Sight & Sound Workshop, CVPR 2020, June 2020

BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

Jesper Haahr Christensen[1] Sascha Hornauer[2] Stella Yu[2]

[1]Technical University of Denmark, [2]University of California, Berkeley / ICSI

Existing Sensors have some Drawbacks

- Vision is valuable sensor but sometimes fails
- Ultrasound, Radar or LIDAR sensors are often costly, complex and provide limited information

Image: Dark shadows by Christopher Kyba is licensed under a Creative Commons Attribution 4.0 International License. https://creativecommons.org/licenses/by/4.0/legalcode

Approach Inspired by Nature

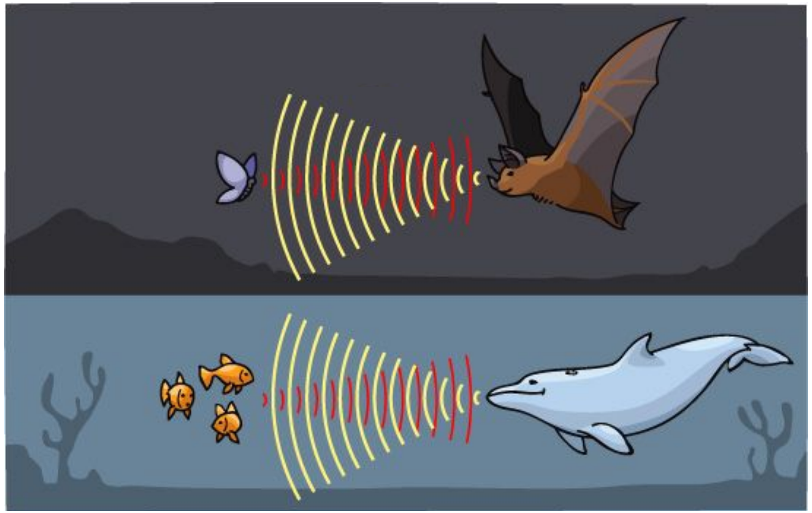
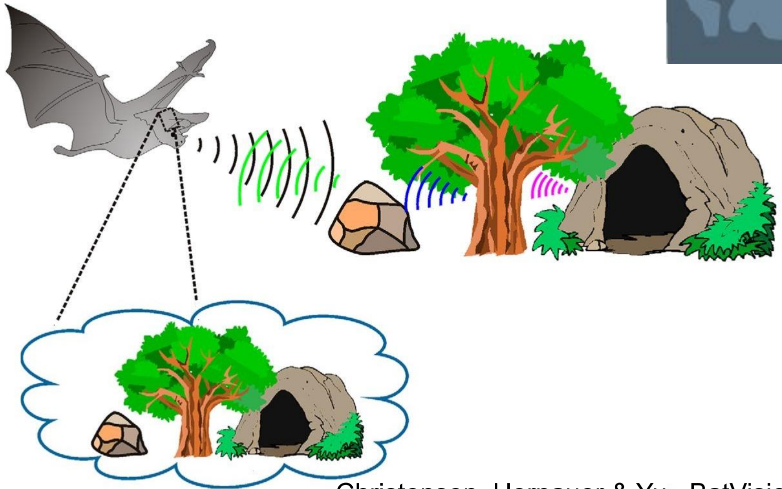
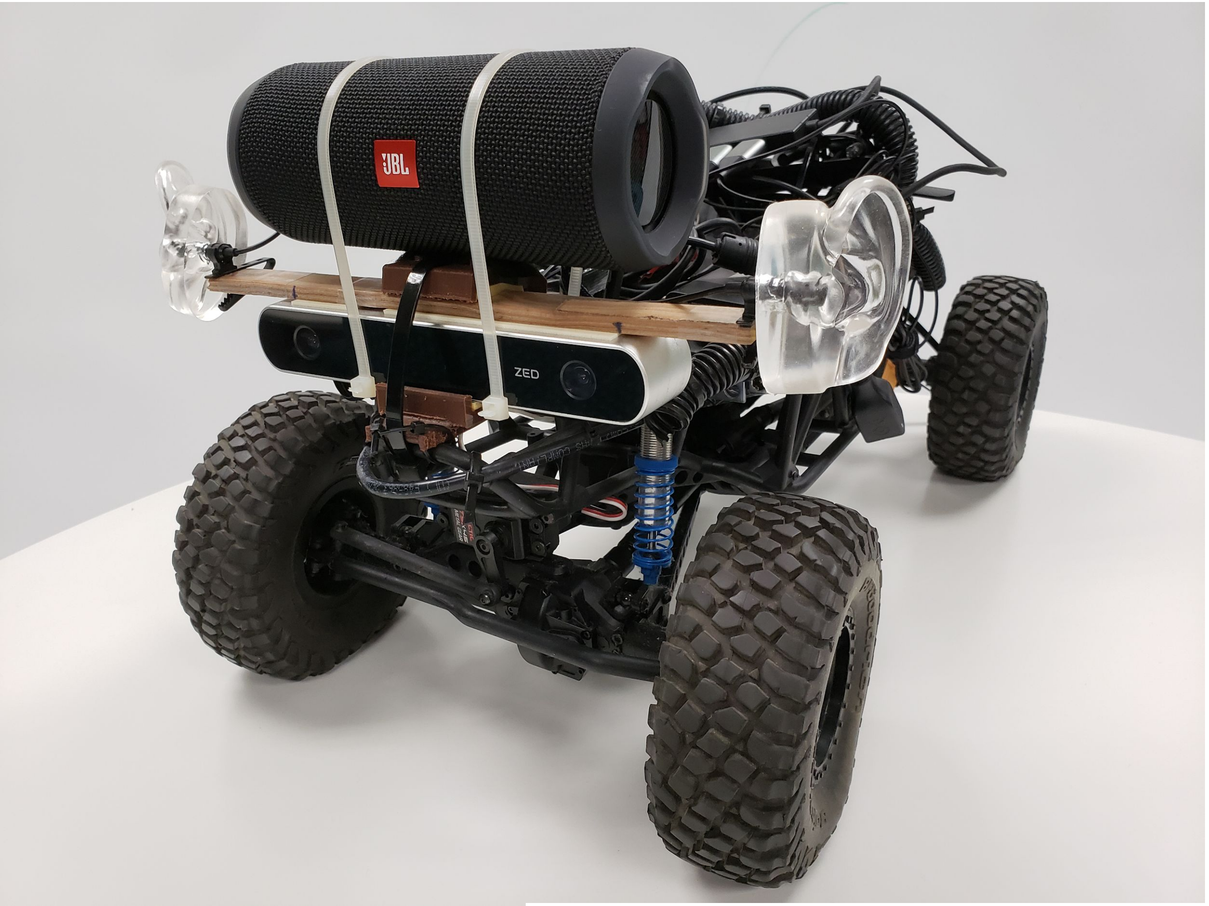


