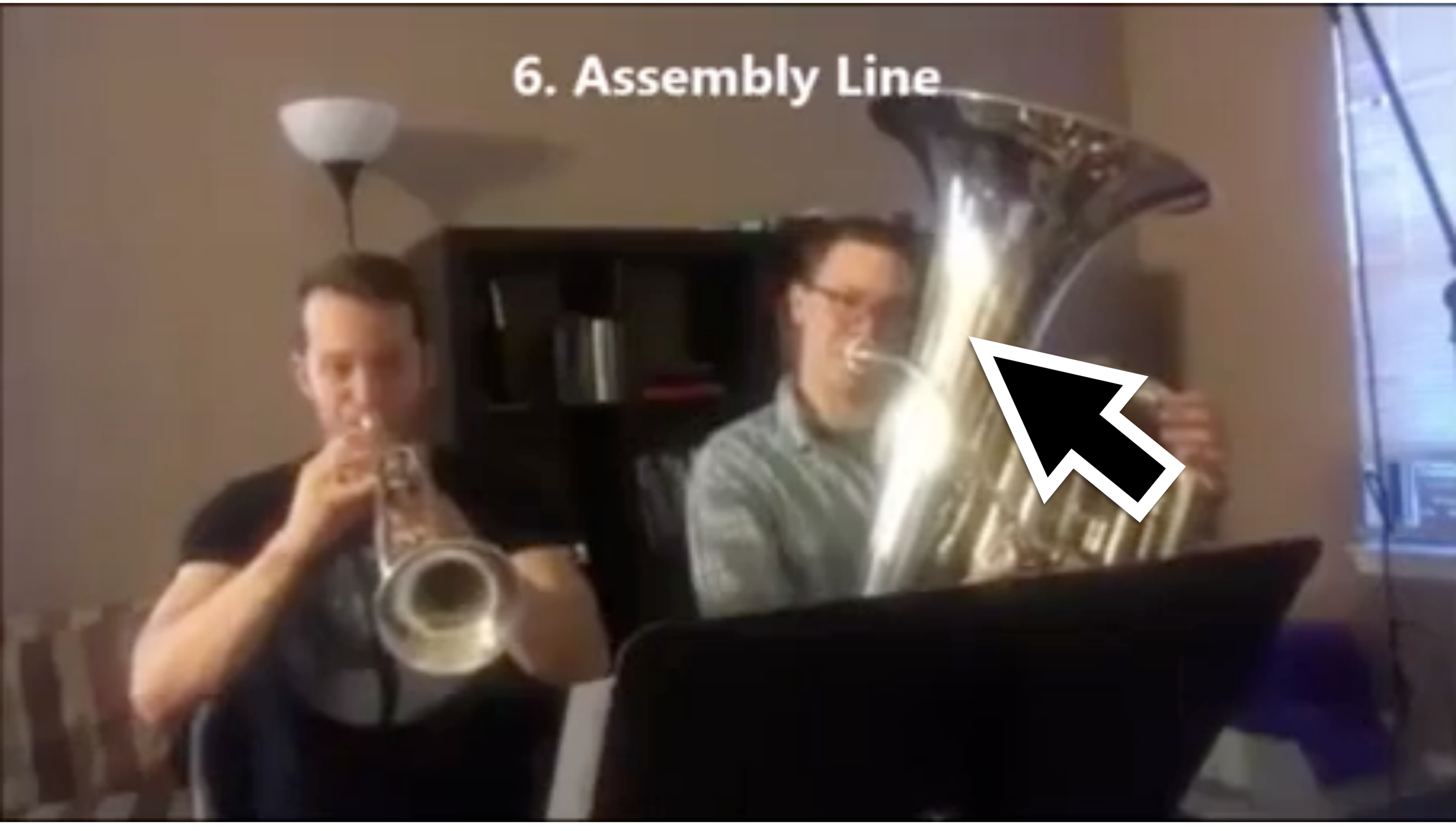
Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

