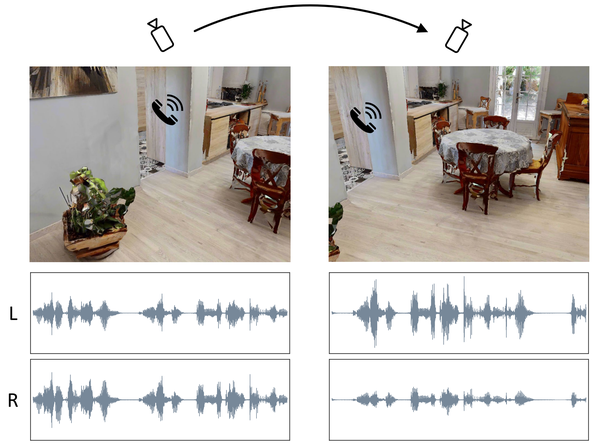
Sound Localization from Motion: Jointly Learning Sound Direction and Camera Rotation

Ziyang Chen, Shengyi Qian, Andrew Owens
University of Michigan

ICCV23 PARIS

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?

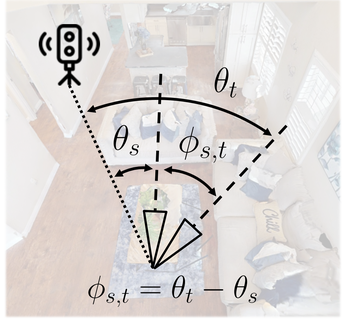


Sound Localization from Motion

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:

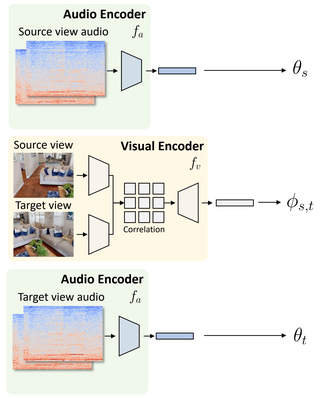
- **Audio model:** predict sound direction from stereo audio.
- **Visual model:** predict camera rotation from two images.



Method

Estimating pose and localizing sound

Cross-modal geometric consistency: $\phi_{s,t} = \theta_t - \theta_s$



$$\mathcal{L}_{\text{geo}} = \min_{\substack{\hat{\mathbf{r}}_s \in \{\mathbf{r}_s, Q\mathbf{r}_s\} \\ \hat{\mathbf{r}}_t \in \{\mathbf{r}_t, Q\mathbf{r}_t\}}} \|\hat{\mathbf{r}}_s - R_{s,t}\hat{\mathbf{r}}_t\|^2, \quad Q = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

$\mathbf{r}_i = \begin{bmatrix} \cos\theta_i & \sin\theta_i \end{bmatrix}^\top$ — sound direction

$R_{s,t} = \begin{bmatrix} \cos\phi_{s,t} & -\sin\phi_{s,t} \\ \sin\phi_{s,t} & \cos\phi_{s,t} \end{bmatrix}$ — camera rotation

$$\mathcal{L}_{\text{binaural}} = \mathcal{L}_{\text{BCE}}\left(\sin\theta_i, d_i\right)$$

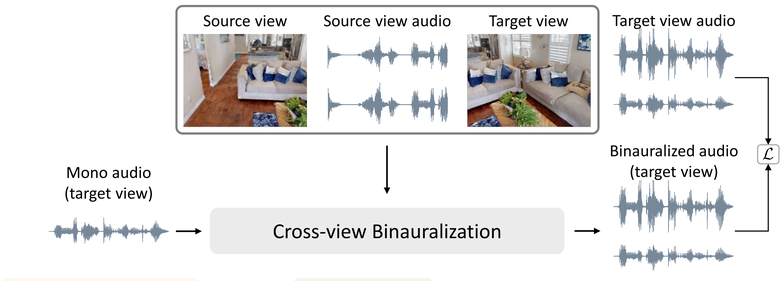
Louder in left (or right) ears

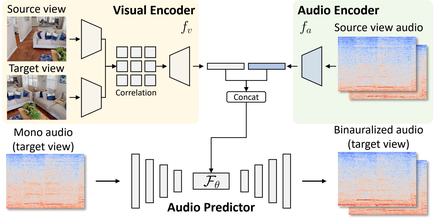
Overall loss:

$$\mathcal{L} = \lambda\mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{binaural}} + \mathcal{L}_{\text{sym}}$$

Learning representation via spatialization

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



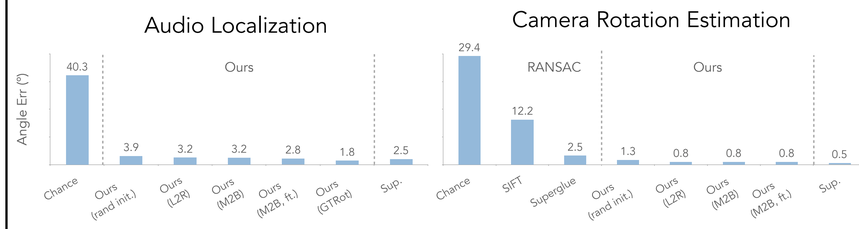


Binauralize audio $\mathbf{A}_t$ with images $(\mathbf{v}_t, \mathbf{v}_s)$ and audio $\mathbf{A}_s$ by optimizing STFT distance:

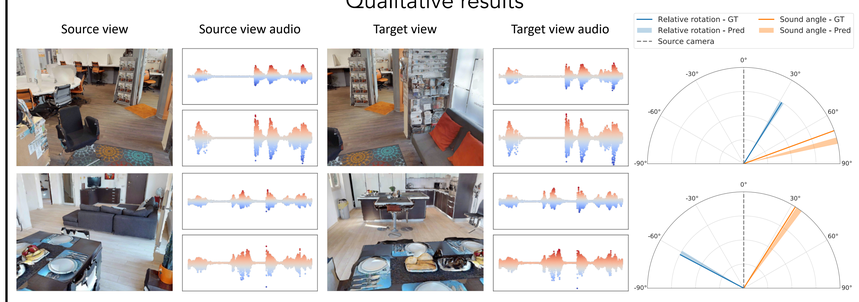
$$\hat{\mathbf{A}}_t = \mathcal{F}_\theta\left(\bar{\mathbf{A}}_t, f_v(\mathbf{v}_s, \mathbf{v}_t), f_a(\mathbf{A}_s)\right)$$

$$\mathcal{L}_{\text{pretext}} = ||\hat{\mathbf{A}}_t - \mathbf{A}_t||_1$$

Experiments

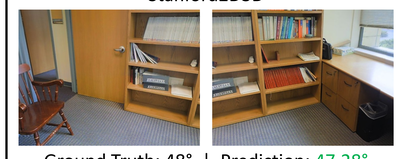


Qualitative results



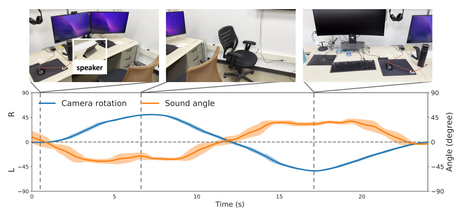
In-the-wild examples

Stanford2D3D



Ground Truth: 48° | Prediction: 47.28°

Real-world demo



Related work

1. Yang, Karren, et al. "Camera Pose Estimation and Localization with Active Audio Sensing." ECCV 2022
2. Chen, Changan, et al. "Novel-View Acoustic Synthesis." *CVPR 2023*.
3. Chen, Changan, et al. "Soundspaces 2.0: A simulation platform for visual-acoustic learning." NeurIPS 2022