Image credit: https://askabiologist.asu.edu/echolocation

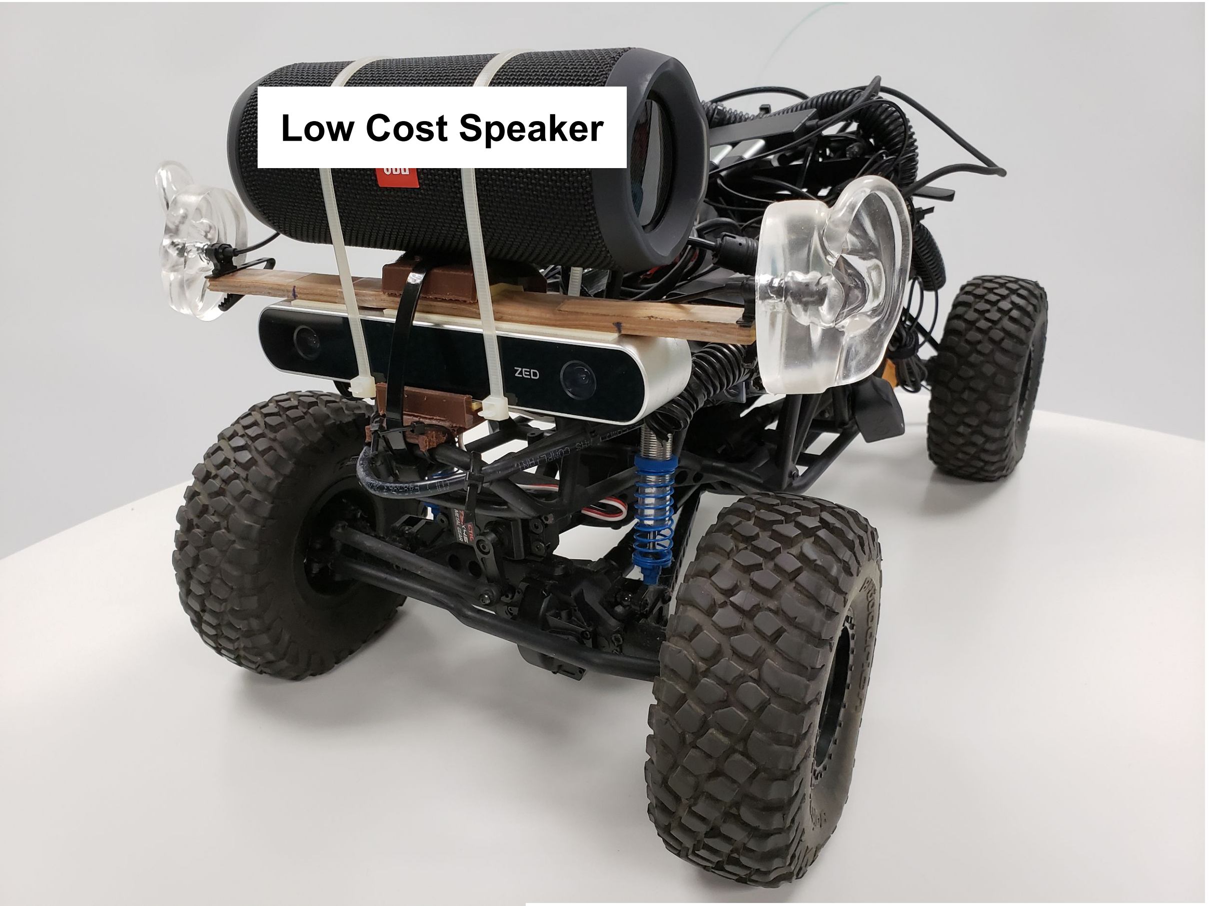


Batvision System



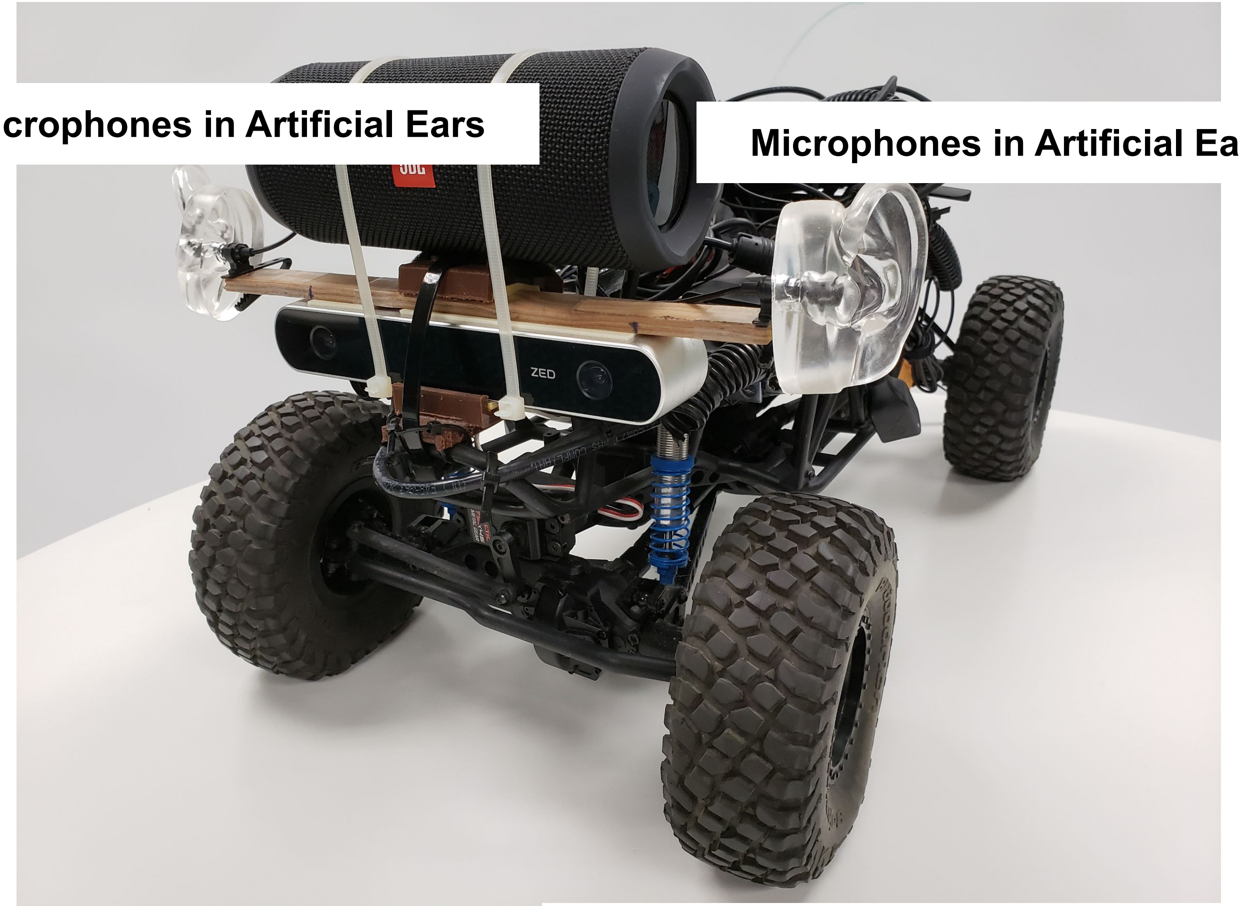
Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

Batvision System



Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

Batvision System

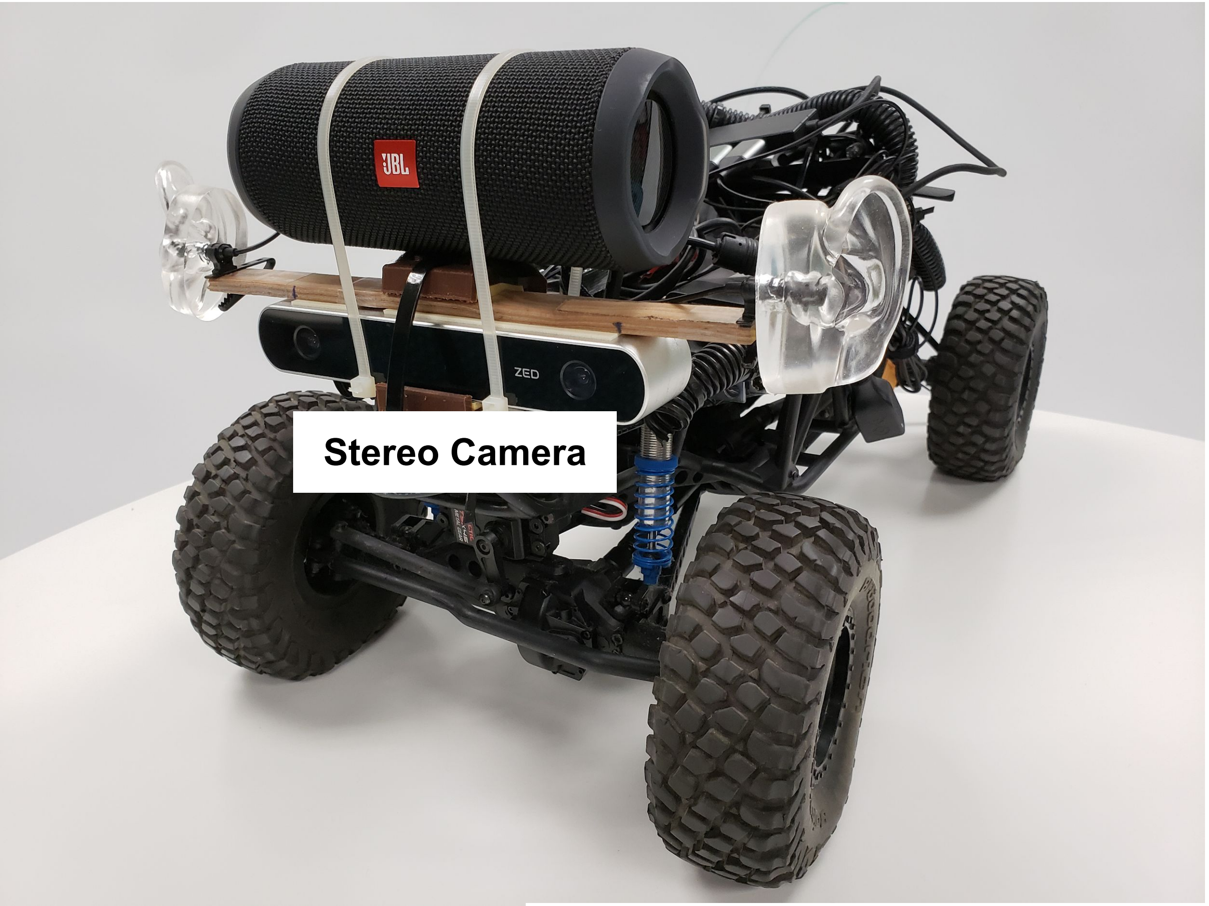


Microphones in Artificial Ears

Microphones in Artificial Ears

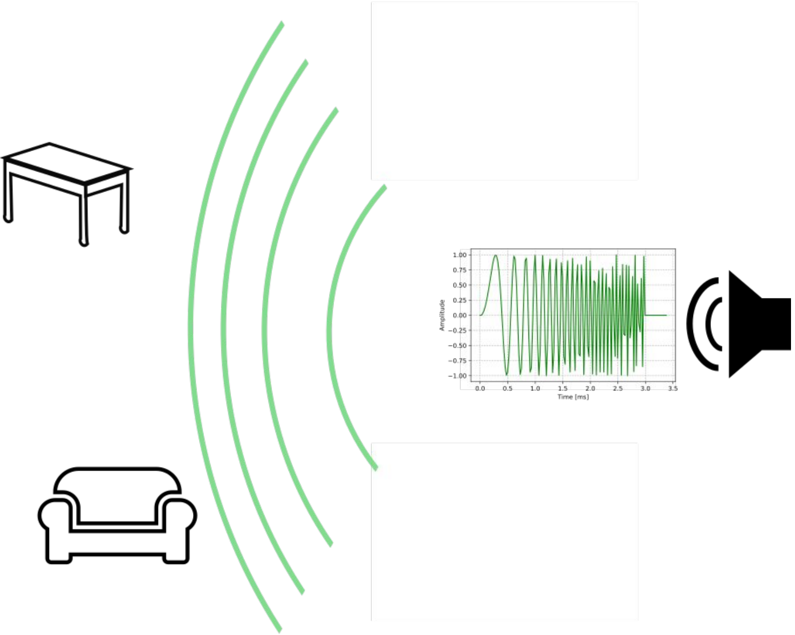
Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