The sound of clicked object



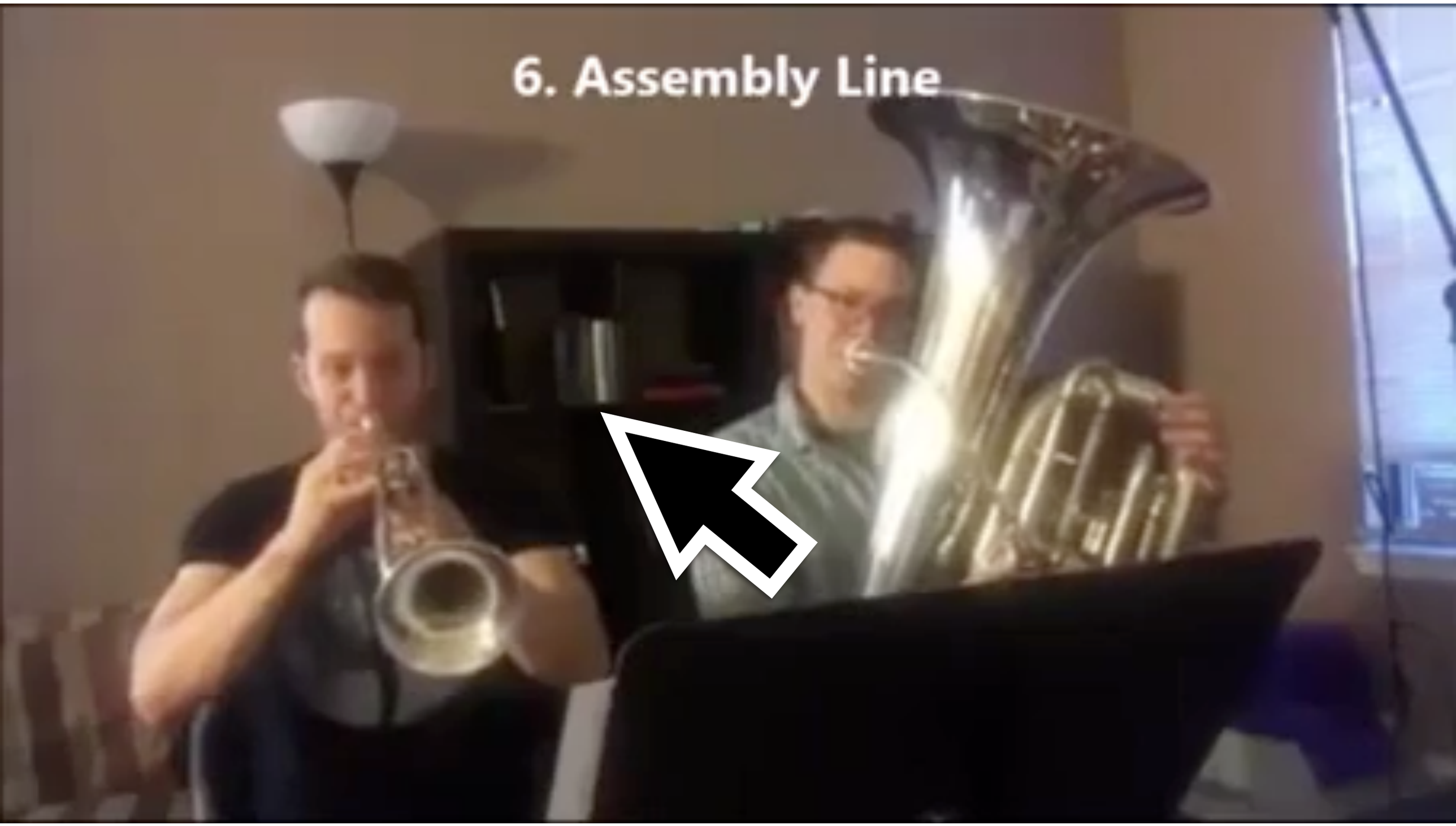
Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

The sound of clicked object



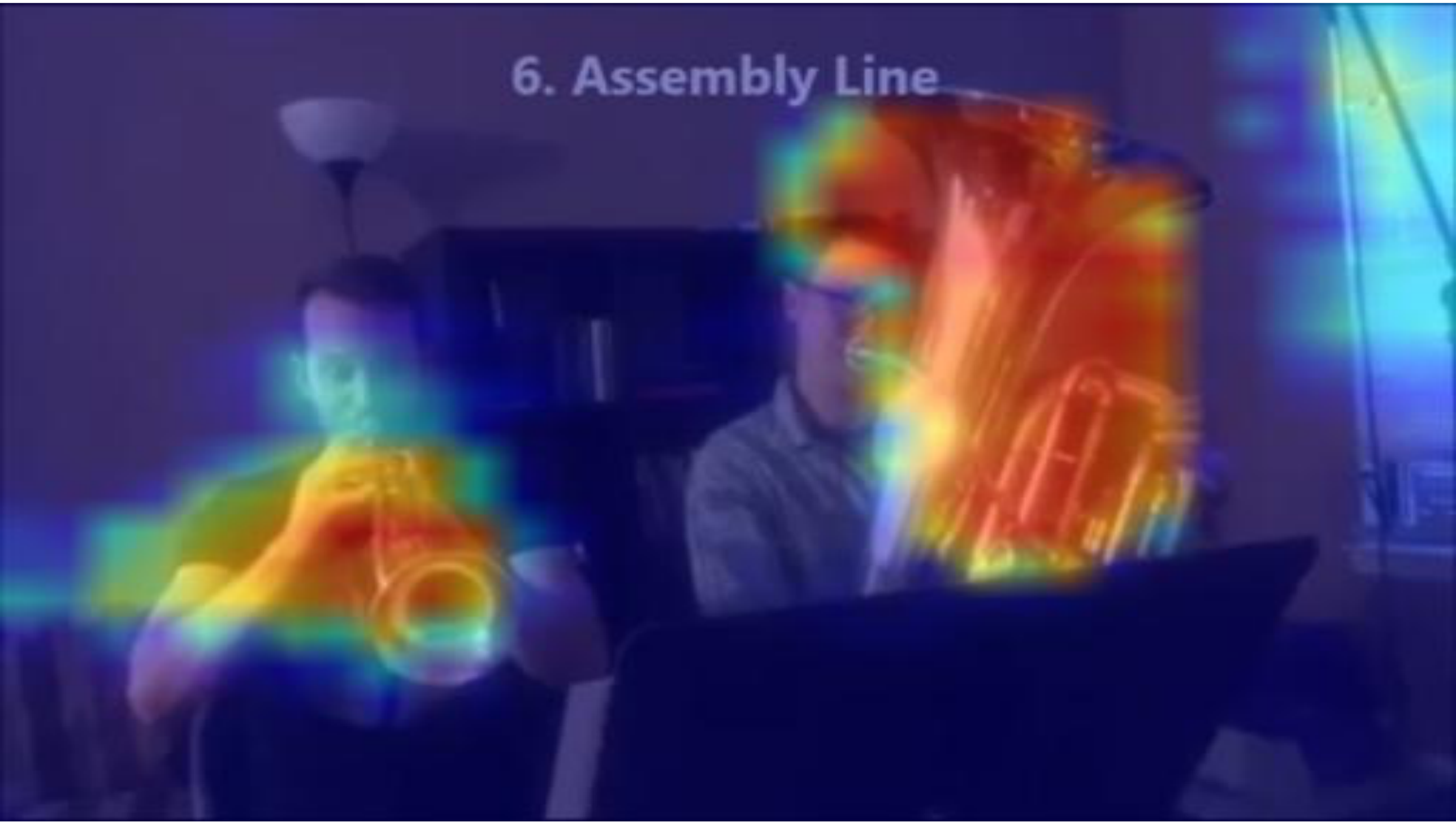
Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

