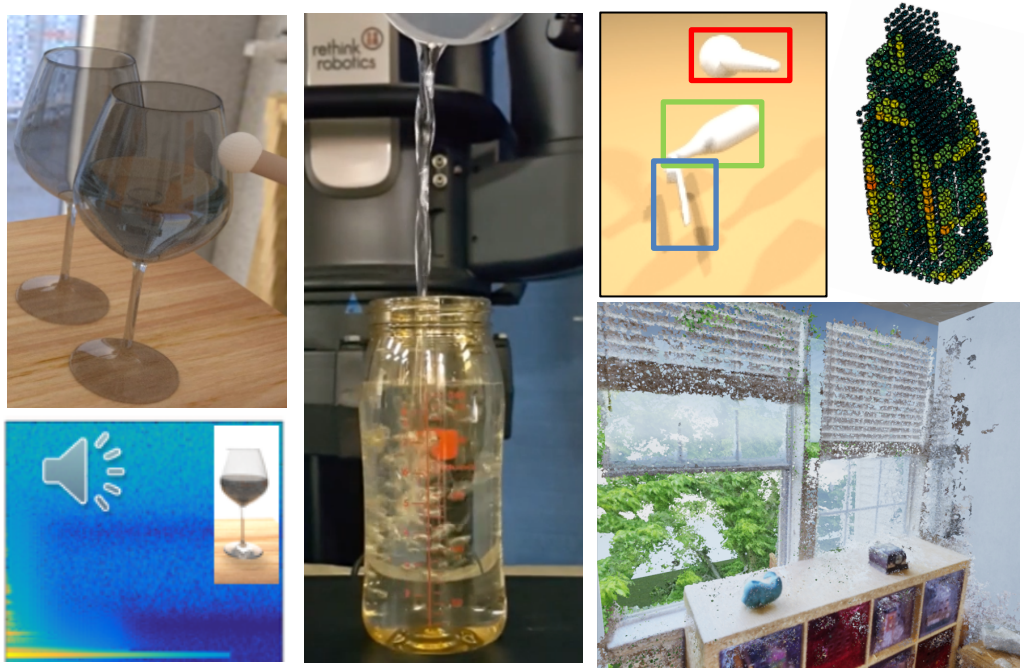
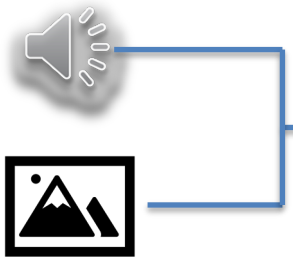
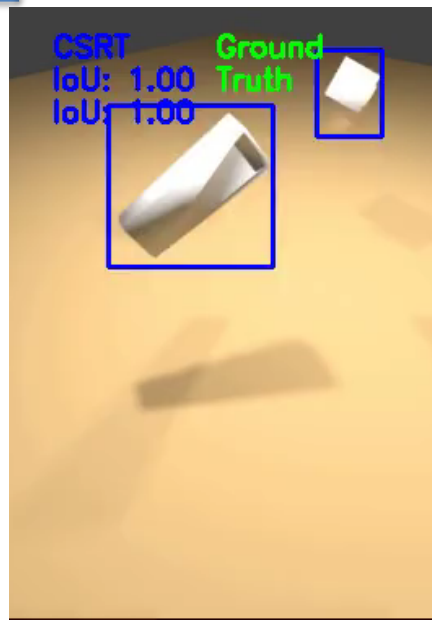
Multimodal Learning for Audio and Visual Processing





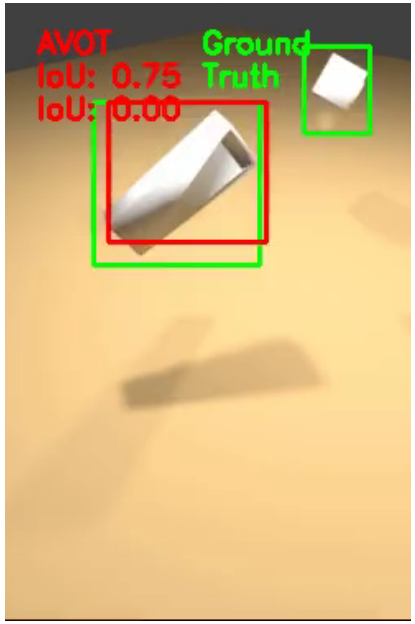
Doctoral Dissertation Defense
Justin Wilson
Advisors: Henry Fuchs and Ming Lin
July 29, 2020

Sights and sounds are all around us



+

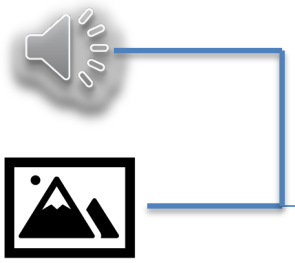
Audio



+

Visual

[1] Sound20K [Zhang et al. 2017]
[2] Example-Guided Physically Based Modal Sound Synthesis [Ren et al. 2013]

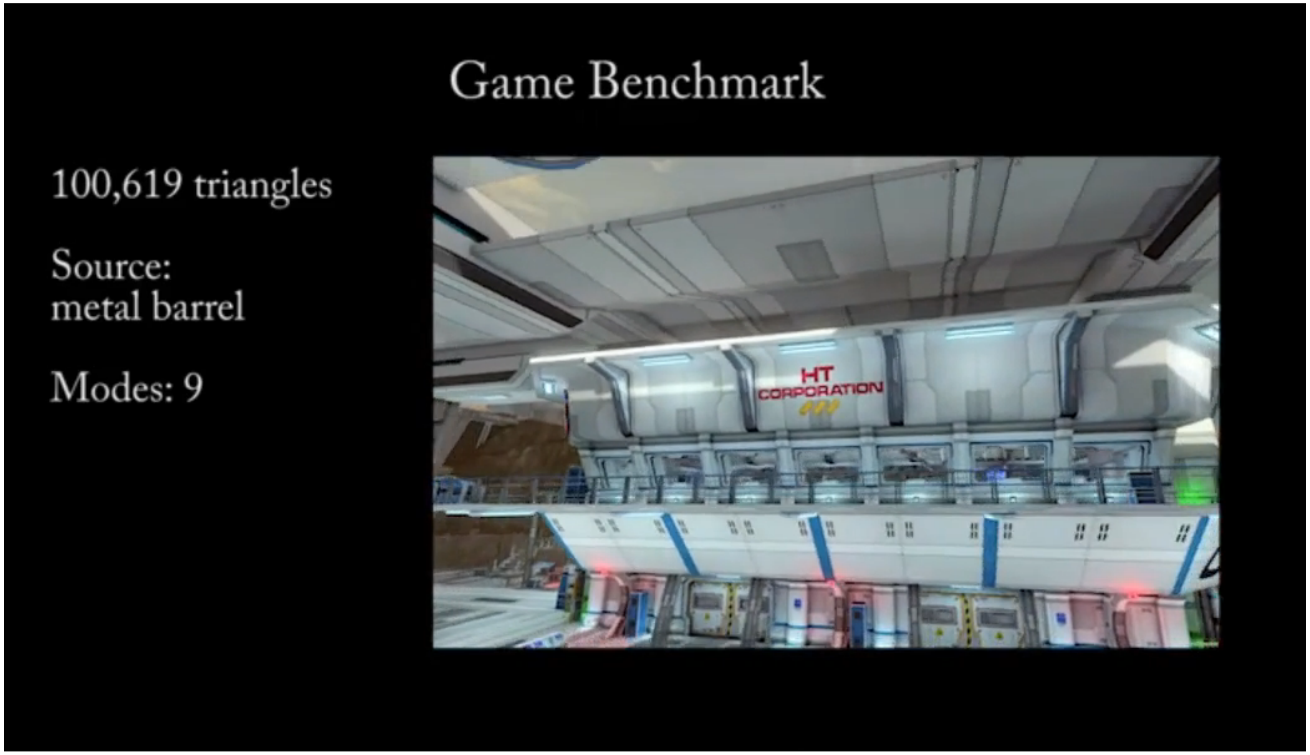


Audio-Visual Research Areas

- Sound synthesis and acoustics
- Sound separation (e.g. speech or noise filtering)
- 3D reconstruction
- Fluid-structure interactions
- Cross-modal biometrics
  - Seeing Voice and Hearing Faces[1]
- Deep fake detection[2]

[1] Seeing Voice and Hearing Faces: Cross-modal biometric matching [Nagrani et al. 2018]
[2] Emotions Don't Lie: A Deepfake Detection Method using Audio-Visual Affective Cues [Mittal et al. 2020]

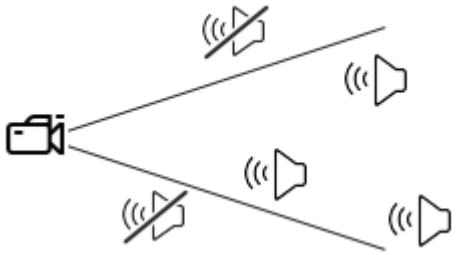
Sound Synthesis and Propagation in Virtual Environments



Source: **SynCoPation: Interactive Synthesis-Coupled Sound Propagation** [Rungta et al. 2016]

Audio-Visual Separation, Zooming, Localization, and More

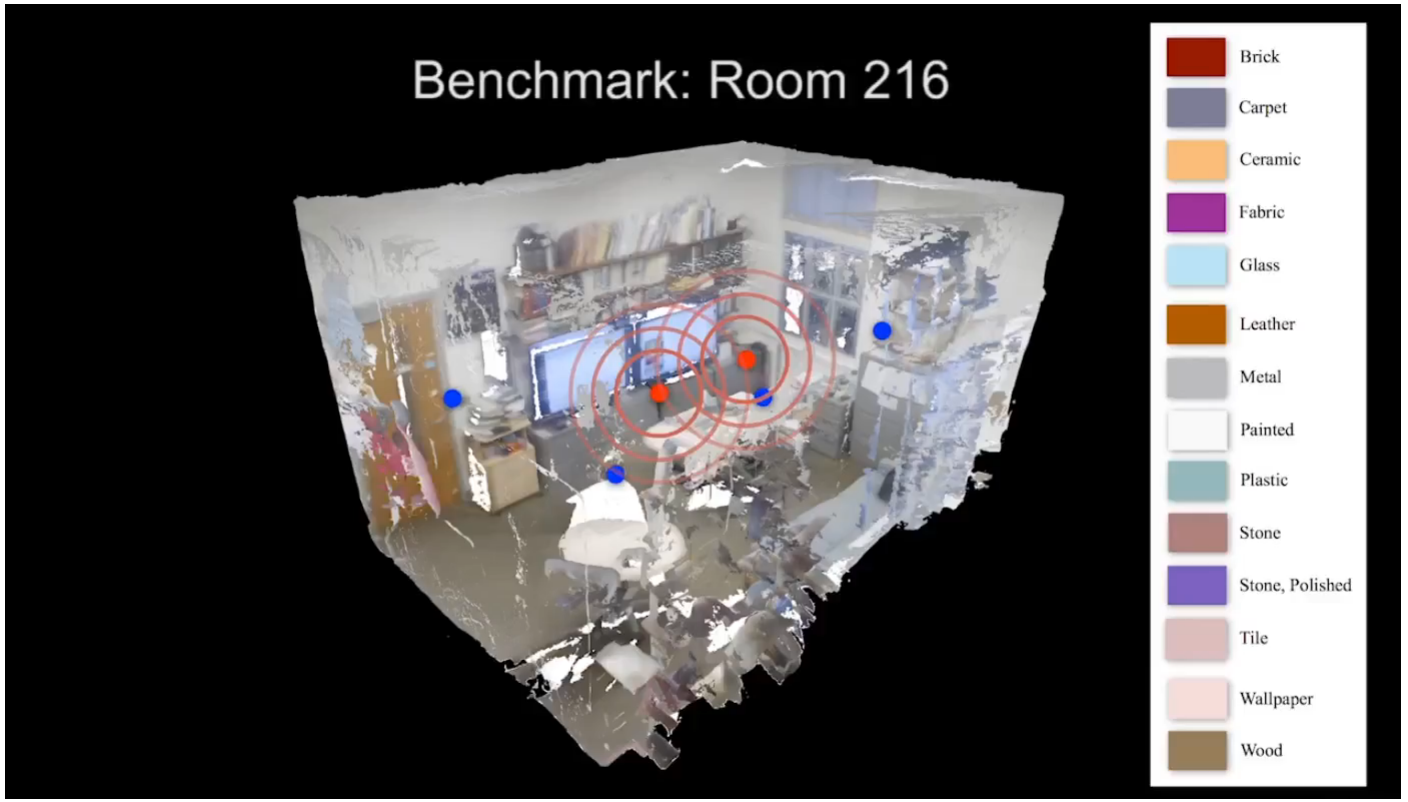
Source: Looking to Listen at the Cocktail Party[1]