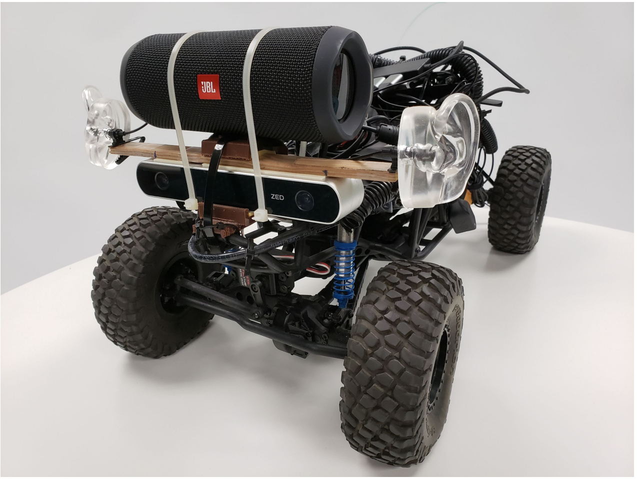
Batvision System



Stereo Camera

Batvision System

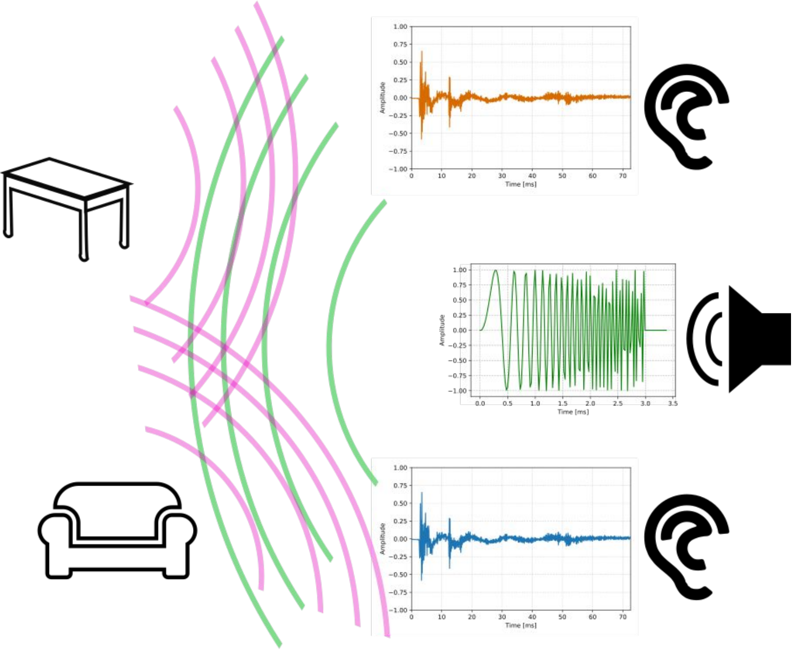


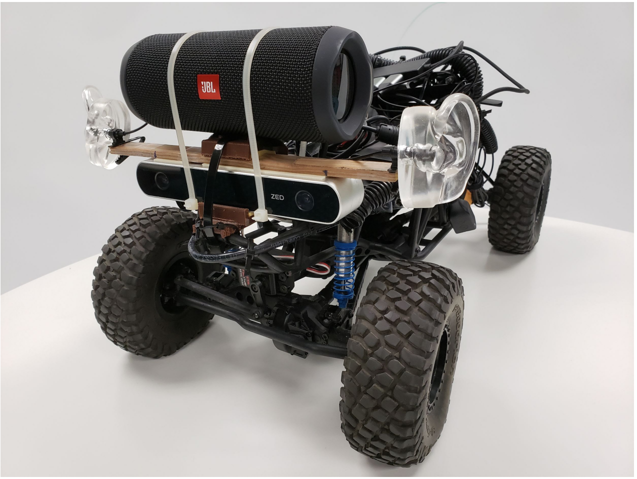


Chirp:

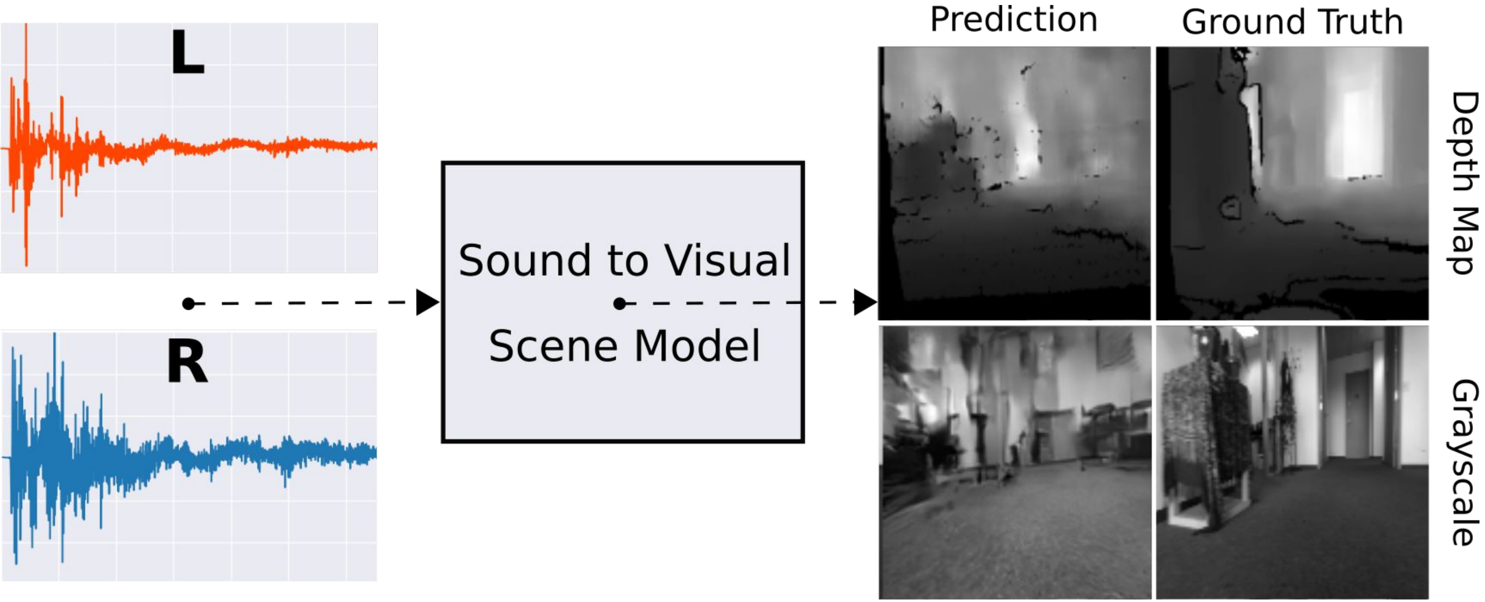
- 3 Milliseconds
- From 20hz to 20kHz

Batvision System





Neural Network Prediction of Visual Layout from Sound



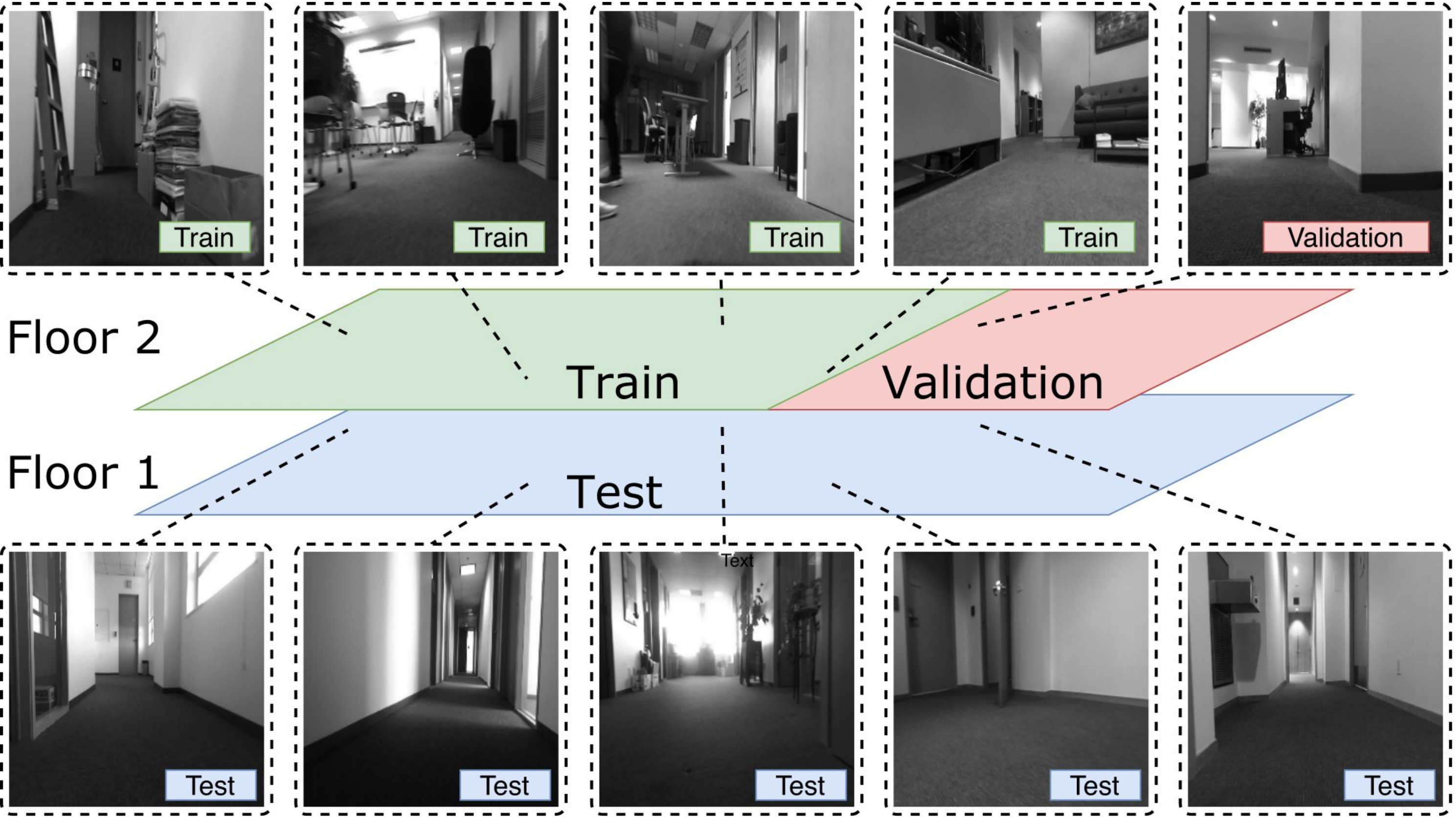
Sound to Visual Scene Model

Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

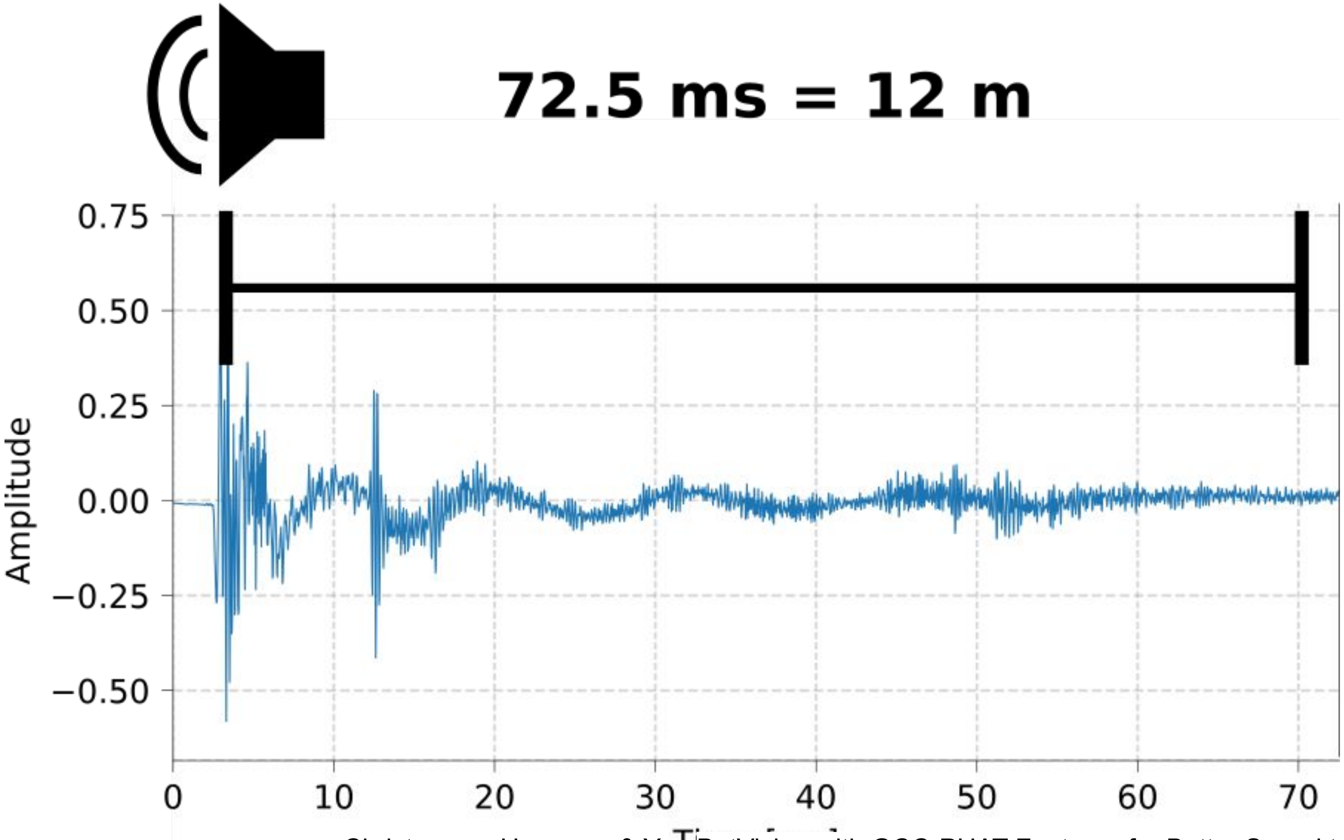
Dataset: Binaural echo to depth

Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

Data Collection in Office Building



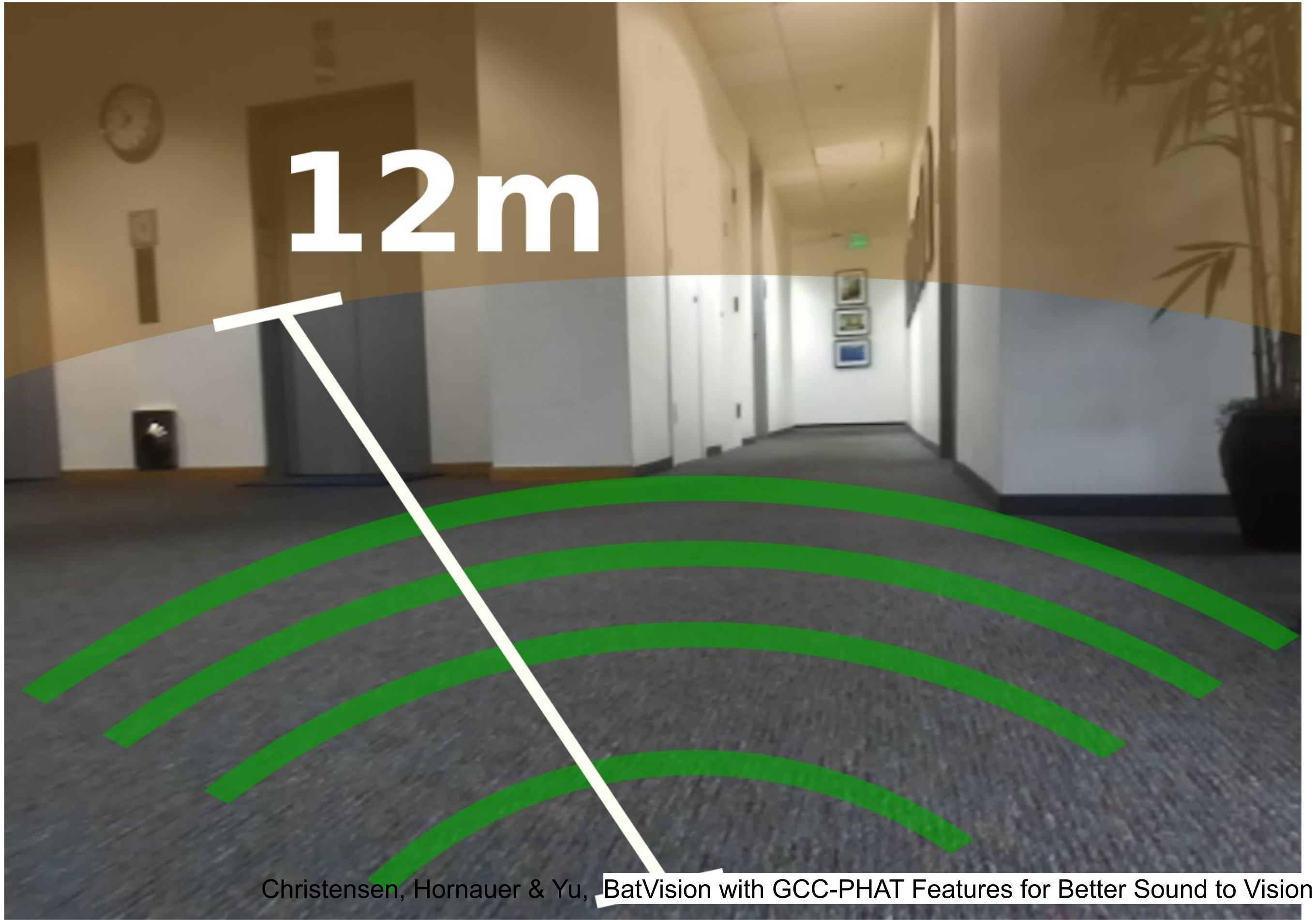
Echos cut off at fixed distance



72.5 ms = 12 m

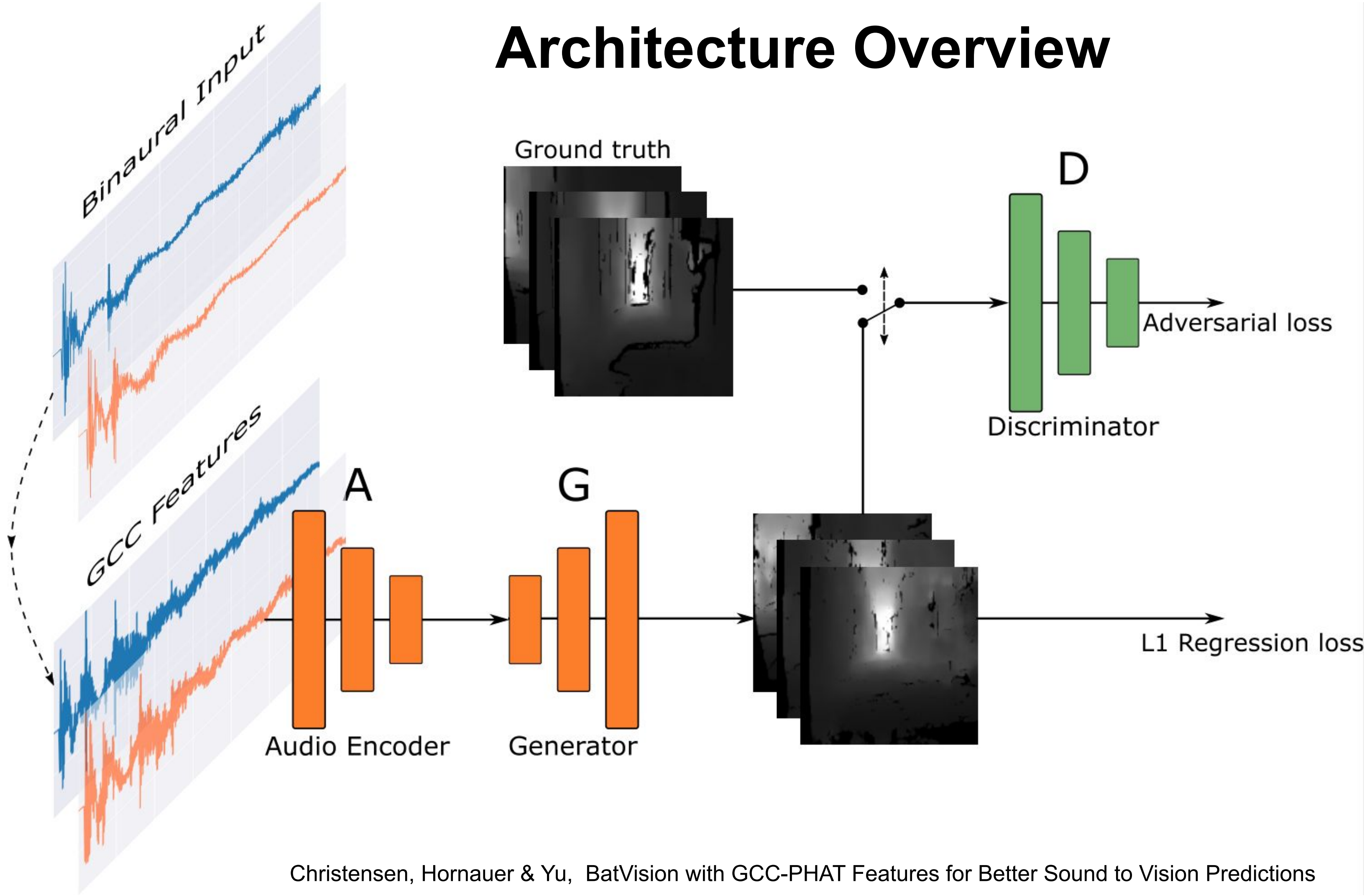
Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