The sound of clicked object



Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

Predicted sound volume

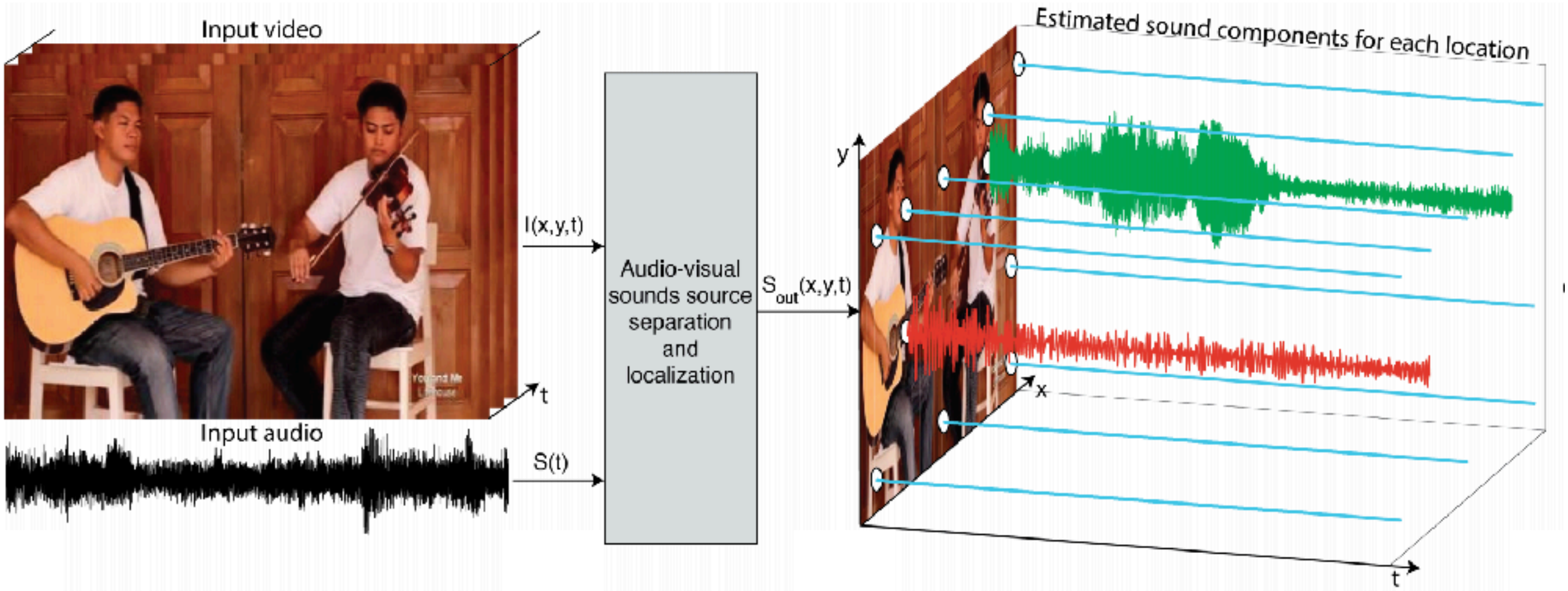


Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

Predicted sound clusters

Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

Audiovisual Grounding



Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba.

