Sound Localization from Motion: Jointly Learning Sound Direction and Camera Rotation
Ziyang Chen, Shengyi Qian, Andrew Owens 
University of Michigan

Motivation
The images and sounds that we perceive undergo subtle but geometrically consistent changes as we rotate our heads. Can we use these cues to learn audio and visual models of space?

Sound Localization from Motion
We propose SLFM: jointly learning sound direction and camera rotation from multi-view audio-visual data.

Idea: learn to enforce cross-modal consistency. Audio and visual predictions should agree with each other:
- Audio model: predict sound direction from stereo audio.
- Visual model: predict camera rotation from two images.

Method

Estimating pose and localizing sound
Cross-modal geometric consistency:
\[
\phi_{s,t} = \theta_t - \theta_s
\]

Learning representation via spatialization
We learn an audio-visual representation that conveys spatial cues by solving a cross-view binauralization task.

Experiments

Related work