Structure Beyond can not be Directly Observed



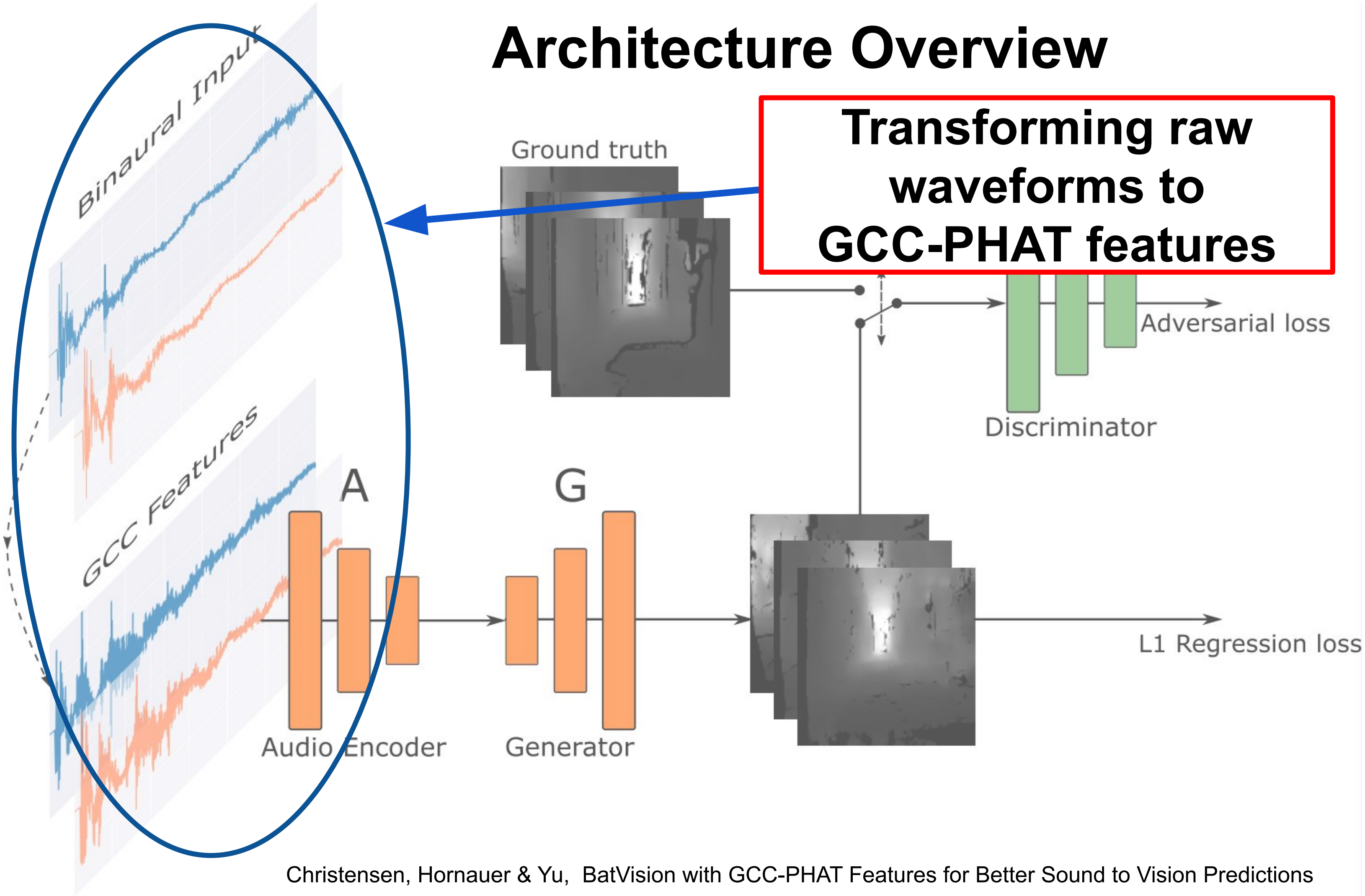
Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