Sounds like a good idea

- **Andrew Owens, Alexei A. Efros.** Audio-Visual Scene Analysis with Self-Supervised Multisensory Features

- **Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, In So Kweon.** Learning to Localize Sound Source in Visual Scenes

- **Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, Michael Rubinstein.** Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation

- **Relja Arandjelovic, Andrew Zisserman.** Objects that Sound

- **Ruohan Gao, Rogerio Feris, Kristen Grauman.** Learning to Separate Object Sounds by Watching Unlabeled Video

- **Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman.** The Conversation: Deep Audio-Visual Speech Enhancement

Collect unlabeled videos

Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba.

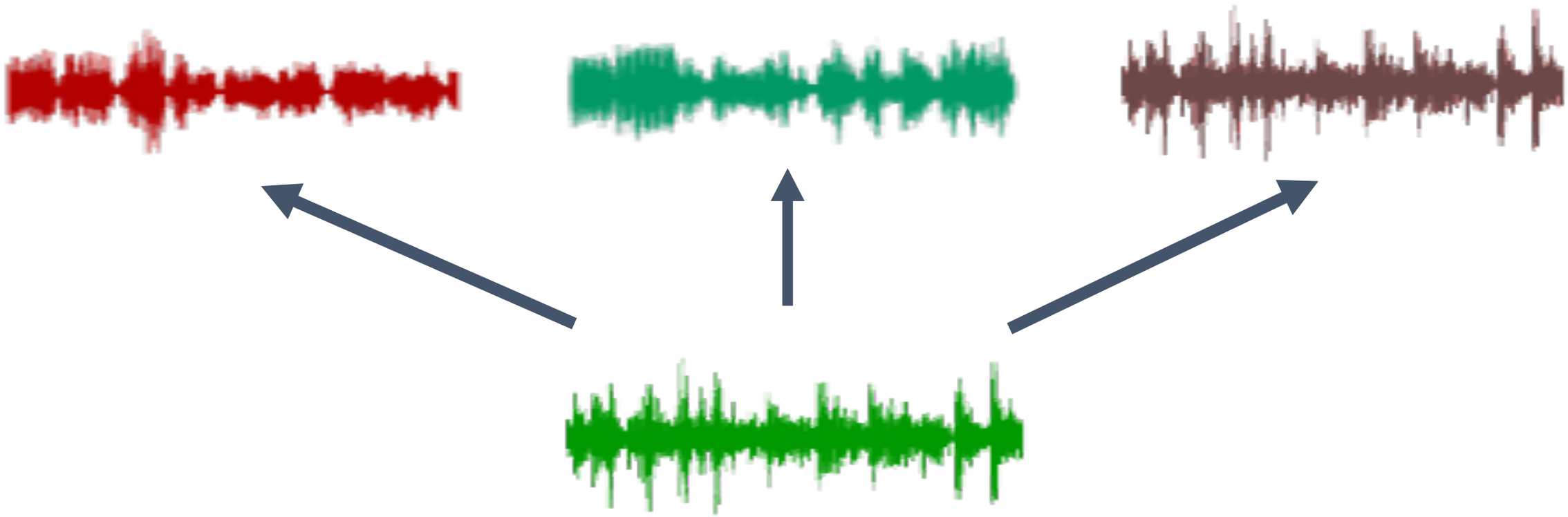
Mix Sound Tracks

Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba.

How to recover originals?

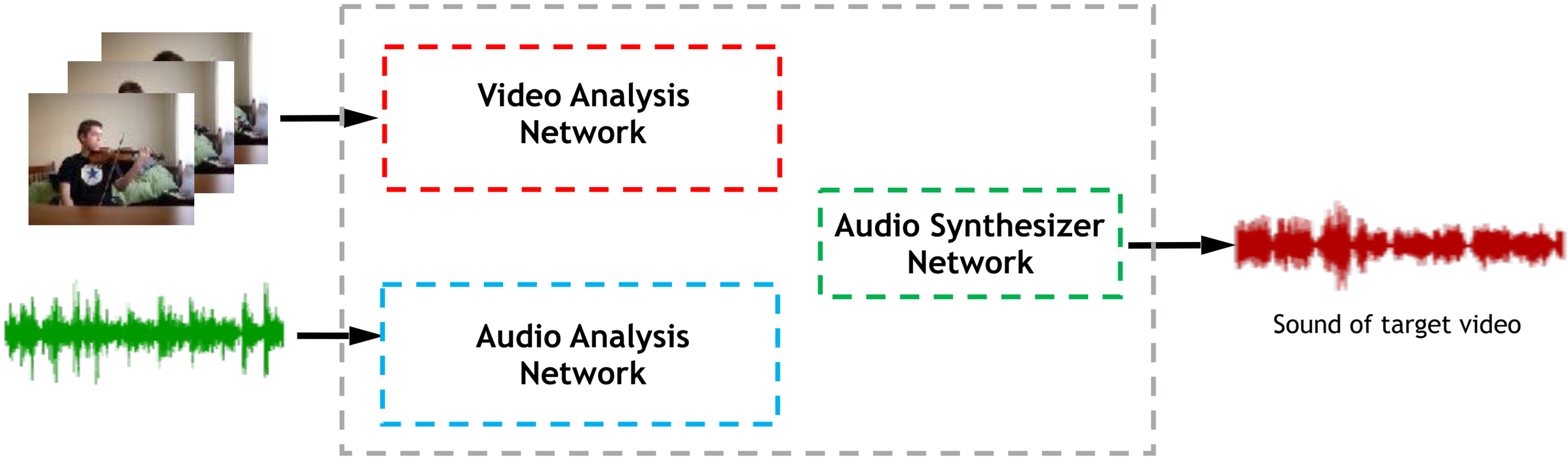
Audio-only:

- Ill-posed
- permutation problem



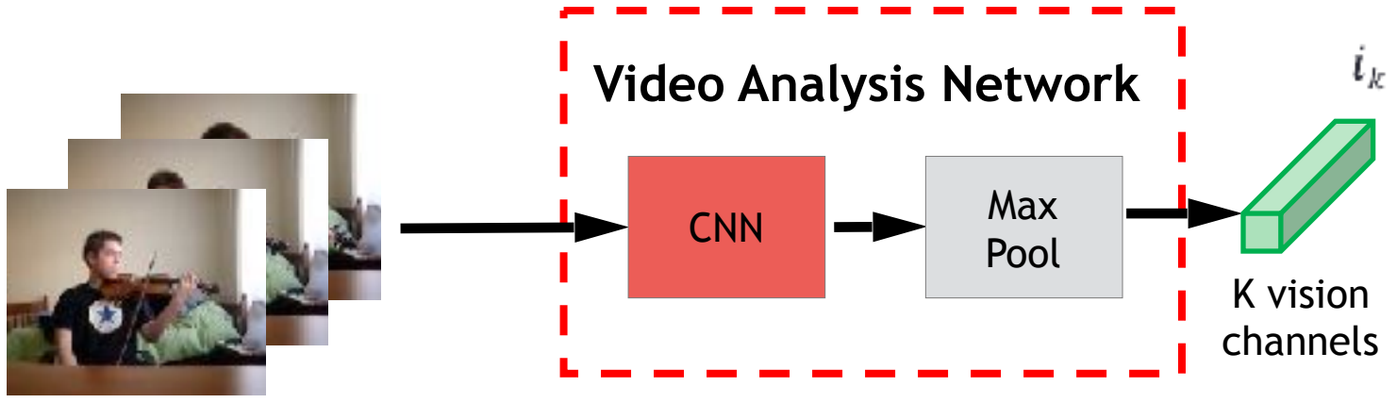
Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba.

Vision can help



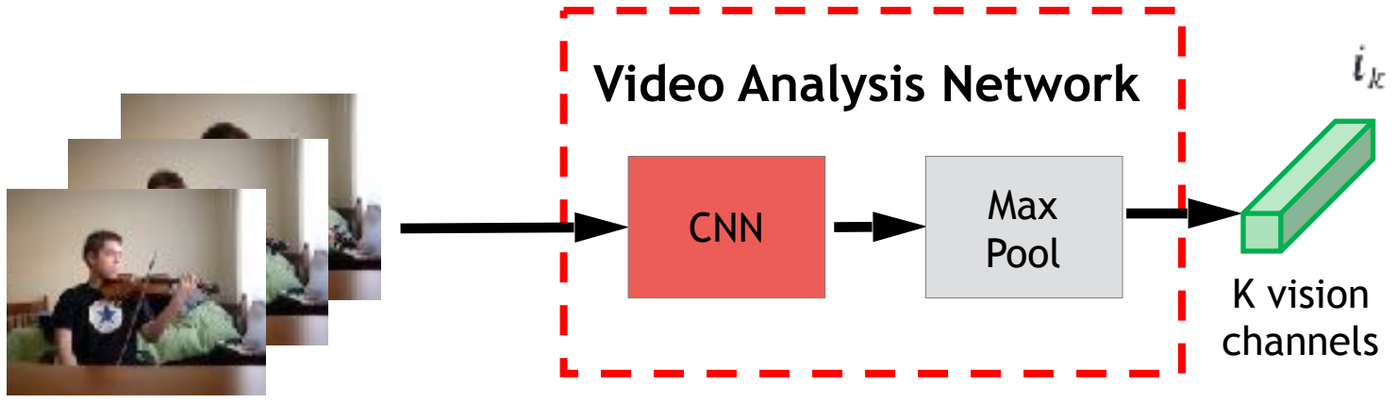
Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba.