Source: Audio-Visual Zooming[2]

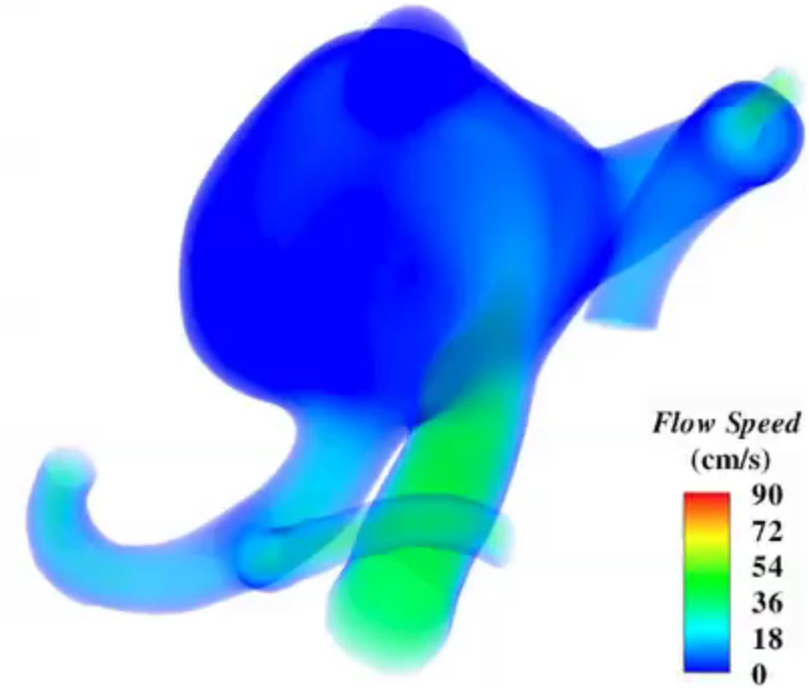
[1] Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation [Ephrat et al. 2018]
[2] Audiovisual Zooming: What You See Is What You Hear [Nair et al. 2019]

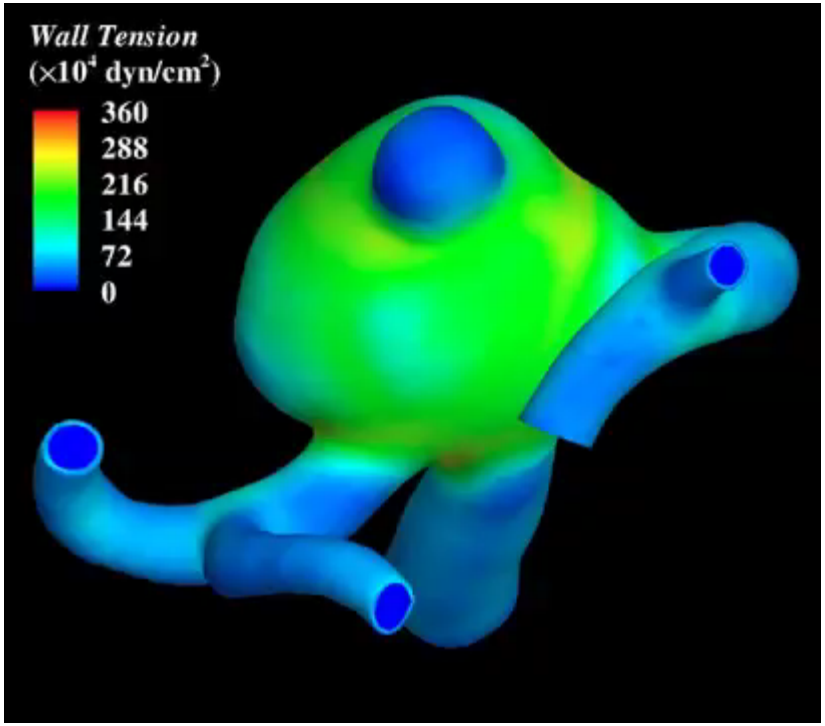
Enhanced Scene Reconstructions using Acoustic Optimization



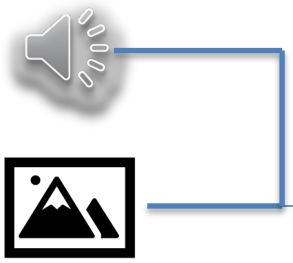
Source: **Acoustic Classification and Optimization for Multi-Modal Rendering of Real-World Scenes** [Schissler et al. 2017]

Coupling Fluid-Structure Interactions



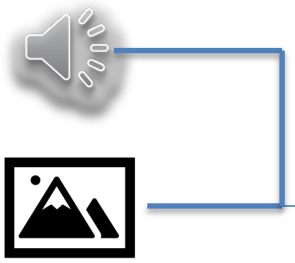


Source: **Computational Vascular Fluid–Structure Interaction** [Bazilevs et al. 2010]



Thesis Statement

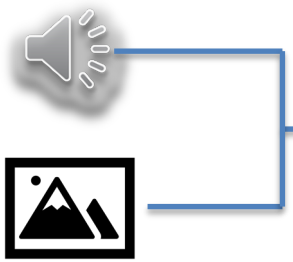
- Coupling multimodal information enhances task performance and processing of audio-visual learning based methods for **fluid-structure sound synthesis**, **liquid pouring sequences**, **object tracking**, and **3D reconstructions** while also allowing for single mode application for special cases.



Outline

- I: Introduction
- II: Fluid-Structure Sound Synthesis
- III: Liquid Pouring Sequences
- IV: Audio-Visual Object Tracking
- V: Audio-Augmented 3D Reconstruction
- VI: Conclusions

Part II



Fluid-Structure Sound Synthesis

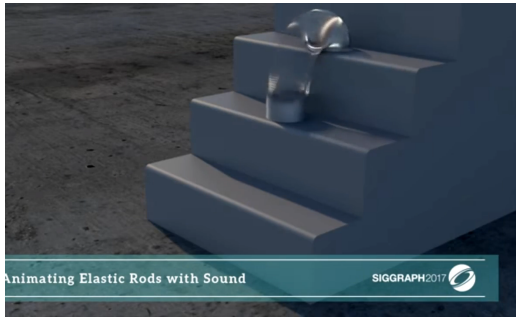
"Glass half full: sound synthesis for fluid-structure coupling using added mass operator." Justin Wilson, Auston Sterling, Nicholas Rewkowski, Ming C. Lin. Computer Graphics International (CGI) 2017, The Visual Computer.

Motivation for Sound Synthesis

- Animating fire with sound
  - Chadwick et al. 2011

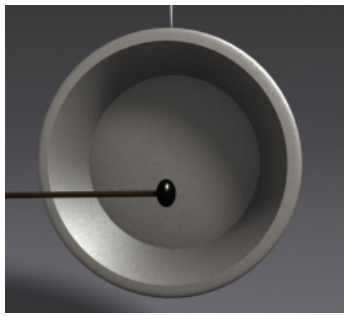
- Sounding liquids
  - Moss et al. 2010

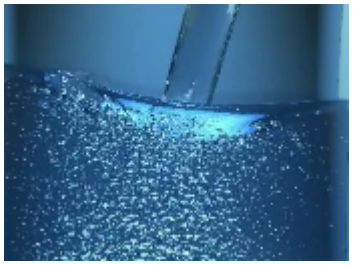
- Animating elastic rods with sound
  - Schweickart et al. 2017

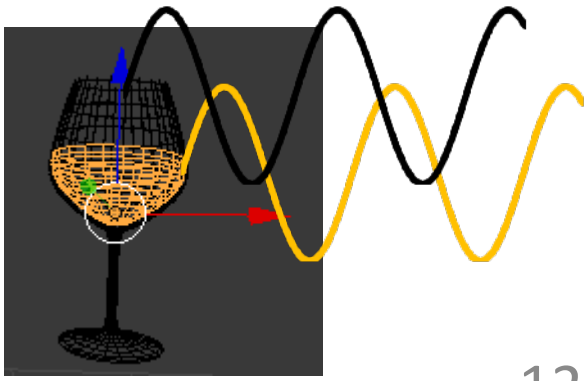


Challenges

- Previous research focuses on single systems only, either solid or liquid but not both
- Not all sound simulations achieve real-time performance
- Computationally expensive to model the variation in sound pressure waves from a coupled vibrating fluid-structure system

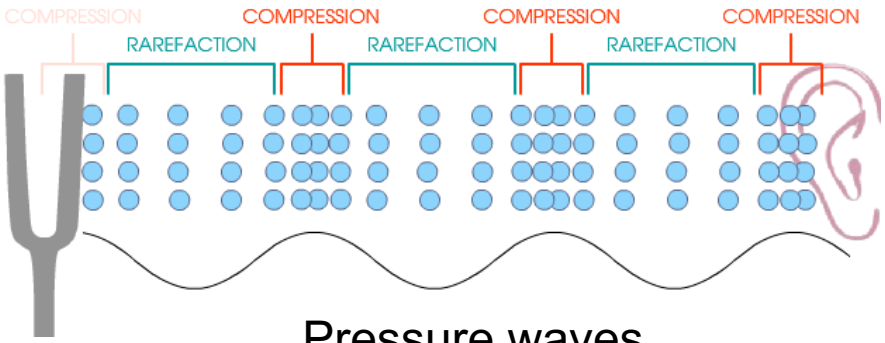






Background: Sound Waves

- When an object is struck, it vibrates and deforms
- Sound is generated by the vibration of pressure waves through a medium (e.g. air) and perceived by the ears as sound
- This harmonic motion can modeled as an underdamped spring mass system with $mx'' + dx' + kx = 0$
  - m = mass
  - d = damping
  - k = stiffness
  - x = displacement