Architecture Overview



Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

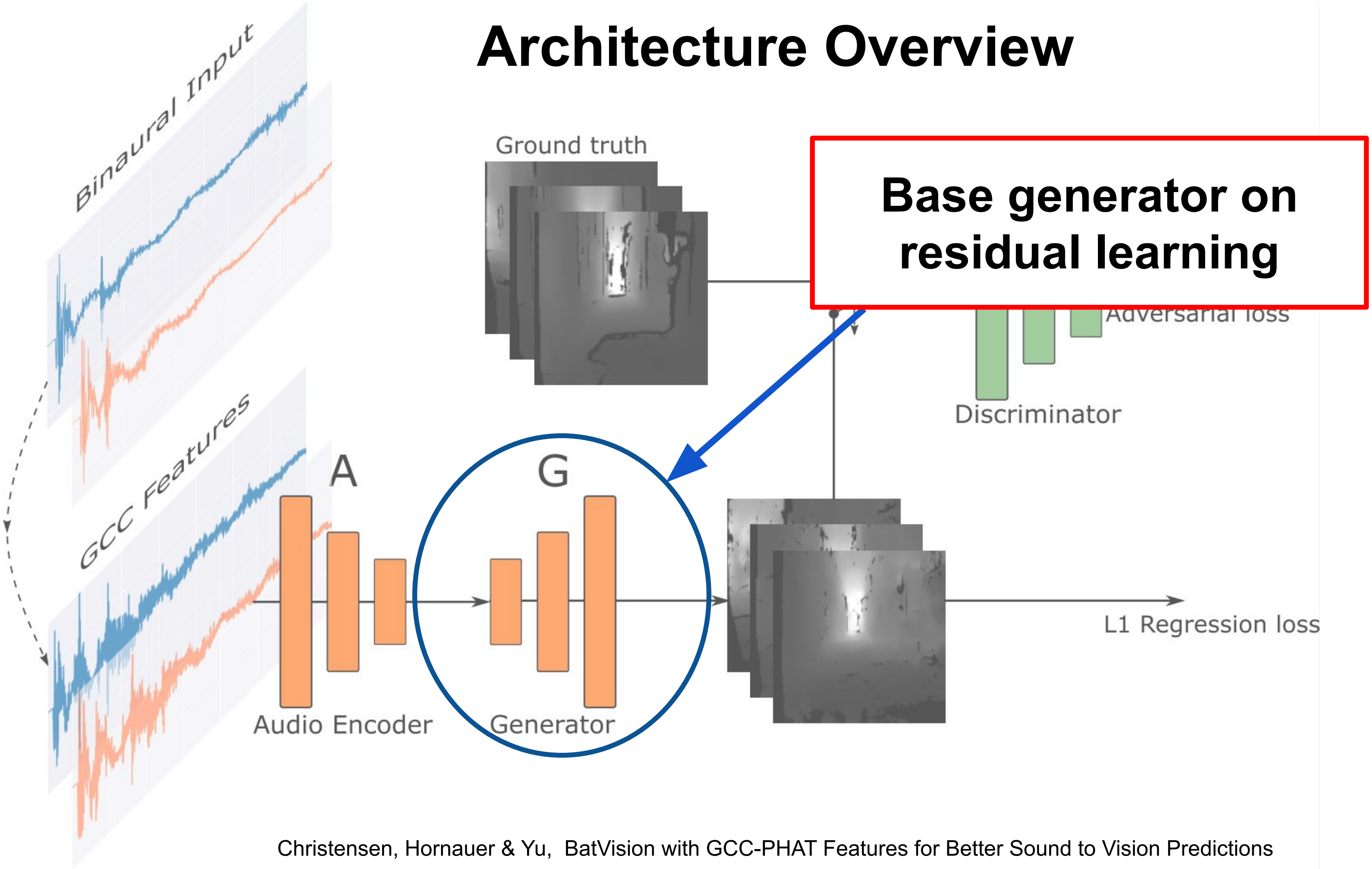
Architecture Overview



Transforming raw waveforms to GCC-PHAT features

Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

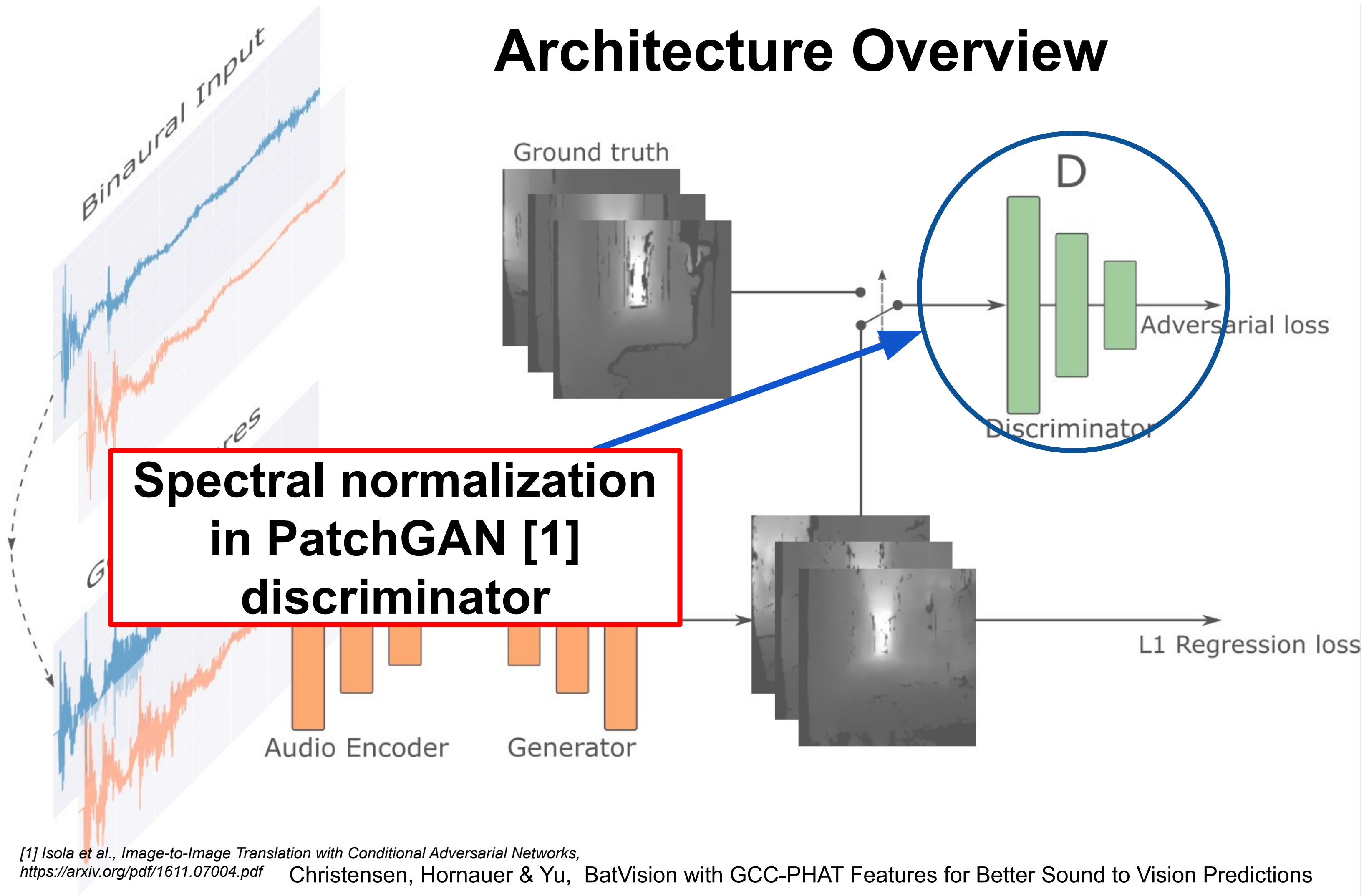
Architecture Overview



Base generator on residual learning

Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

Architecture Overview

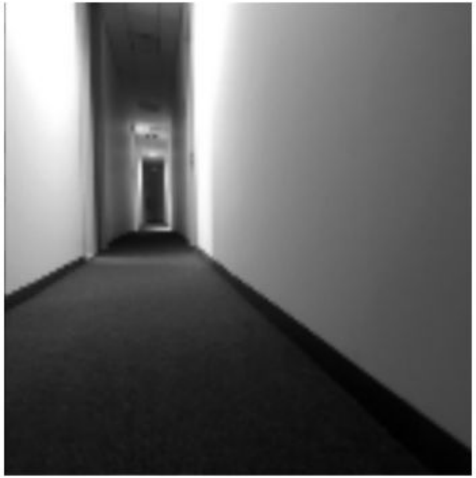


Spectral normalization in PatchGAN [1] discriminator

[1] *Isola et al., Image-to-Image Translation with Conditional Adversarial Networks, https://arxiv.org/pdf/1611.07004.pdf*

Results: Depth map prediction

Generated from a single binaural echo only!

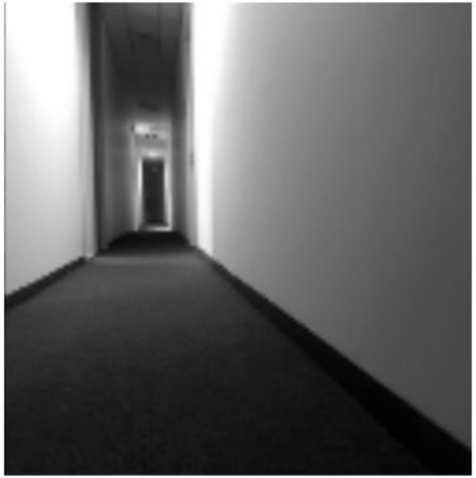


Grayscale image
from camera

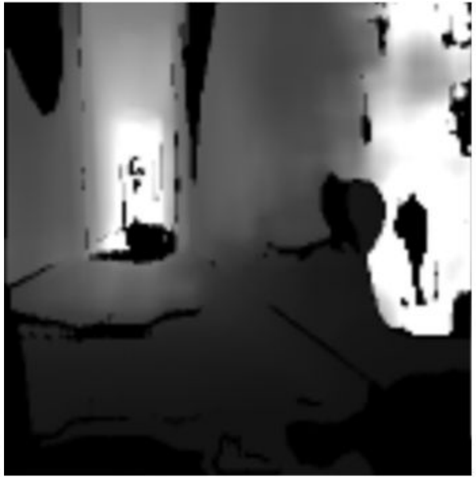
Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

Results: Depth map prediction

Generated from a single binaural echo only!



Grayscale image
from camera

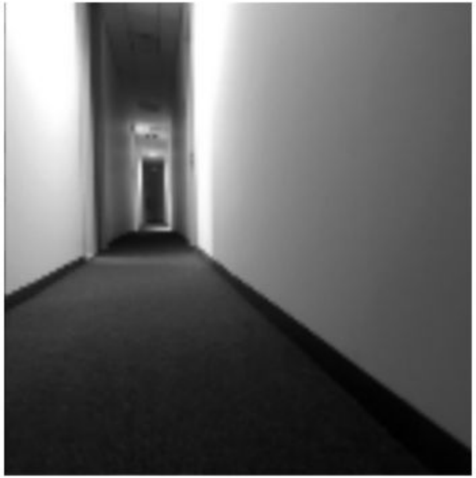


Depth map
from stereo camera

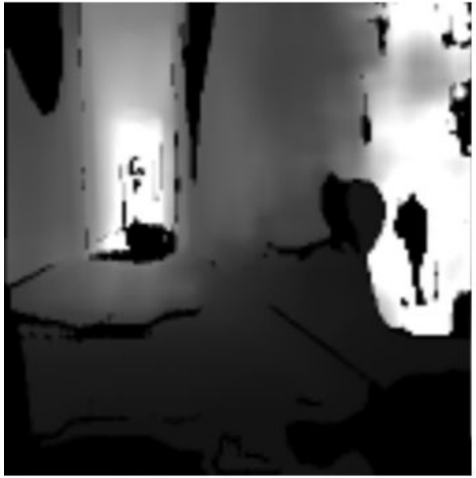
Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

Results: Depth map prediction

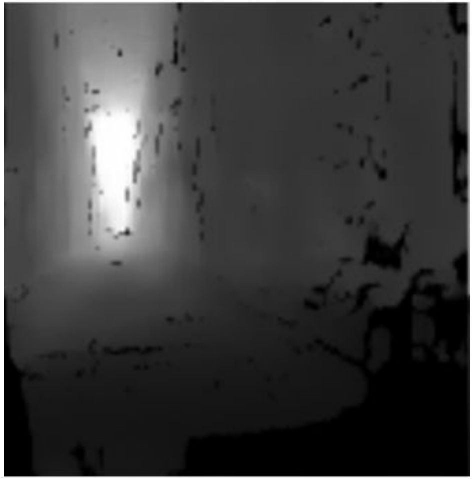
Generated from a single binaural echo only!



Grayscale image
from camera



Depth map
from stereo camera

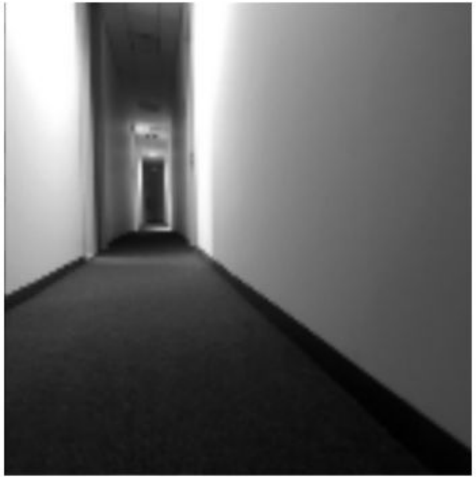


**Predicted depth map
from BatVision [1]**

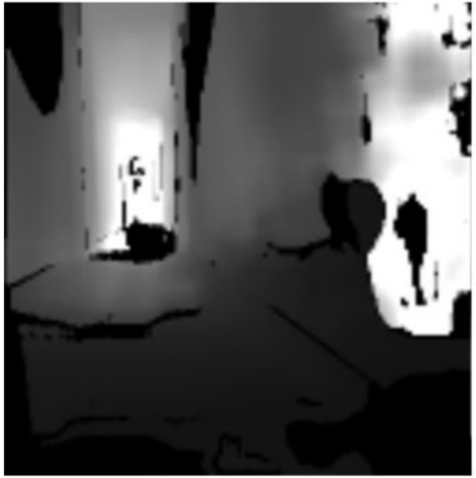
[1] Christensen, Hornauer, Yu. BatVision: Learning to See 3D Spatial Layout with Two Ears. In ICRA 2020.

Results: Depth map prediction

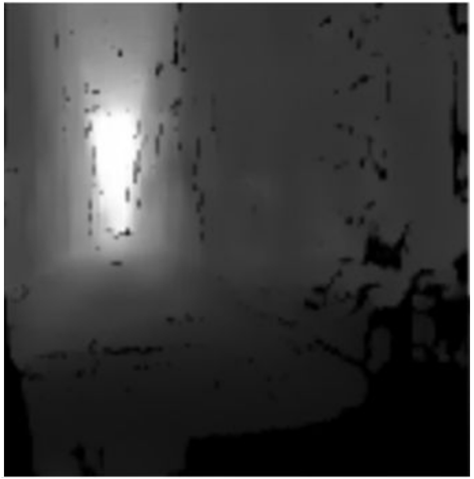
Generated from a single binaural echo only!



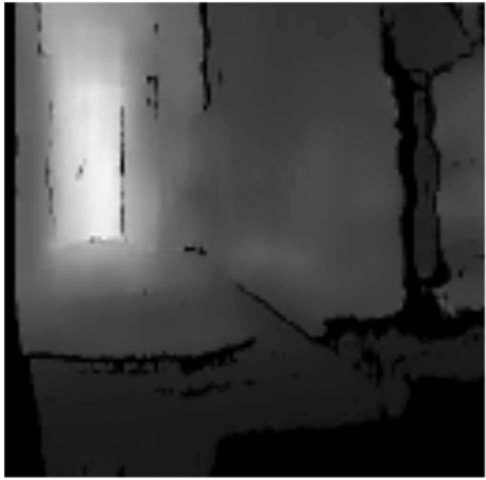
Grayscale image
from camera



Depth map
from stereo camera



**Predicted depth map
from BatVision [1]**



**Predicted depth map
from our improved
BatVision model**

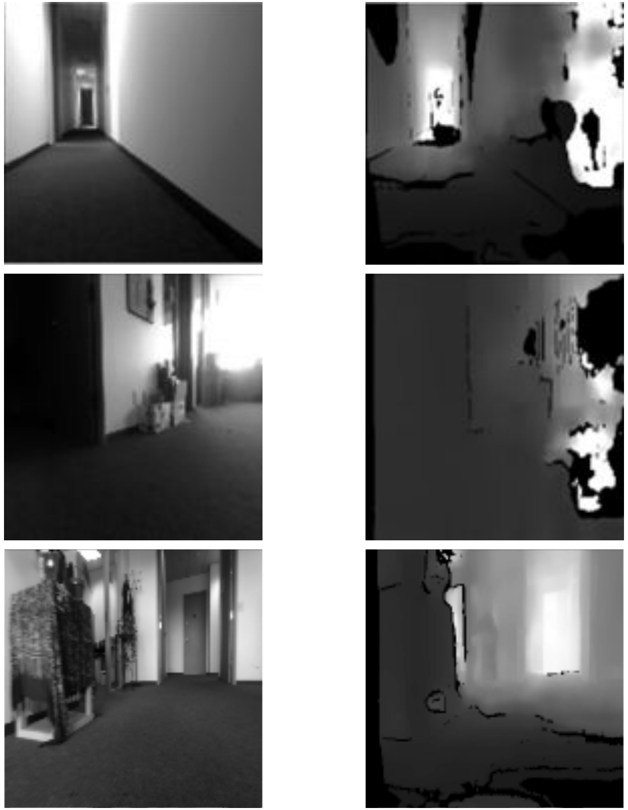
[1] Christensen, Hornauer, Yu. BatVision: Learning to See 3D Spatial Layout with Two Ears. In ICRA 2020.