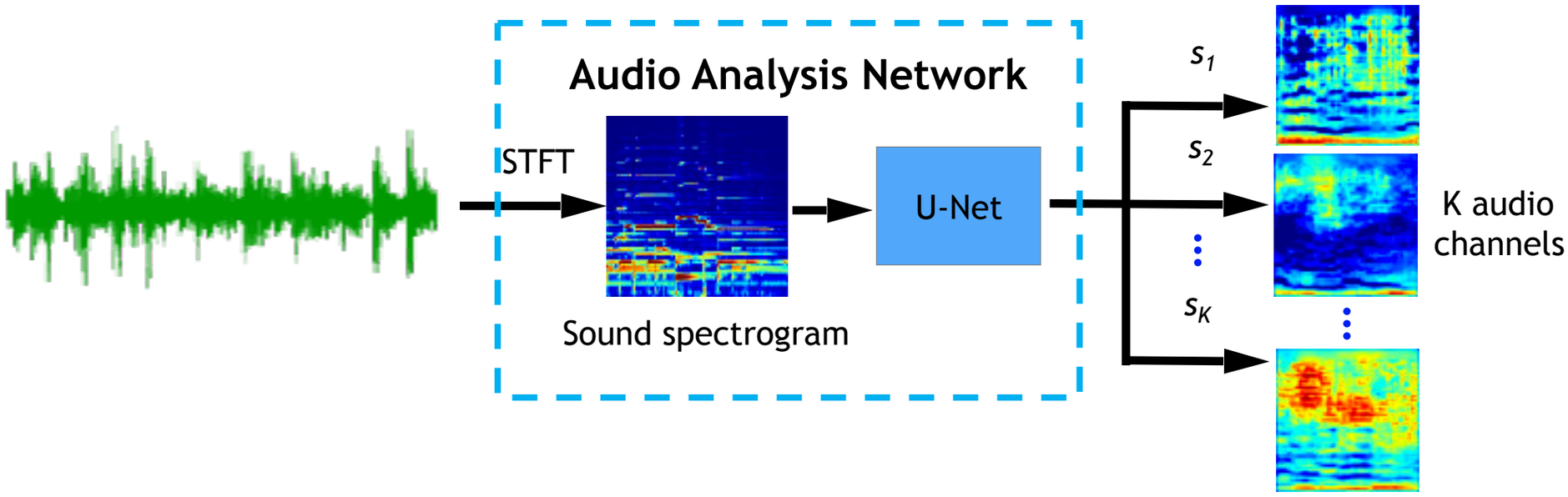
Audiovisual Model



Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba.