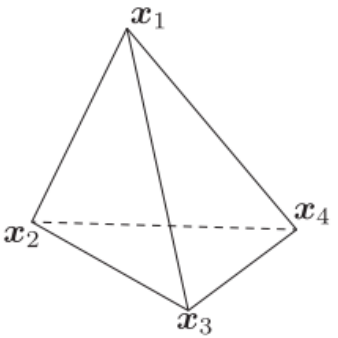
Pressure waves

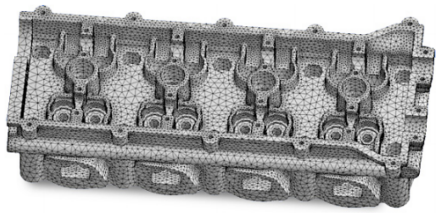
Source: digitalsoundandmusic.com

Background: Sound Dynamics Equation and Object Representation

- Object is a volumetric solid body represented by a finite element mesh
  - Tetrahedral mesh has n nodes where i=1…n
- Mu'' + Du' + Ku = f(t)
  - High-dimensional equivalent to spring-mass $mx''(t) + dx'(t) + kx(t) = f(t)$
- M, D, and K are size 3n x 3n sparse matrices
  - M = mass matrix
  - u = displacement vector of each element
  - D = viscous damping matrix
  - K = stiffness matrix
  - f = vector of forces

Single tetrahedron of finite element mesh



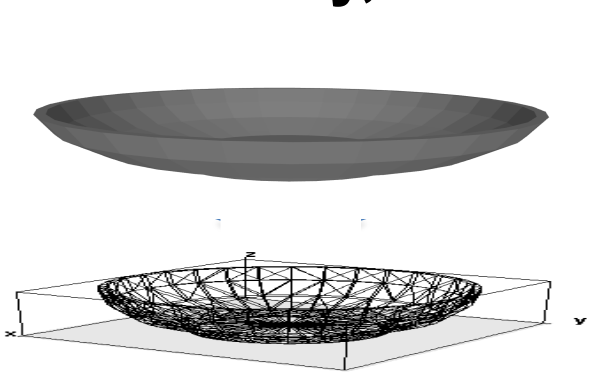


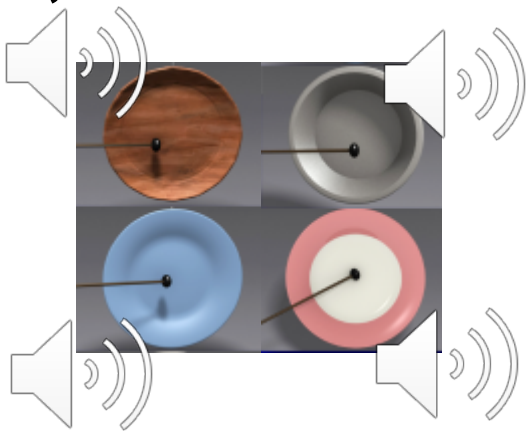
Source: SIGGRAPH 2016 Course Notes

Related Work: Modal Analysis and Sound Synthesis

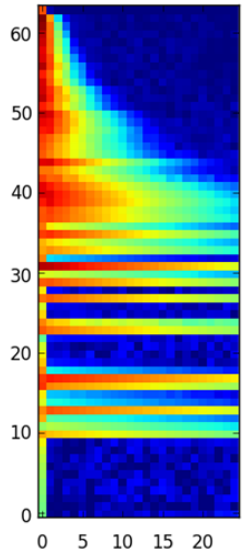
- Solution of sound dynamic equation is damped sinusoids
  - $q_i = a_i e^{-d_i t} \sin(2\pi f_i t + \theta_i)$ where a_i depends on run-time impulse and d_i and f_i depend on the geometry/material

Geometry, Material, and Hit Point





Sound



[1] Example-Guided Physically Based Modal Sound Synthesis [Ren et al. 2013]
Source: SIGGRAPH 2016 Course Notes

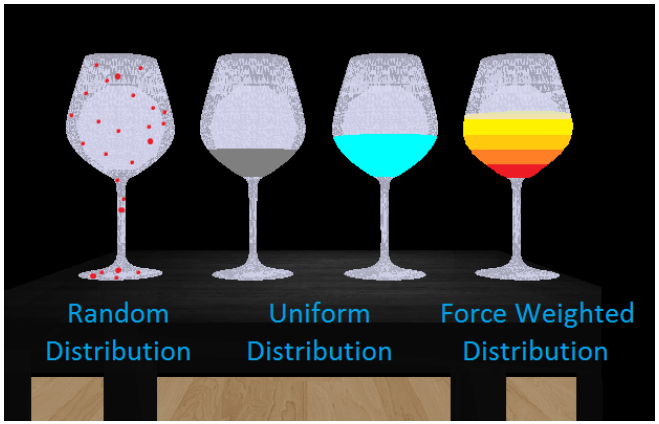
Contribution: Added Mass Operator

- Since the liquid must move with the same phase as the structure's motion, this may be referred to as a rigid double body[1]
- Added Mass is the additional (drag) force resulting from fluid acting on a structure

$$M\ddot{u} + D\dot{u} + Ku = f - m_a\ddot{u}$$

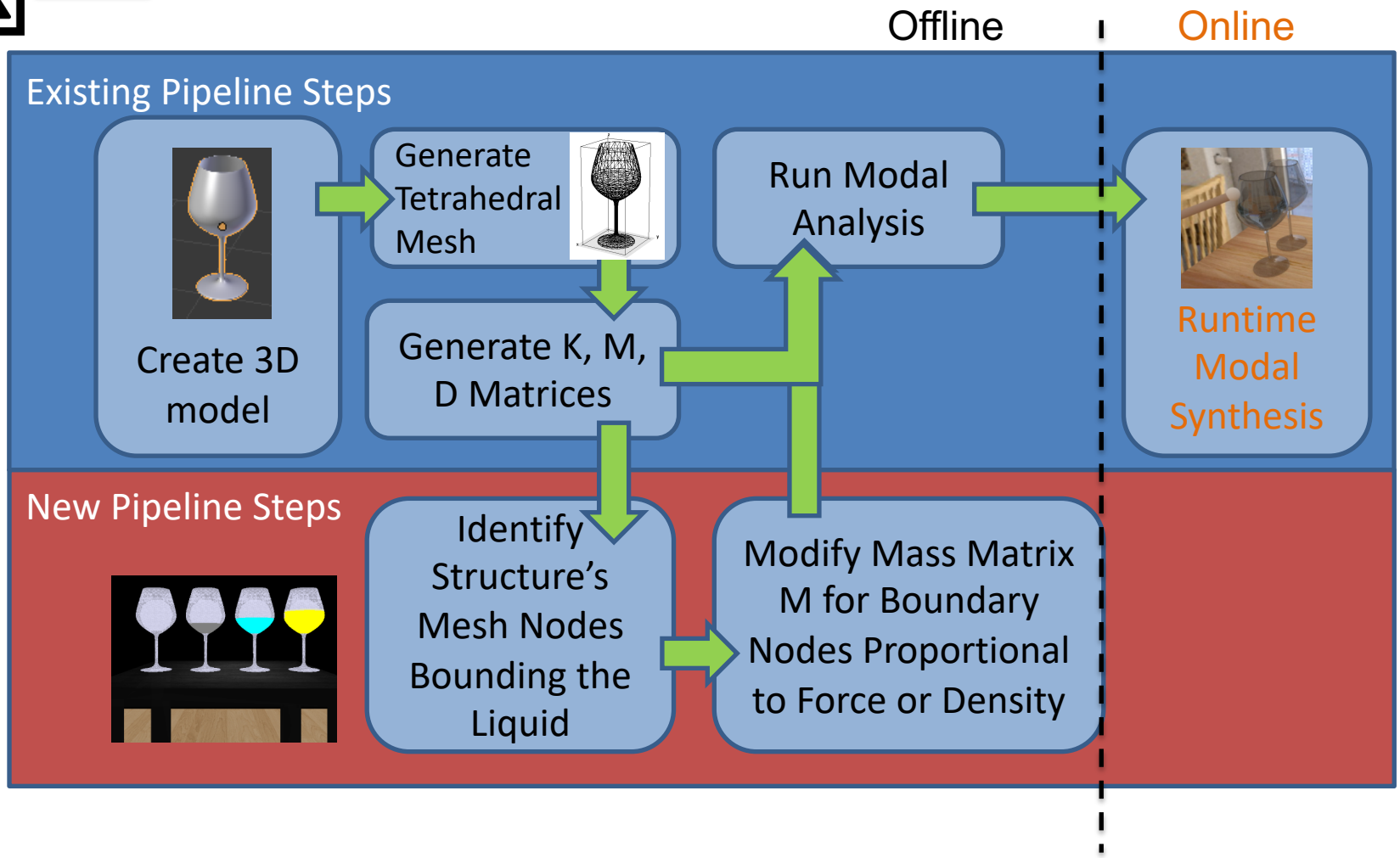
$$(M + m_a)\ddot{u} + D\dot{u} + Ku = f$$

$$m_a = \rho_{fluid} \cdot volume_{fluid}/2$$



[1] Marine Hydrodynamics [Newman 1977]