Results: Improved predictions and less noise

Scene

Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

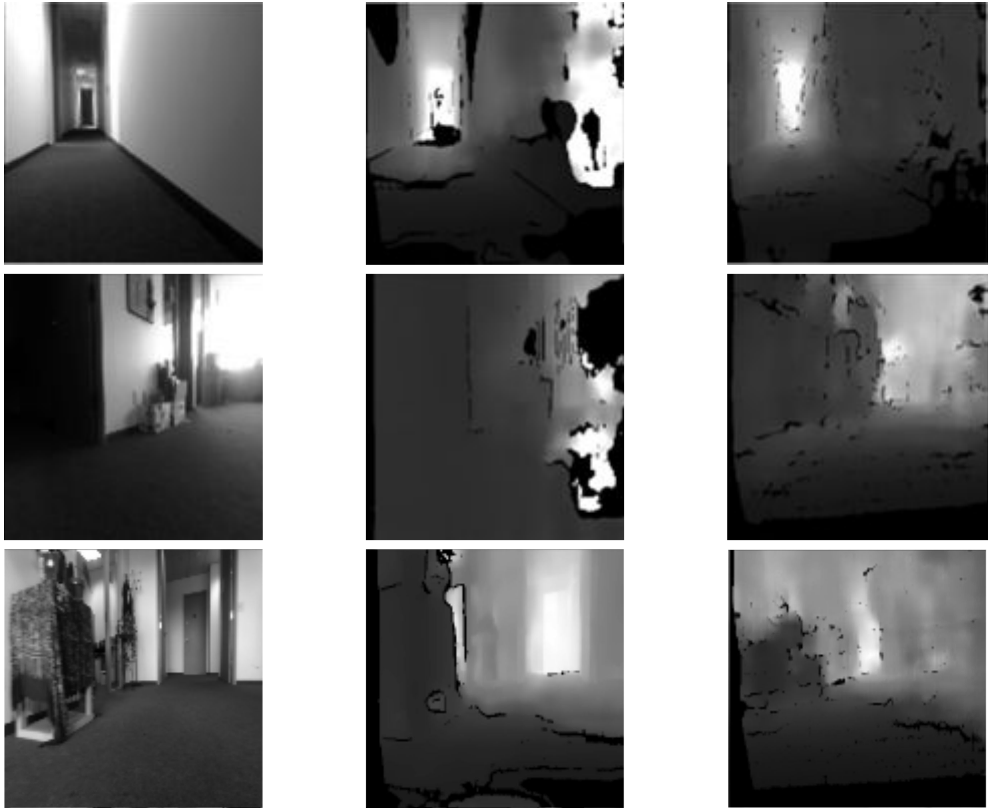
Results: Improved predictions and less noise



Scene **Stereo GT**

Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

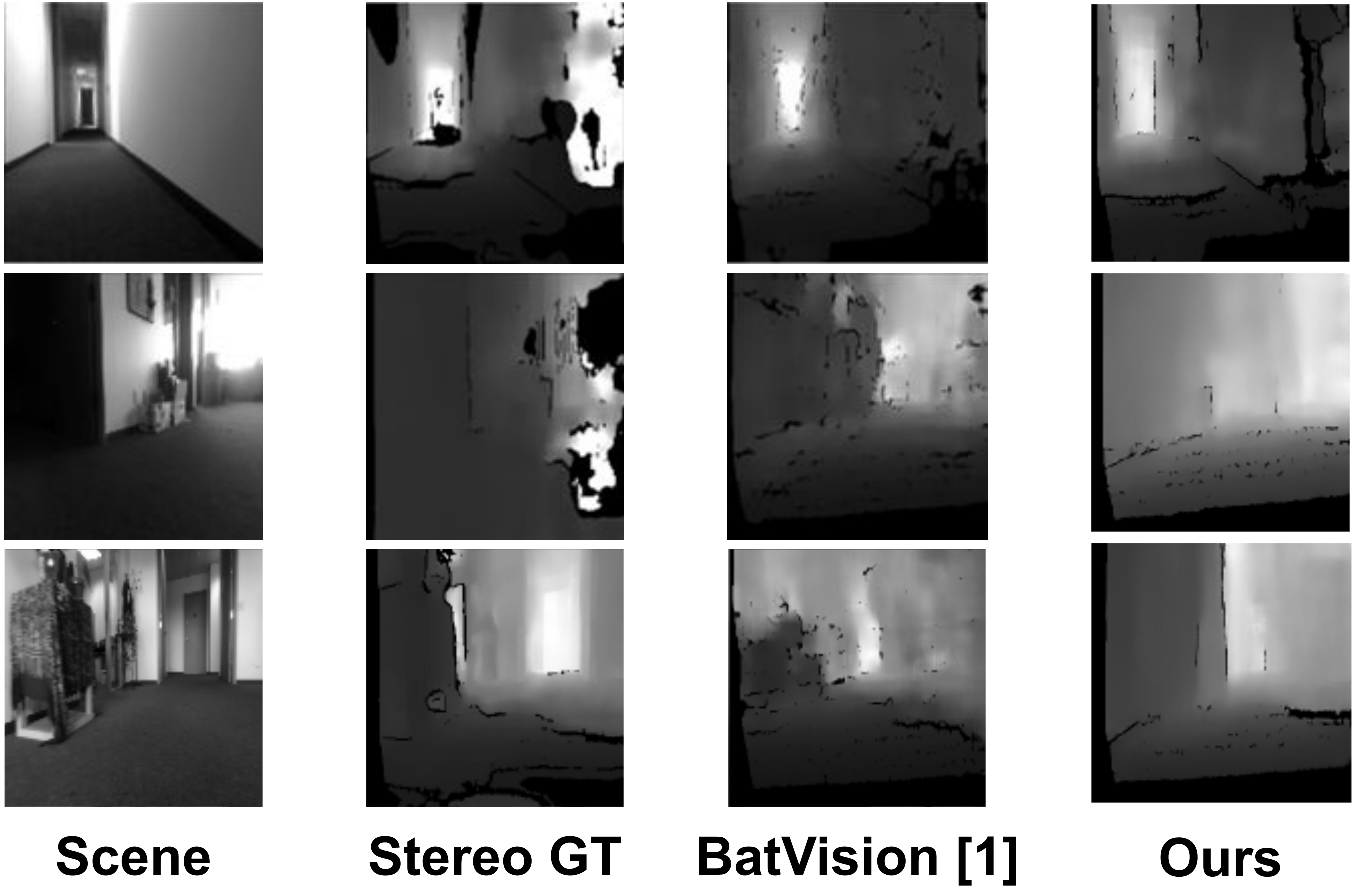
Results: Improved predictions and less noise



Scene **Stereo GT** **BatVision [1]**

[1] Christensen, Hornauer, Yu. BatVision: Learning to See 3D Spatial Layout with Two Ears. In ICRA 2020.

Results: Improved predictions and less noise



Scene

Stereo GT

BatVision [1]

Ours

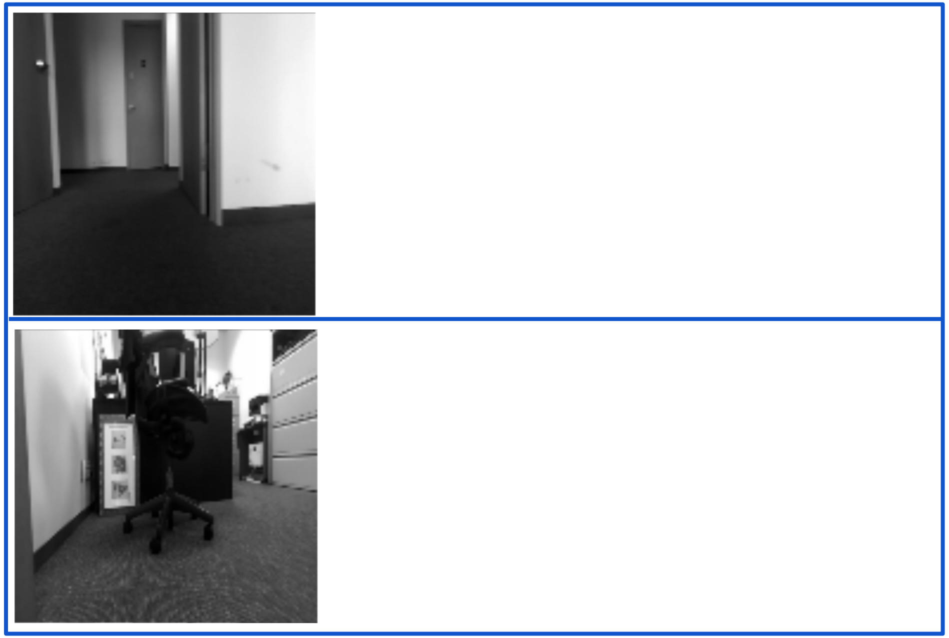
[1] Christensen, Hornauer, Yu. BatVision: Learning to See 3D Spatial Layout with Two Ears. In ICRA 2020.

Results: Grayscale layout

No depth used for training!