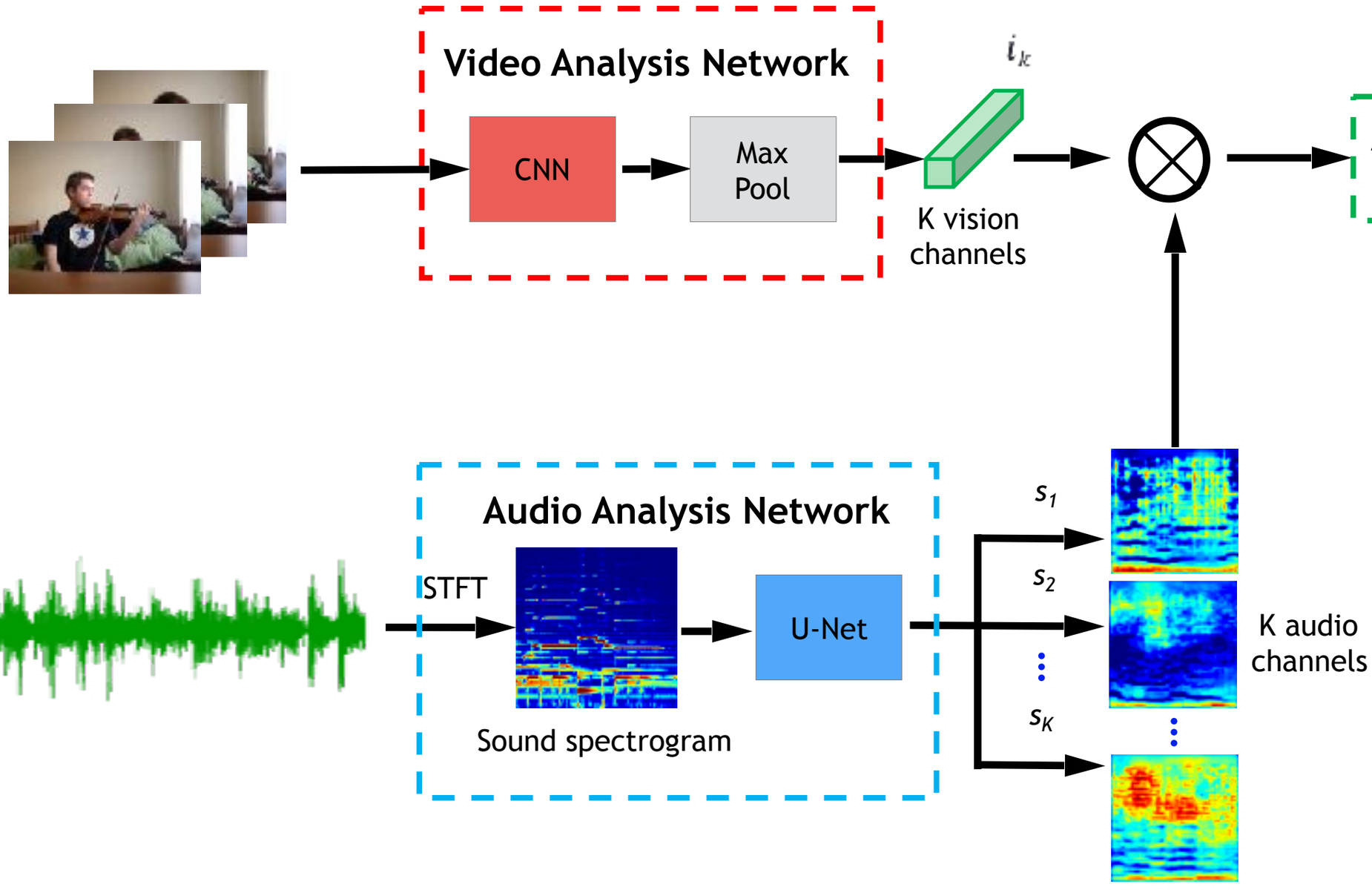
Audiovisual Model





Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba.

Audiovisual Model



Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba.

Original Audio

Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

What does this sound like?

Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

What does this sound like?

Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

What does this sound like?

Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

What regions are making sound?

Original Video



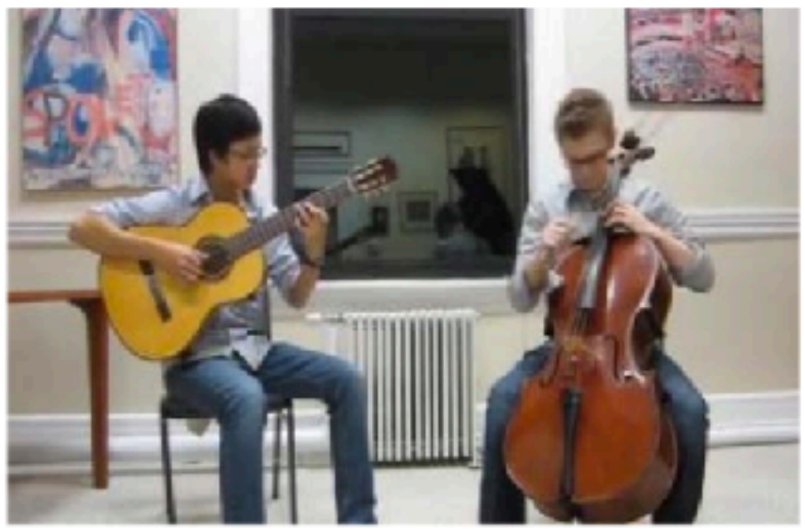



Estimated Volume

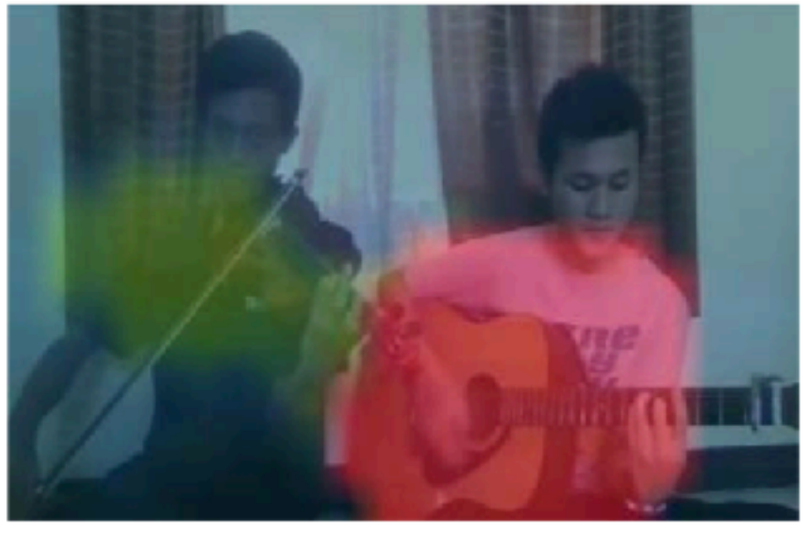
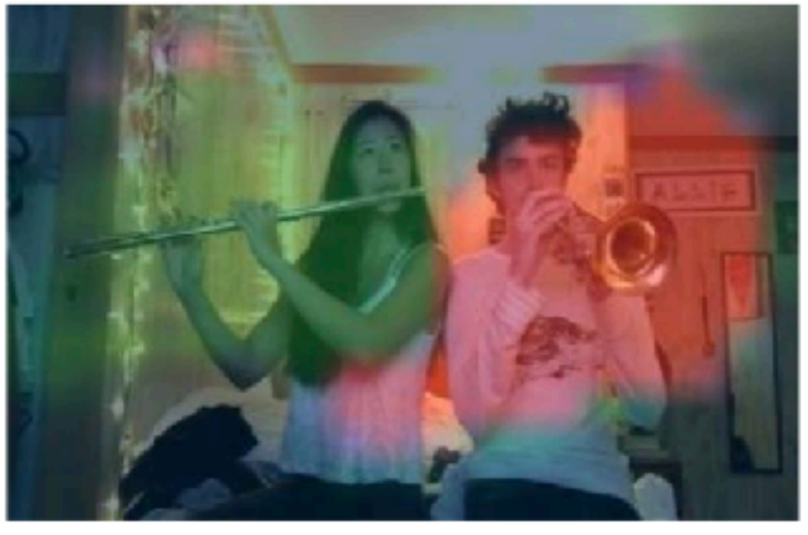
Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