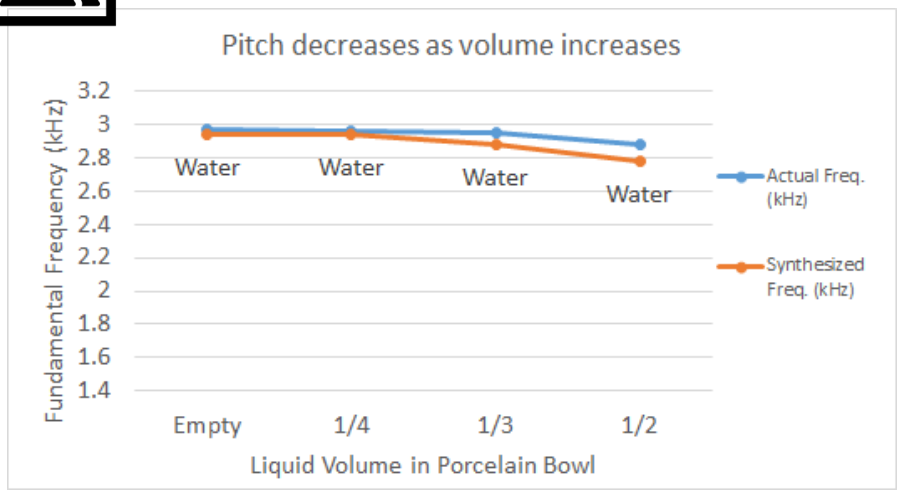
System Overview

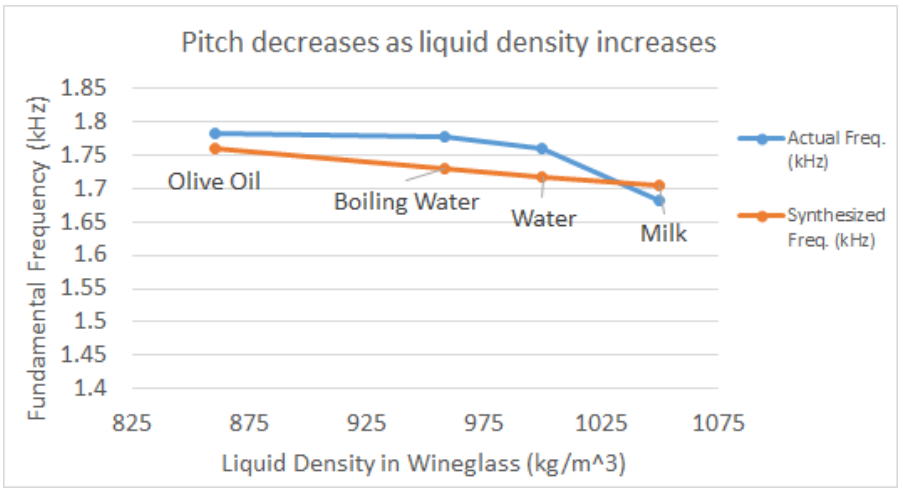


Real Water Xylophone

Simulated Water Xylophone

Validation Results < 5% Relative Error

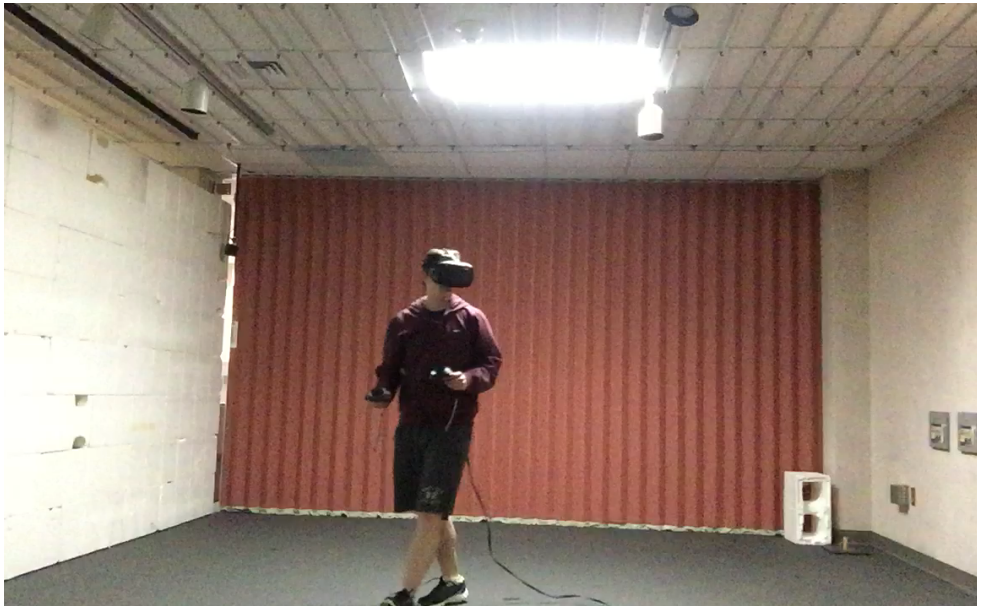


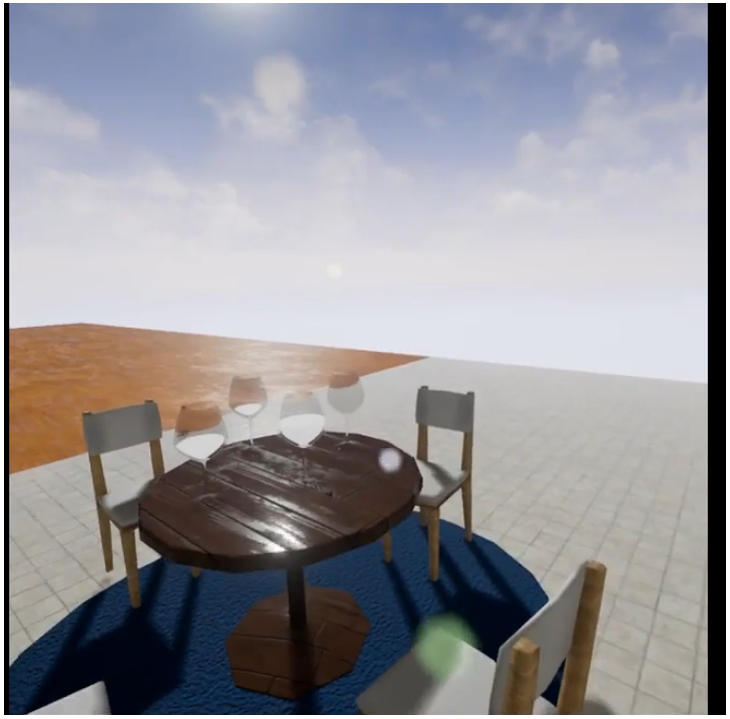


Vol	Syn freq (kHz)	Act freq (kHz)	Rel error (%)
Empty	2.9419	2.9709	0.98
1/4	2.9389	2.9597	0.70
1/3	2.8816	2.9453	2.16
1/2	2.7810	2.8759	3.30

Vol	Syn freq (kHz)	Act freq (kHz)	Rel error (%)
Milk	1.7037	1.6823	1.27
Water	1.7176	1.7607	2.45
Hot water	1.7297	1.7771	2.67
Olive oil	1.7597	1.7824	1.28

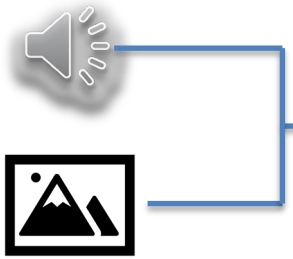
System Demonstration in VE



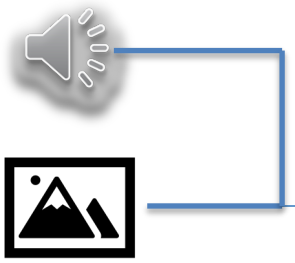


Part III

Analyzing Liquid Pouring Sequences with Audio-Visual



"Analyzing liquid pouring sequences via Audio-Visual Neural Networks." Justin Wilson, Auston Sterling, Ming C. Lin. International Conference on Intelligent Robots and Systems (IROS) 2019.

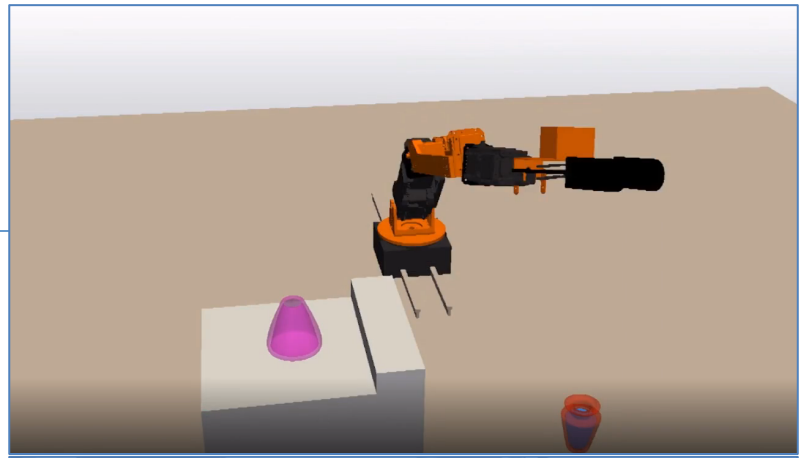


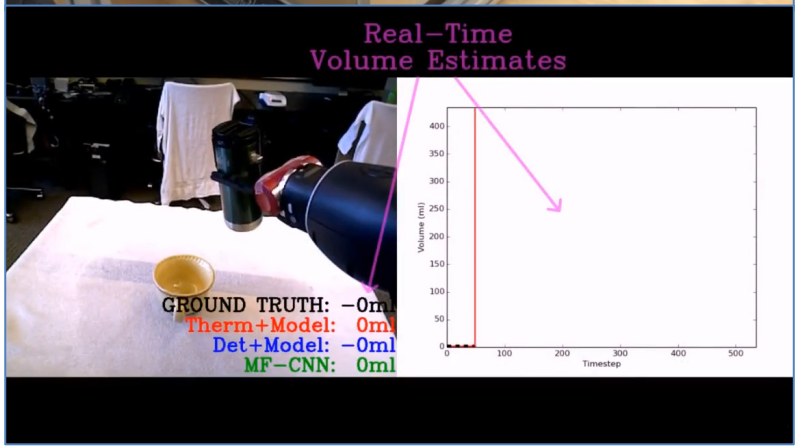
Outline

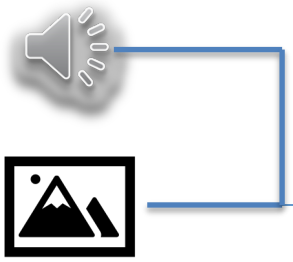
- I: Introduction
- II: Fluid-Structure Sound Synthesis
- III: Liquid Pouring Sequences
- IV: Audio-Visual Object Tracking
- V: Audio-Augmented 3D Reconstruction
- VI: Conclusions

Liquid Pouring

- Motion planning
  - Robot Motion Planning for Pouring Liquids, Pan et al. 2016
- Learning based methods
  - Learning Audio Feedback for Estimating Amount and Flow of Granular Material, Clarke et al. 2018
- Visual control
  - Visual Closed-Loop Control for Pouring Liquids, Schenck et al. 2017