Plausible layout of free space / obstacles

Scene



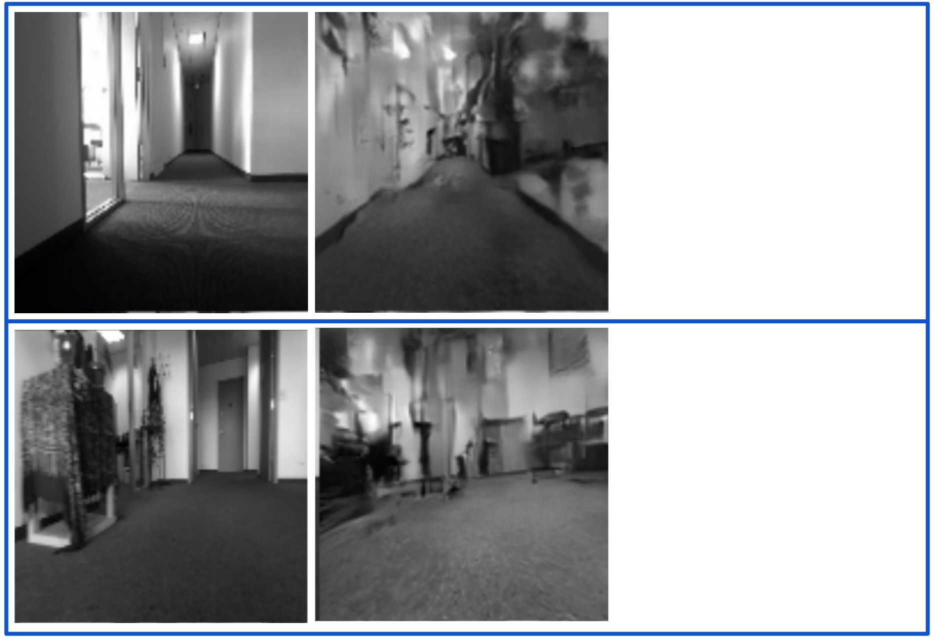
Scene

Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions

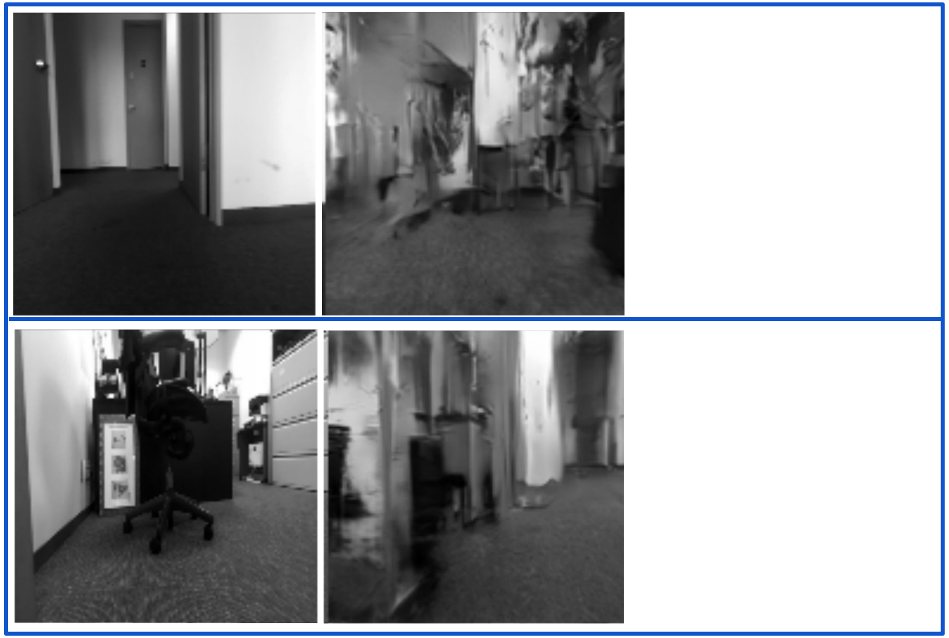
Results: Grayscale layout

No depth used for training!

Plausible layout of free space / obstacles



Scene BatVision [1]



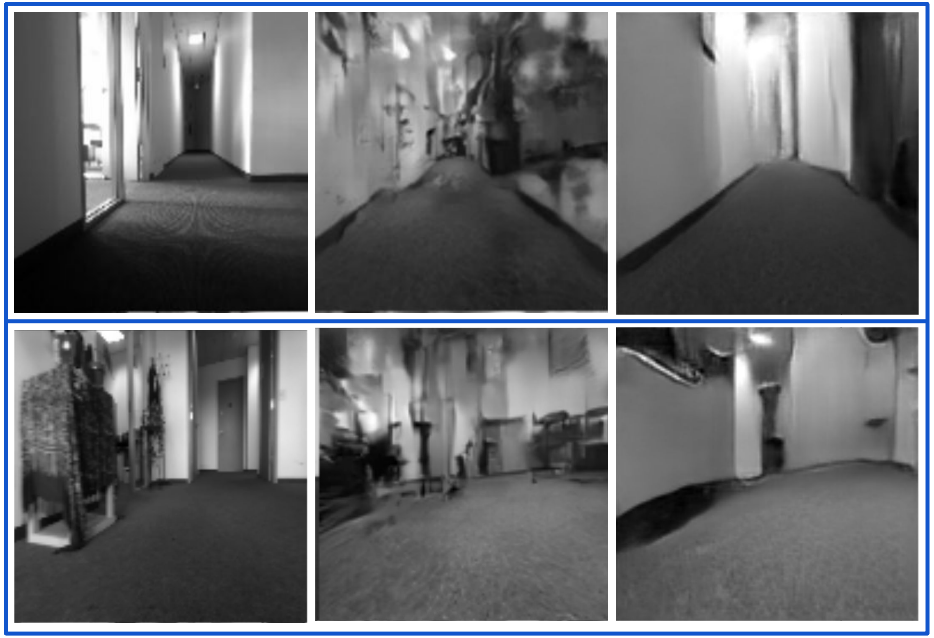
Scene BatVision [1]

[1] Christensen, Hornauer, Yu. BatVision: Learning to See 3D Spatial Layout with Two Ears. In ICRA 2020.

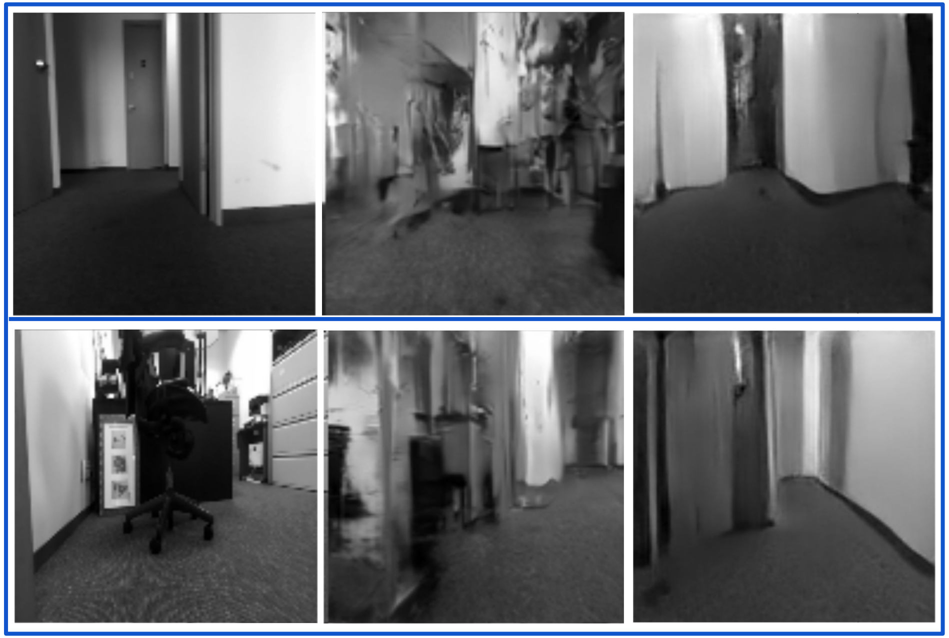
Results: Grayscale layout

No depth used for training!

Plausible layout of free space / obstacles



Scene **BatVision [1]** ***Ours***



Scene **BatVision [1]** ***Ours***

[1] Christensen, Hornauer, Yu. BatVision: Learning to See 3D Spatial Layout with Two Ears. In ICRA 2020.

Conclusion

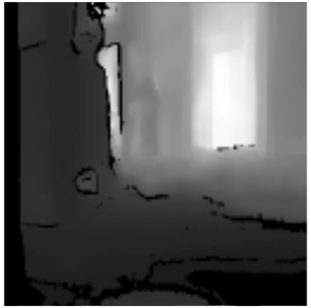
➜ Improving sound-to-**vision**

➜ Increasing perceptual quality and quantitative measures

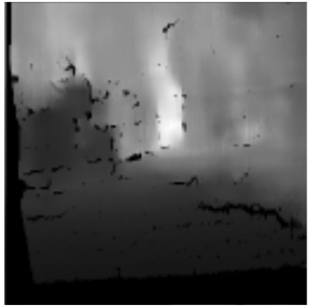
➜ More stable training process

➜ Less noisy depth and layout predictions

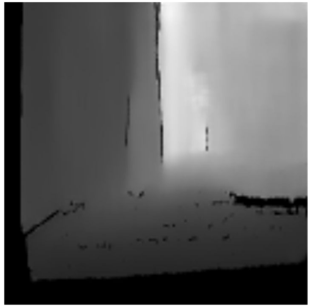
Stereo depth



BatVision

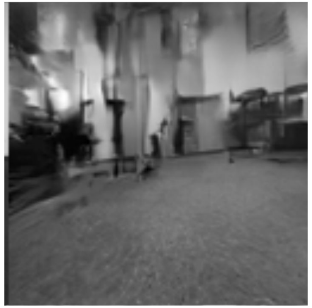


Our improved BatVision

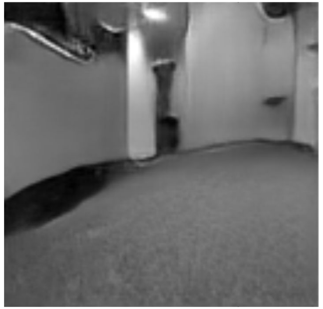


Grayscale image

BatVision layout



Our improved BatVision layout



Christensen, Hornauer & Yu, BatVision with GCC-PHAT Features for Better Sound to Vision Predictions