What sounds are they making?

Original Video



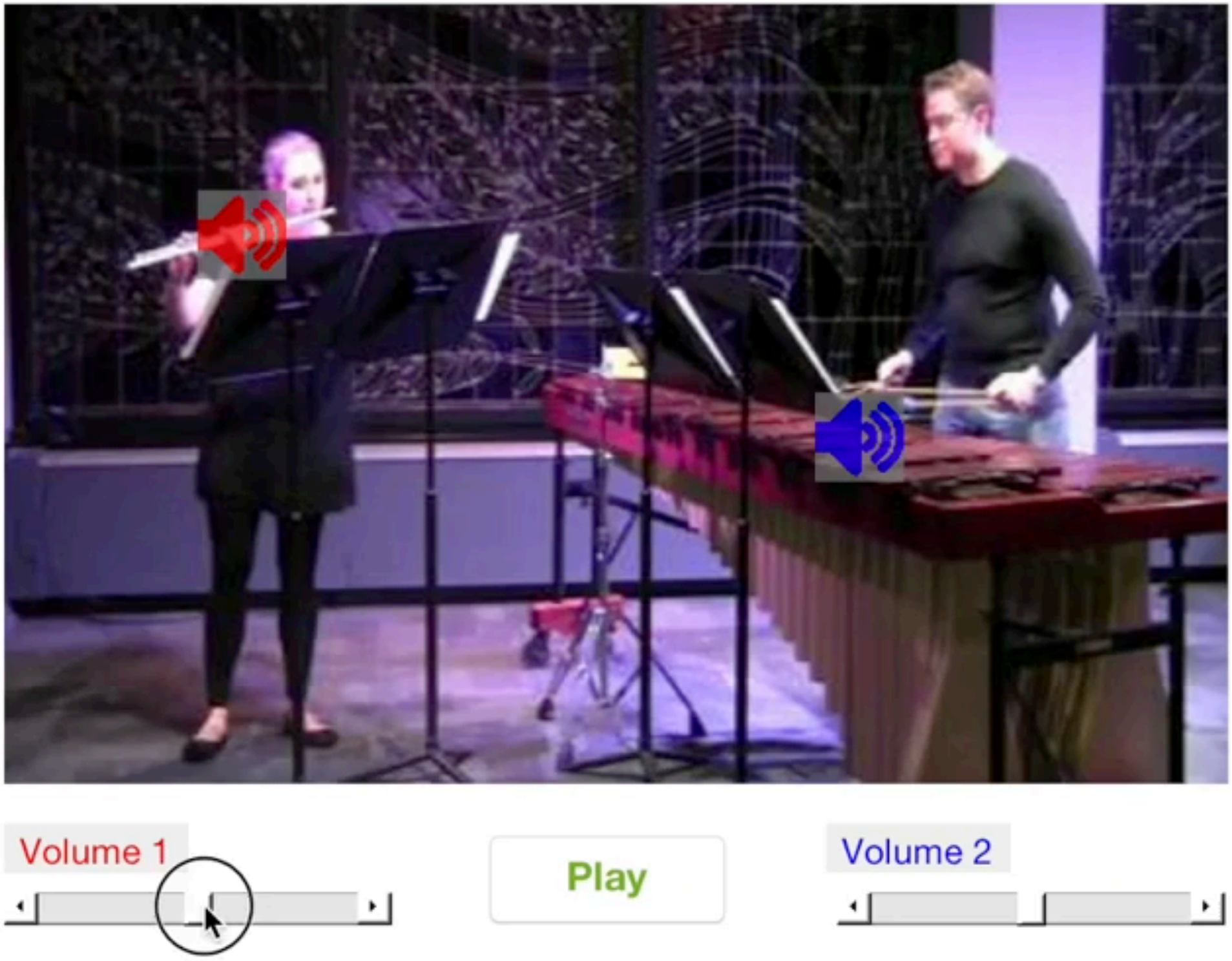


Embedding (projected and visualized as color)

Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

Adjusting Volume



Zhao, Gan, Rouditchenko, Vondrick, McDermott, Torralba. In submission.

Learning from Unlabeled Video

Carl Vondrick