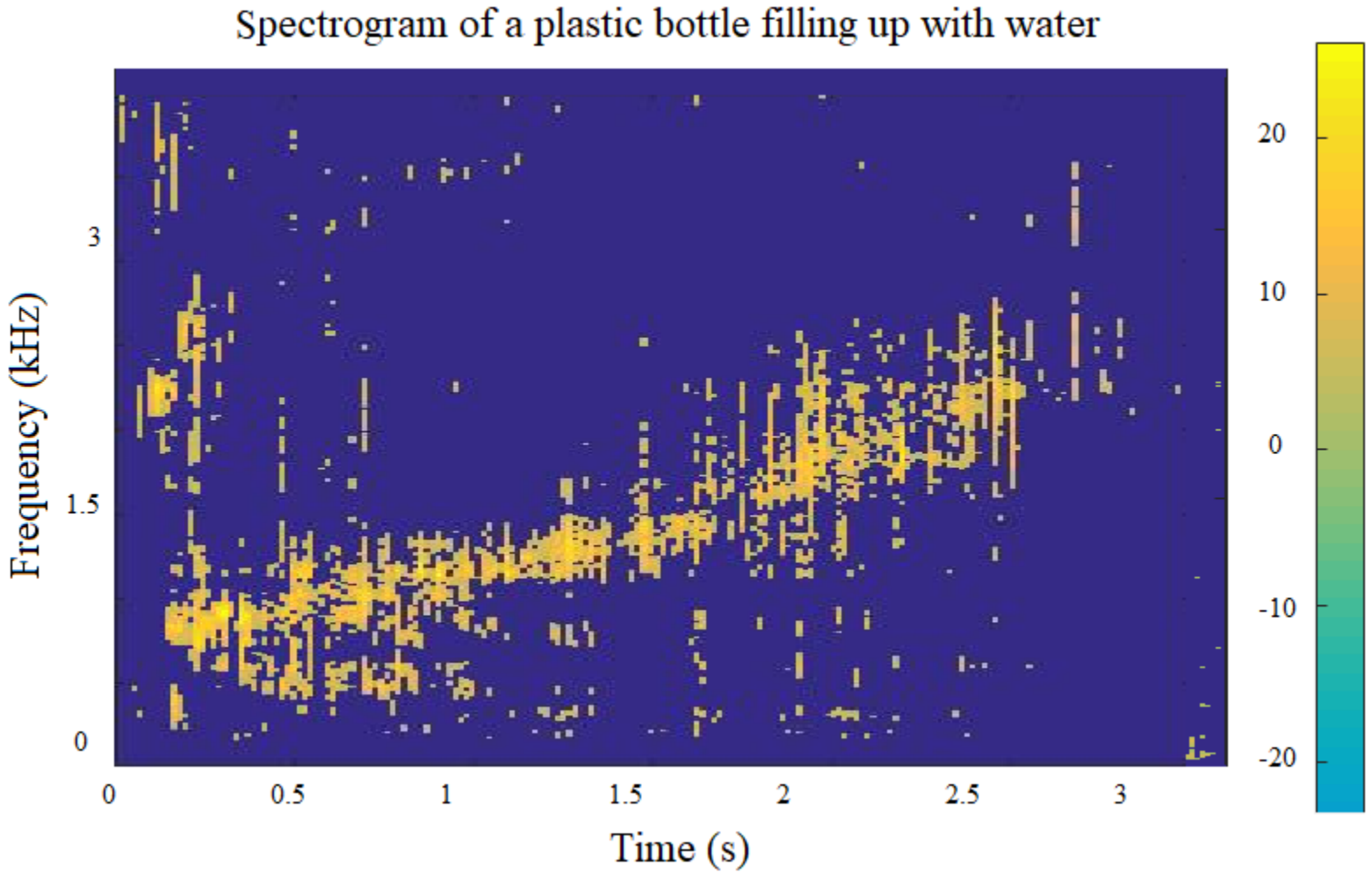




Extensions to Existing Work

- In addition to RGB and contact microphones, how can we also use audio from pouring sequences to estimate poured amount?
- How can we augment audio based learning by fusing with its associated visual data?
- How can we generate audio-visual ground truth data that is easy to replicate and available for future research?

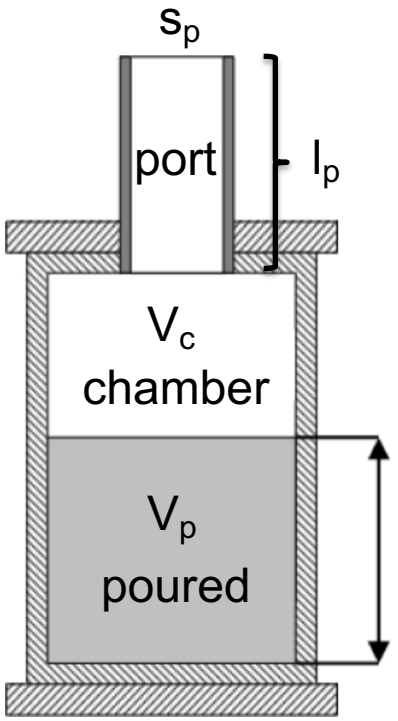
Example Spectrogram of Liquid Pouring Sequence



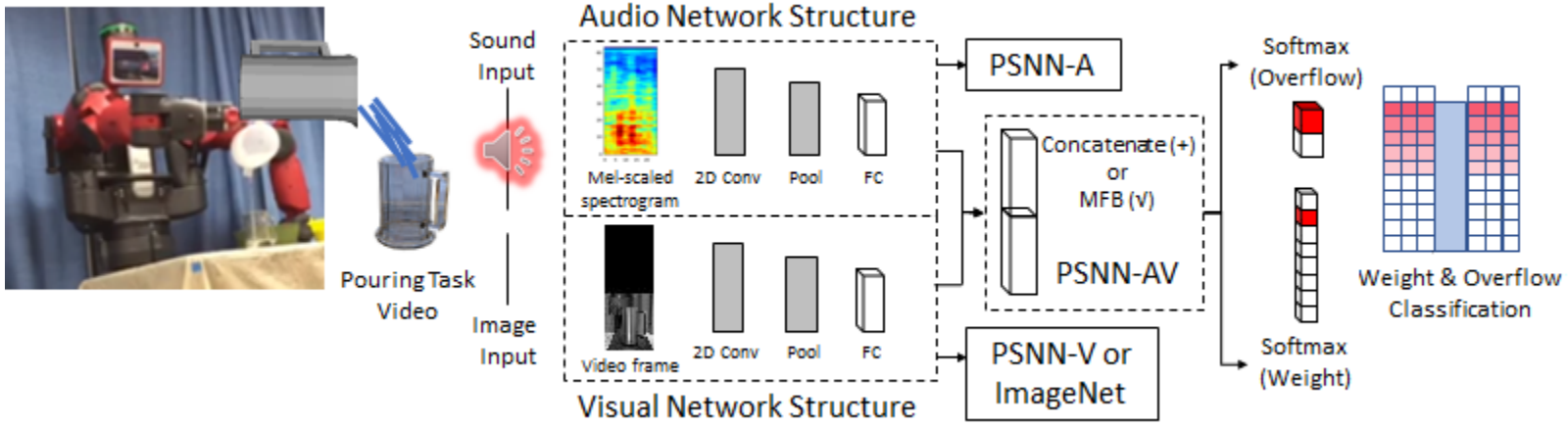
Helmholtz Resonance Frequency

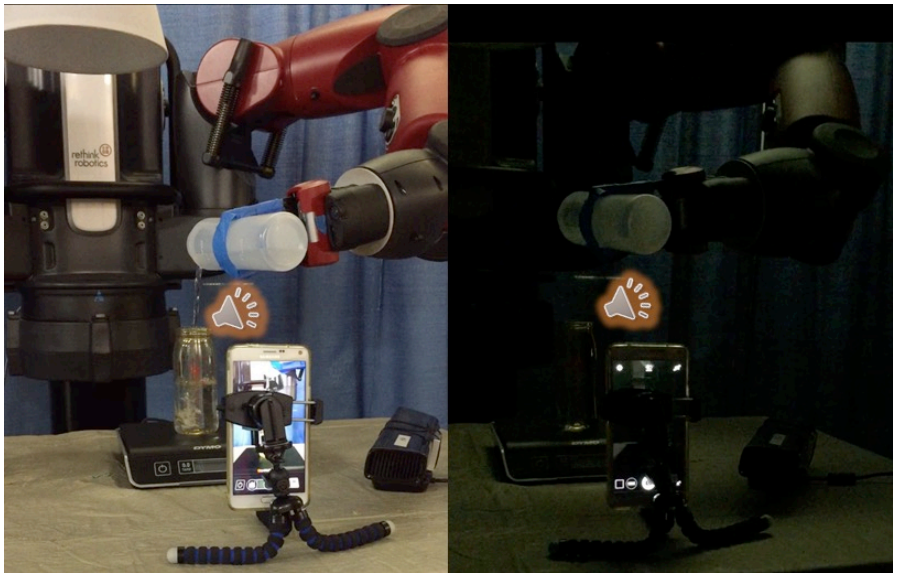
- Audio frequency increases as a container fills up with liquid
- This increase in frequency can be modeled based on the Helmholtz resonance (also referred to as a resonant cavity)
- This resonant frequency f_{res} and placed/poured volume

$$f_{res} = \frac{c}{2\pi}\sqrt{\frac{s_p}{V_c l_p}} \qquad V_p = Vc - \frac{s_p}{l'_p\left(\frac{2\pi fres}{c}\right)^2}$$

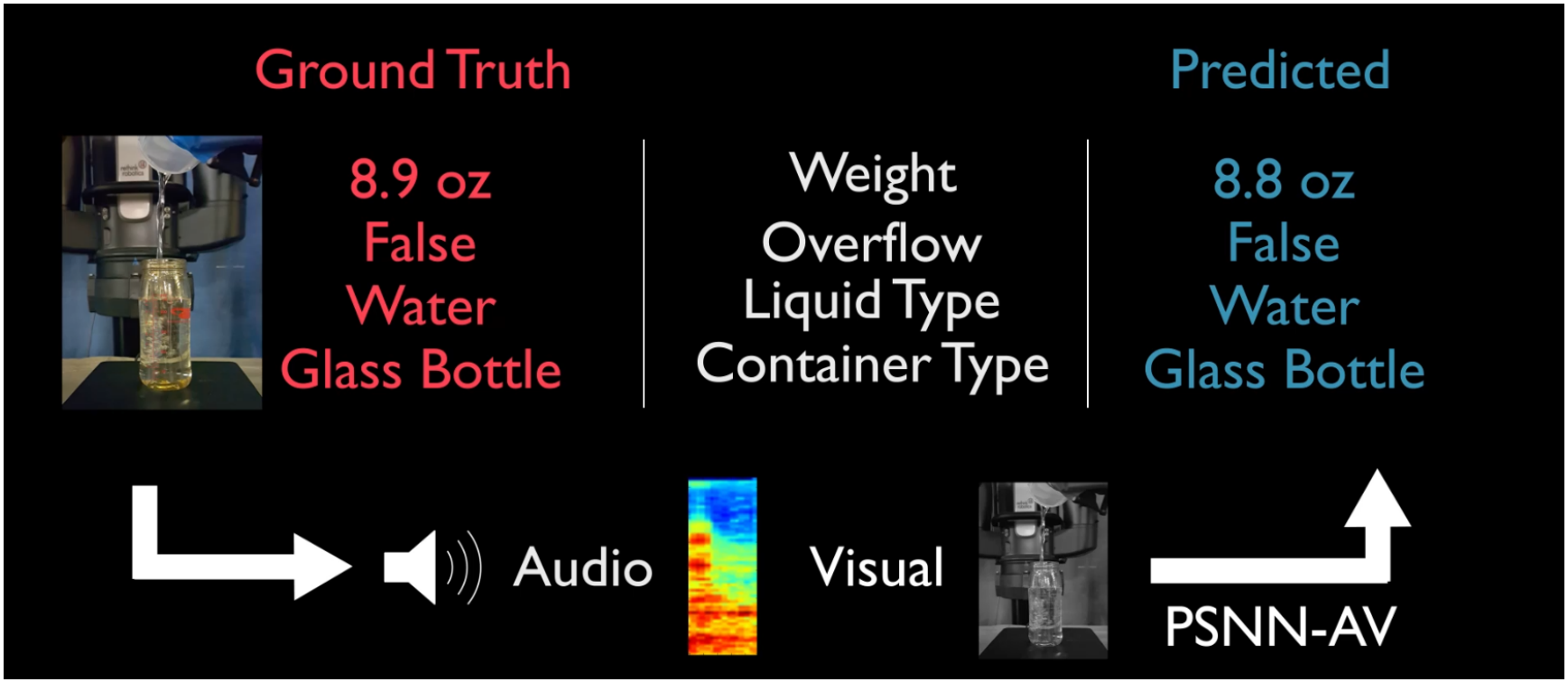


Analyzing Liquid Pouring Sequences via Audio-Visual Neural Networks

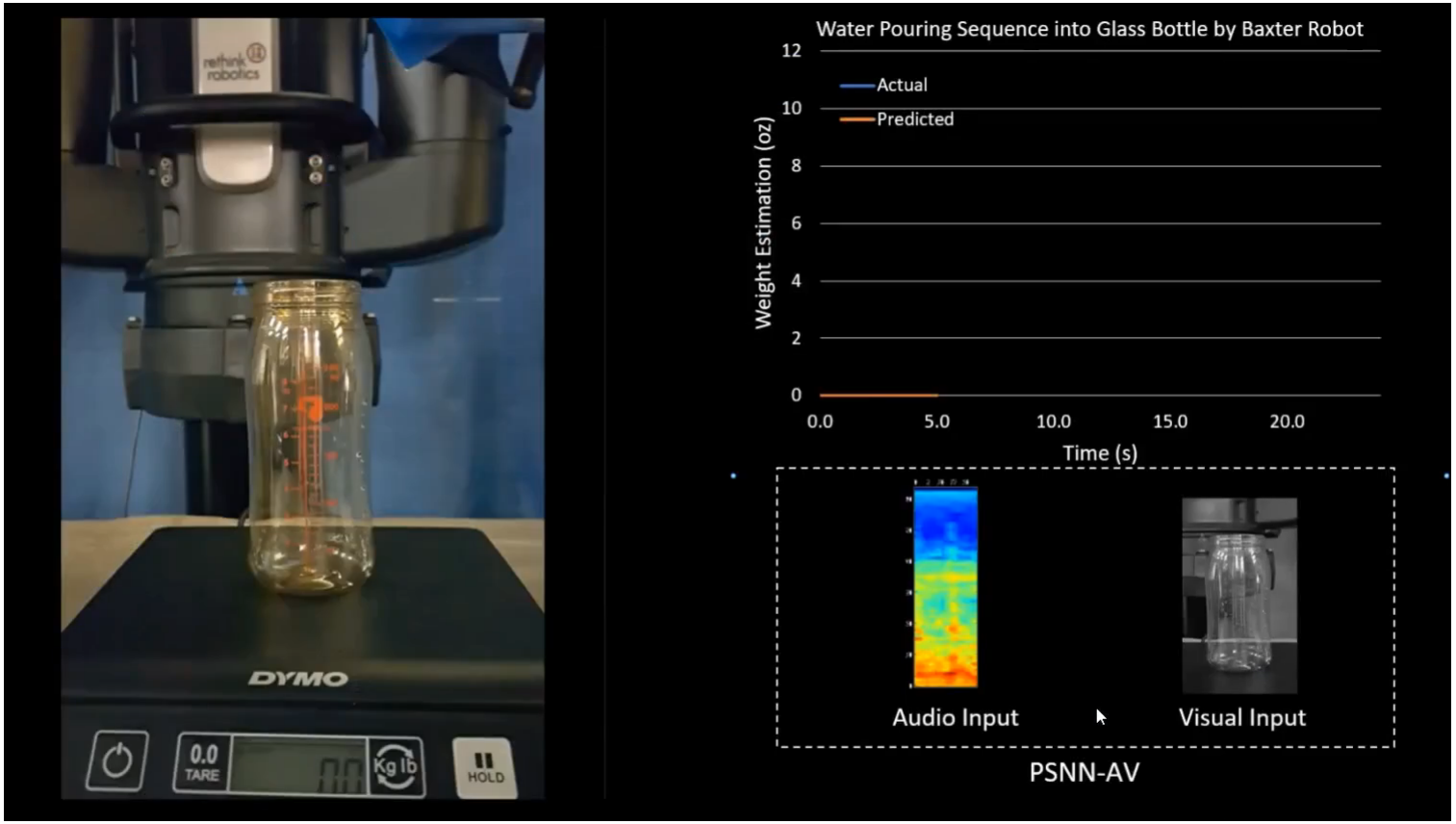




Pouring Sequence Neural Network – Audio-Visual (PSNN-AV)



PSNN-AV Robot Pouring Sequence (Water)



Our Results vs. Baseline

- Audio only PSNN-A correctly classified 88.0% for robot and 75.8% for human sequences respectively
- Audio-augmented PSNN-AV correctly classified poured weight within 0.4 oz for up to 91.5% for robot and 86.4% for human pouring sequences

Weight Estimation and Overflow Detection Accuracy by Method for Robot and Human Experimenter Water Pouring Sequences

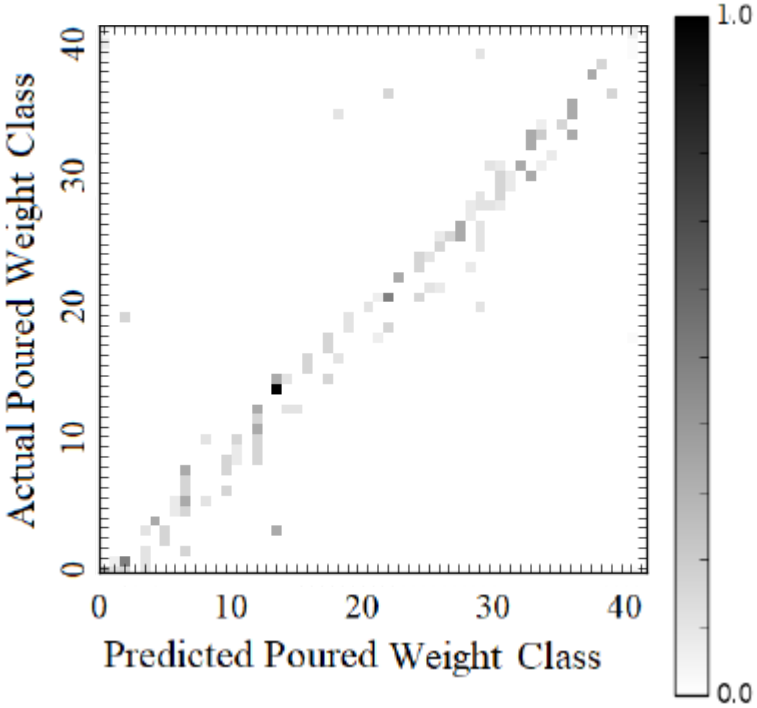
		Glass Bottle, Robot Pour, N=20			Glass Bottle, Human Pour, N=20			Combined Container Dataset, N=40		
Method	Input	+/- 0.4 oz	Ave Err	Overflow	+/- 0.4 oz	Ave Err	Overflow	+/- 0.4 oz	Ave Err	Overflow
kNN [11]	A	66.4%	1.9 oz	71.9%	54.2%	2.7 oz	62.5%	58.8%	2.4 oz	77.1%
Linear SVM [5]	A	4.6%	3.8 oz	50.0%	13.6%	4.3 oz	50.0%	12.7%	4.0 oz	60.4%
SoundNet5 [3]	A	46.0%	1.9 oz	50.0%	42.4%	3.6 oz	50.0%	21.2%	3.3 oz	50.0%
SoundNet8 [3]	A	11.2%	3.3 oz	50.0%	29.2%	4.7 oz	50.0%	35.4%	4.4 oz	50.0%
TCN [26]	A	78.4%	0.9 oz	50.0%	40.1%	3.7 oz	50.0%	49.6%	2.6 oz	50.0%
PSNN-A (Ours)	A	**88.0%**	**0.5 oz**	**78.1%**	**75.8%**	**1.9 oz**	**64.3%**	**80.8%**	**1.3 oz**	**83.3%**
ImageNet [23]	V	**83.8%**	**0.3 oz**	–*	**71.2%**	**0.4 oz**	–*	68.1%	1.1 oz	–*
PSNN-V (Ours)	V	79.9%	0.6 oz	–*	66.5%	0.6 oz	–*	**78.0%**	**0.4 oz**	–*
PSNN-AV Cat (Ours)	AV	**91.5%**	**0.2 oz**	–*	**86.4%**	**0.2 oz**	–*	82.0%	0.3 oz	–*
PSNN-AV MFB (Ours)	AV	88.8%	**0.2 oz**	–*	71.2%	2.1 oz	–*	**86.7%**	**0.2 oz**	–*

1 oz = 29.5735 ml

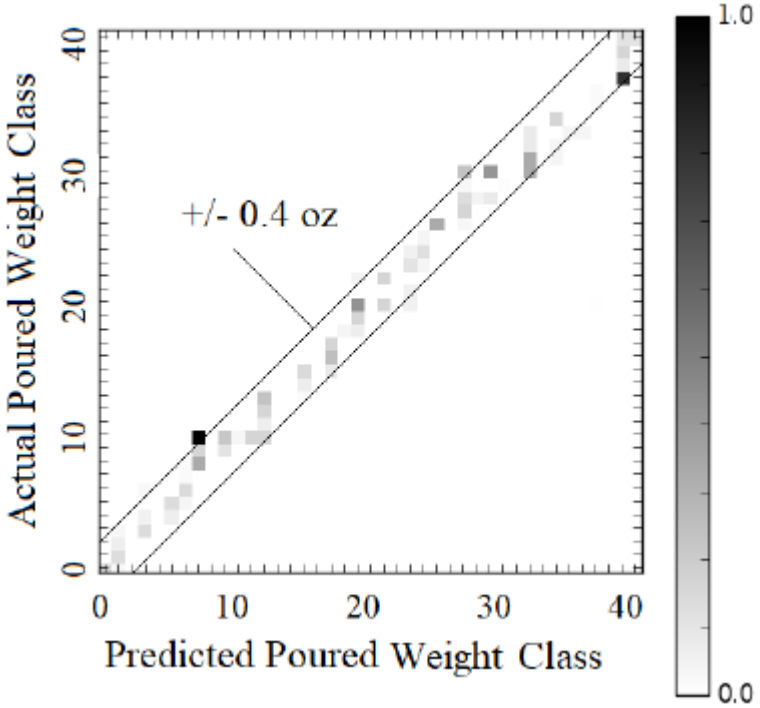
* Only audio-based neural networks were evaluated for overflow as visual information oversimplified the task

Example Confusion Matrices

PSNN-A Confusion Matrix

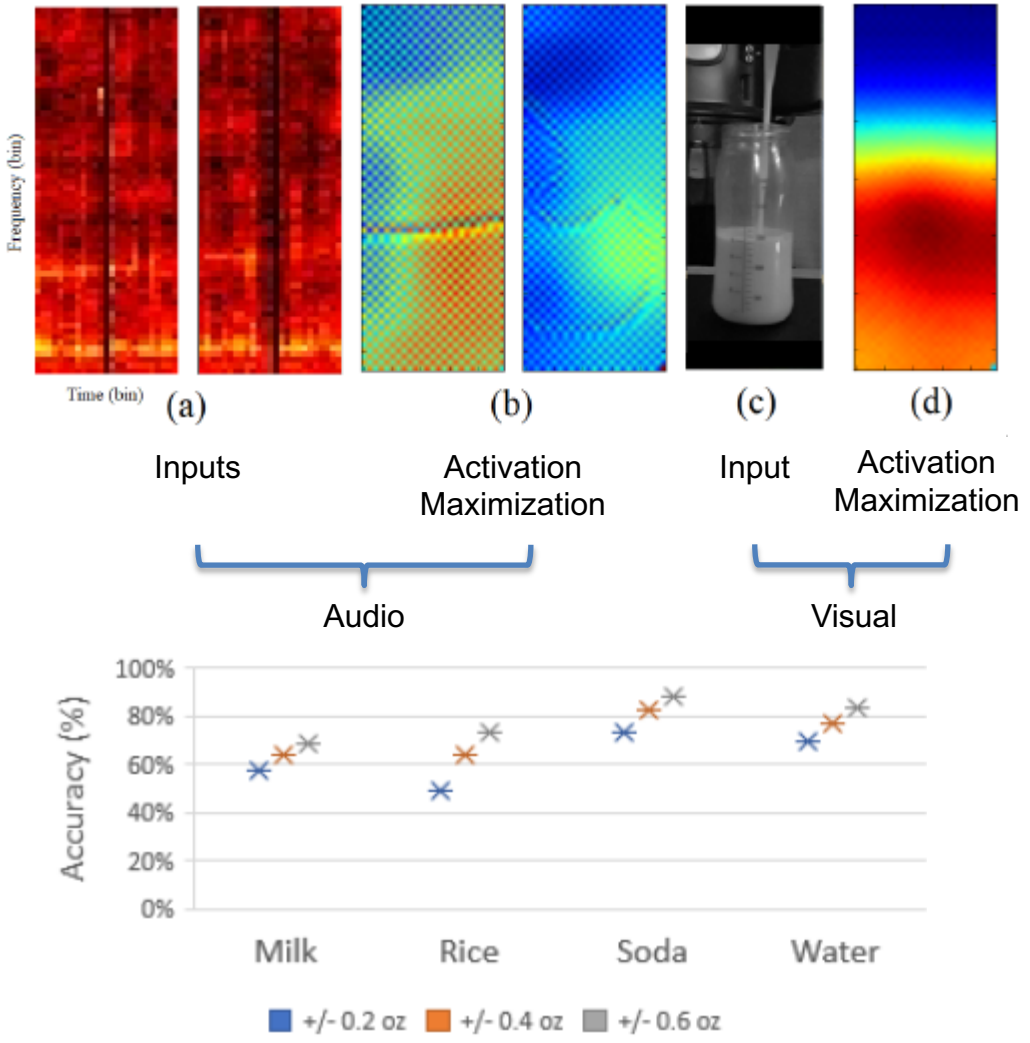


PSNN-AV Confusion Matrix

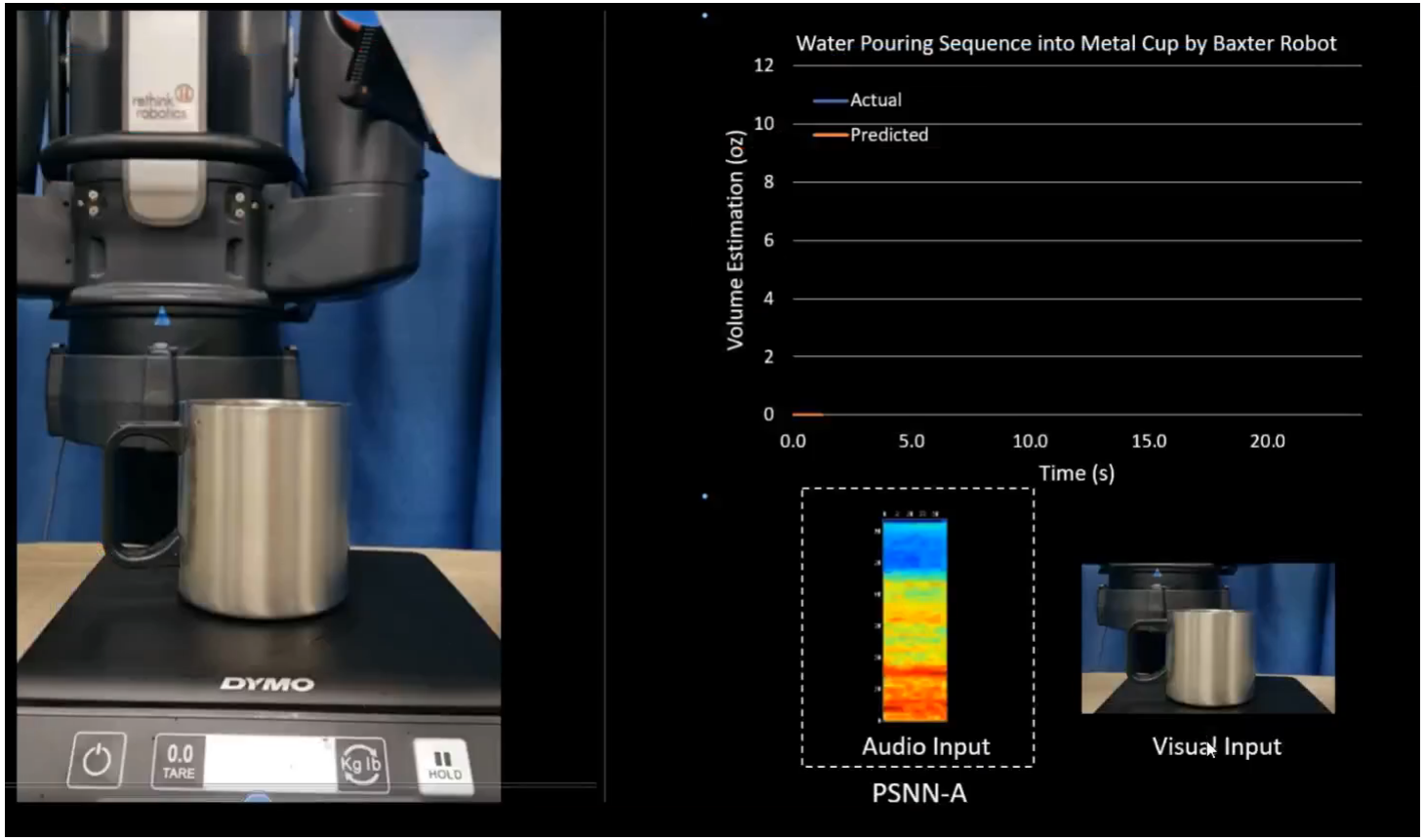


Activation Maximizations

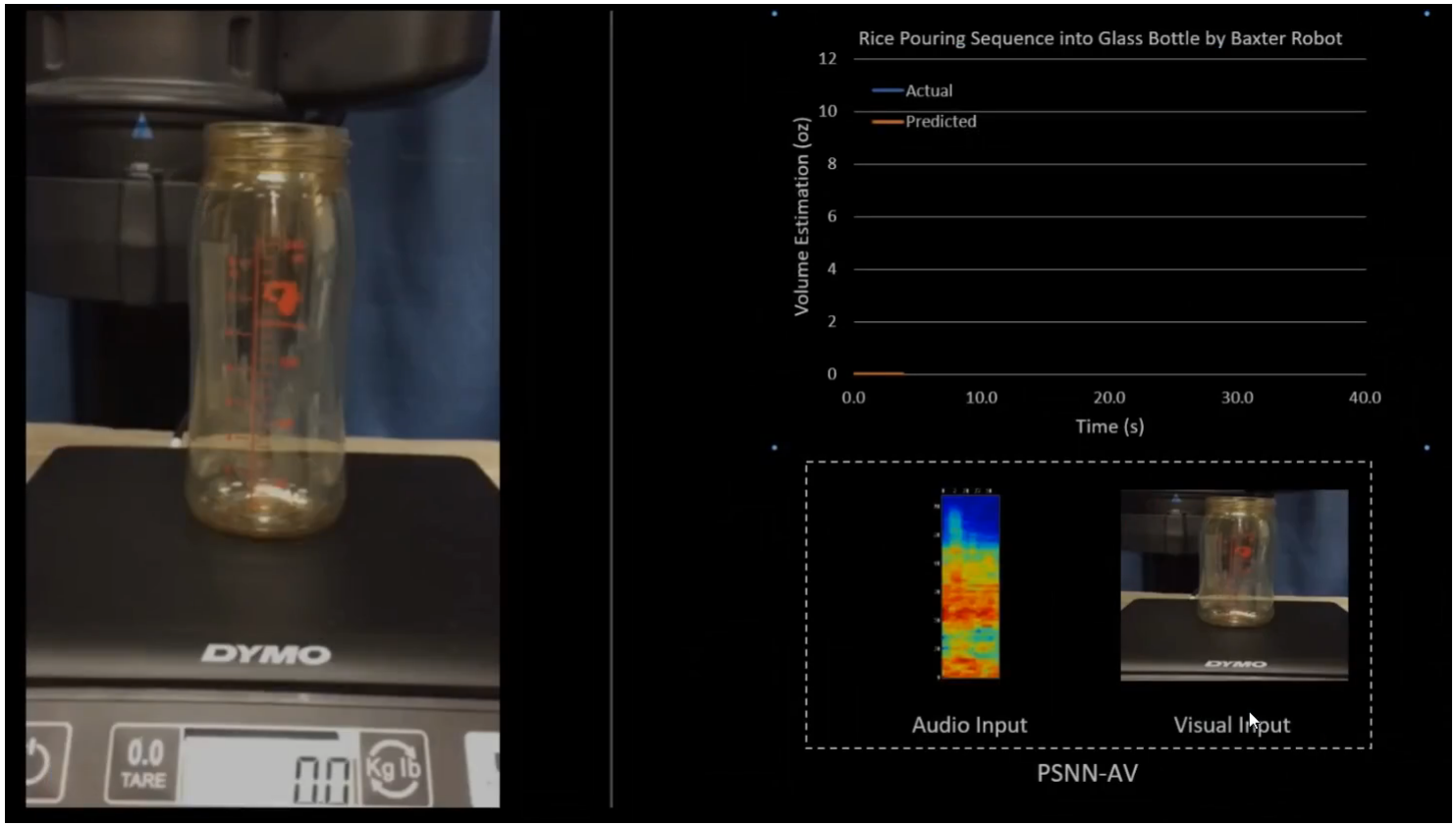
- Inputs that would maximize PSNN activations
  - Demonstrates ability to learn changes in frequency and height
- Various pouring contents evaluated with PSNN



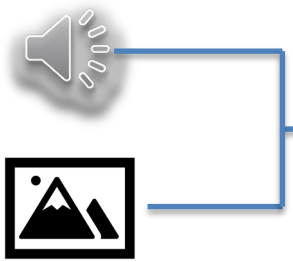
PSNN-A Pouring Sequence Demo



PSNN-AV Robot Pouring Sequence (Rice)

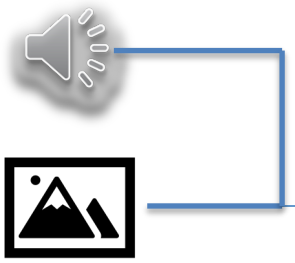


Part IV



Audio-Visual Object Tracking

"AVOT: Audio-Visual Object Tracking of Multiple Objects for Robotics." Justin Wilson and Ming C. Lin. International Conference on Robotics and Automation (ICRA) 2020.



Outline

- I: Introduction
- II: Fluid-Structure Sound Synthesis
- III: Liquid Pouring Sequences
- IV: Audio-Visual Object Tracking
- V: Audio-Augmented 3D Reconstruction
- VI: Conclusions

Motivation

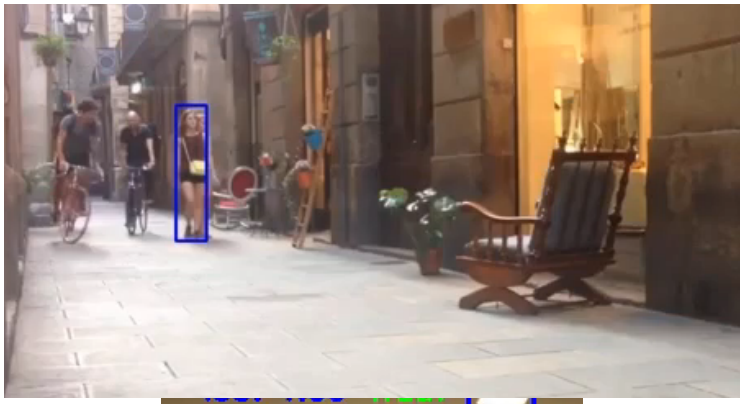
- Object tracking methods are used for:
  - Autonomous driving[1]
  - Mobile robotics[2]
  - Speaker recognition[3]

[1] Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite [Geiger et al. 2012]
[2] Tracking multiple moving objects with a mobile robot [Schulz 2001]
[3] Multi-speaker tracking from an audio-visual sensing device [Qian et al. 2019]

38

Challenges

- Tracking the challenging cases of:
  - Colliding or occluding objects
  - Similar object categories
  - Smaller objects

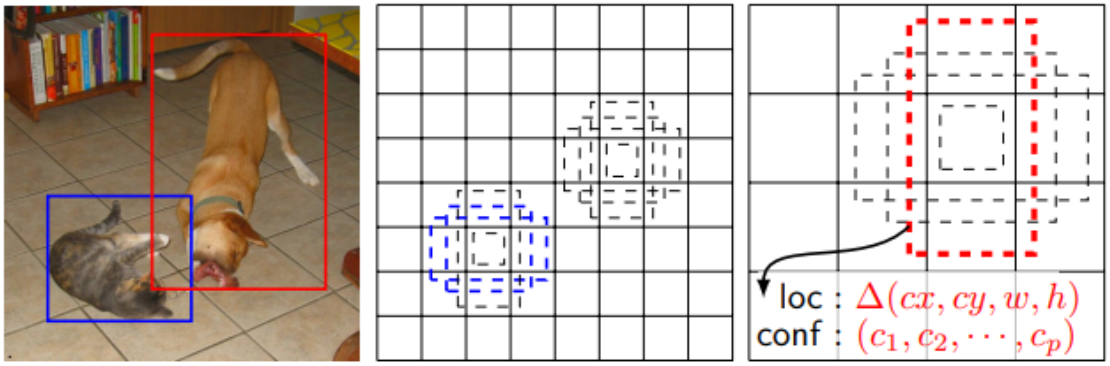


Top: https://www.pyimagesearch.com/2018/11/12/yolo-object-detection-with-opencv/
Middle: https://www.learnopencv.com/object-tracking-using-opencv-cpp-python/
Bottom: CSRT algorithm based on a virtual scene from Sound-20K [Zhang et al. 2017]

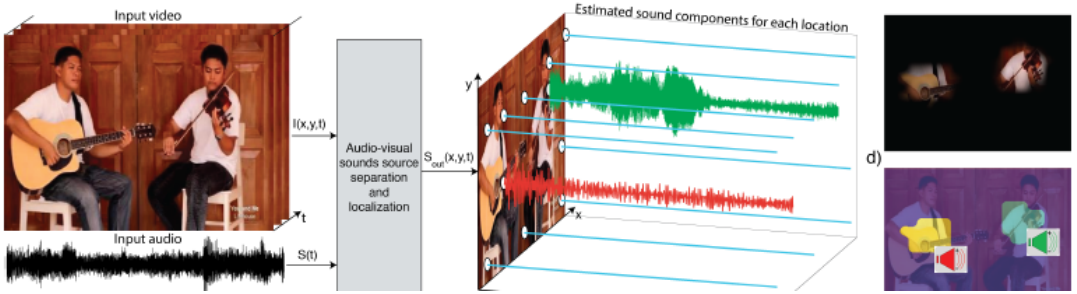
39

Related Work: Object Detection and Tracking

- Object Tracking
  - Frame skipping[1]
  - Deep Learning
    - Faster R-CNN[2]
    - YOLO[3]
    - SSD[4]
- Audio-Visual

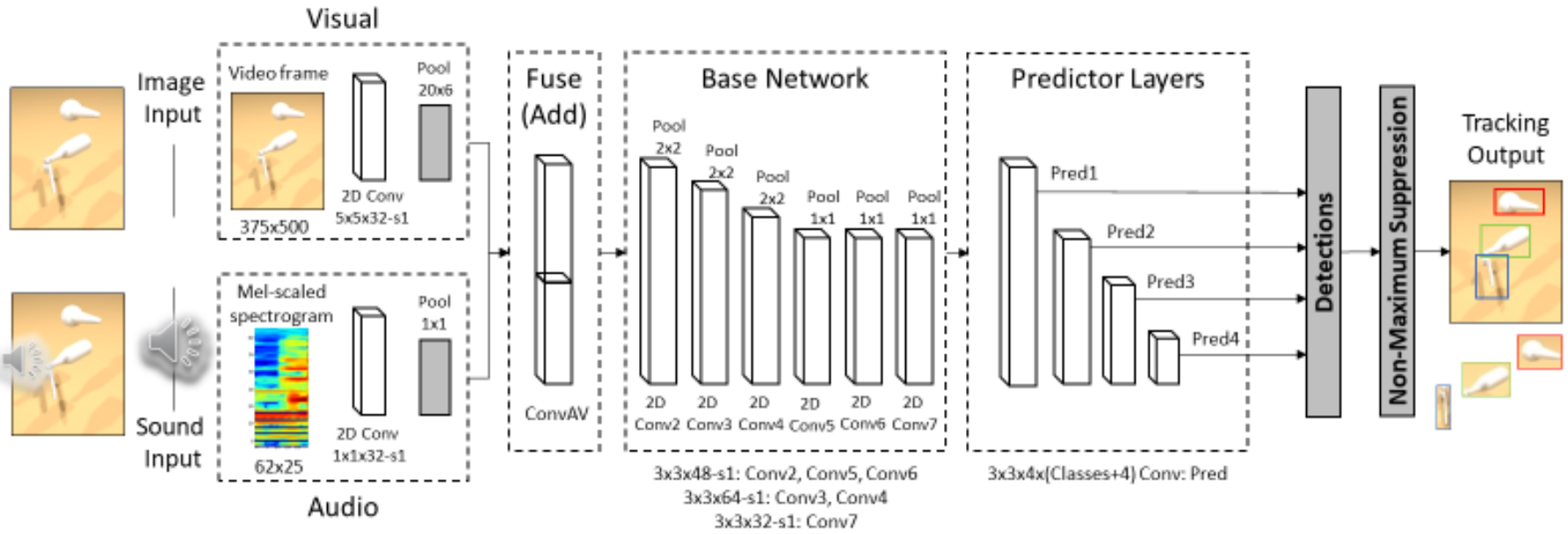


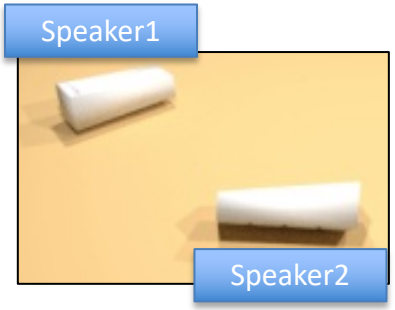
(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map



[1] W. Detect or Track: Towards Cost-Effective Video Object Detection/Tracking [Luo et al. 2018]
[2] Faster R-CNN: Towards Realtime Object Detection with Region Proposal Networks [Ren et al. 2015]
[3] You Only Look Once: Unified, Real-Time Object Detection [Redmon et al. 2015]
[4] SSD: Single Shot MultiBox Detector [Ren et al. 2015]
[5] The Sound of Pixels [Zhao et al. 2018]
[6] Look, Listen, and Learn [Arandjelovic et al. 2017]

Contribution: Audio-Visual Object Tracker (AVOT)

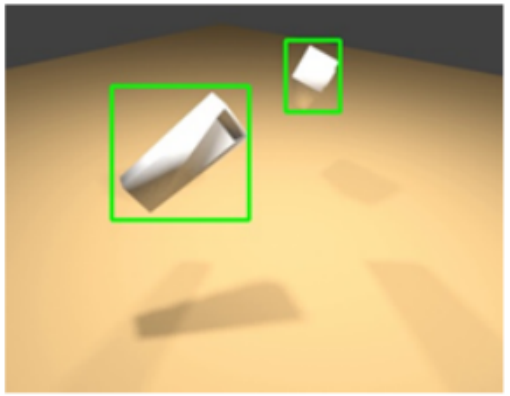




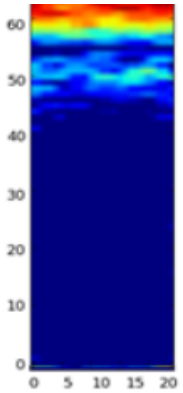
We define an object based on its geometry and material (e.g. glass, wood, etc.) to track objects with the same geometry but different materials

Annotated Audio-Visual Dataset

- Our experiments consist of bounding box annotated virtual scenes from the Sound20K audio-visual dataset
- 1,752 audio-visual segments
  - 18 objects (geometry/mat)
  - 103 frames per 3-sec video
- Training and test data



Ground truth bounding boxes



Mel-scaled spectrogram

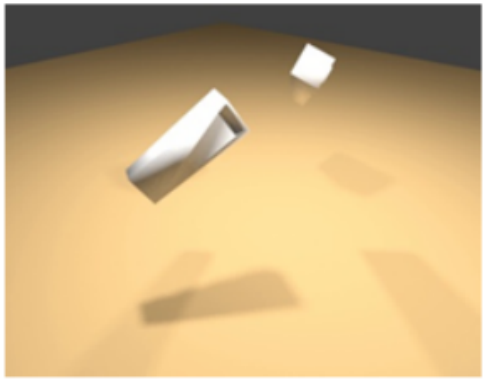
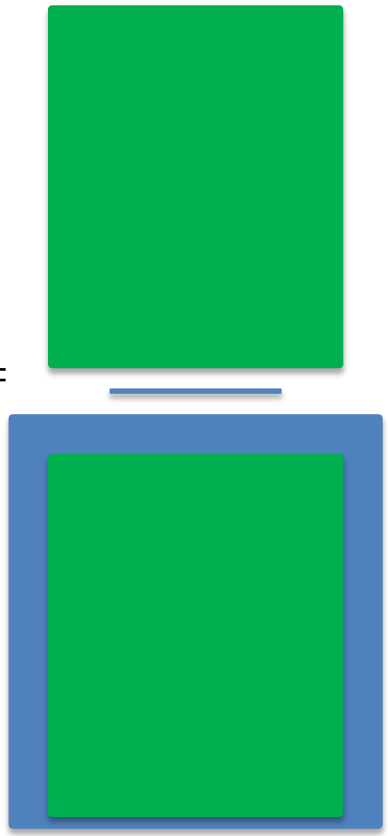
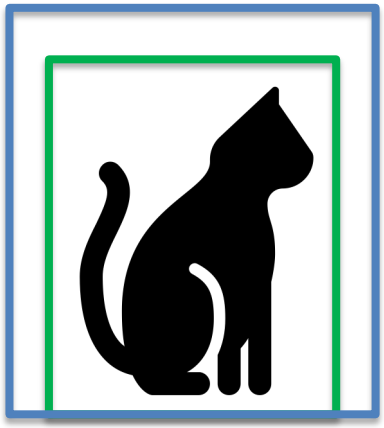


Image frame

Evaluation Metrics

- Frames per Second (FPS)
- Intersection over Union (IoU)
  - IoU = $\frac{\text{Area of Overlap}}{\text{Area of Union}}$ =



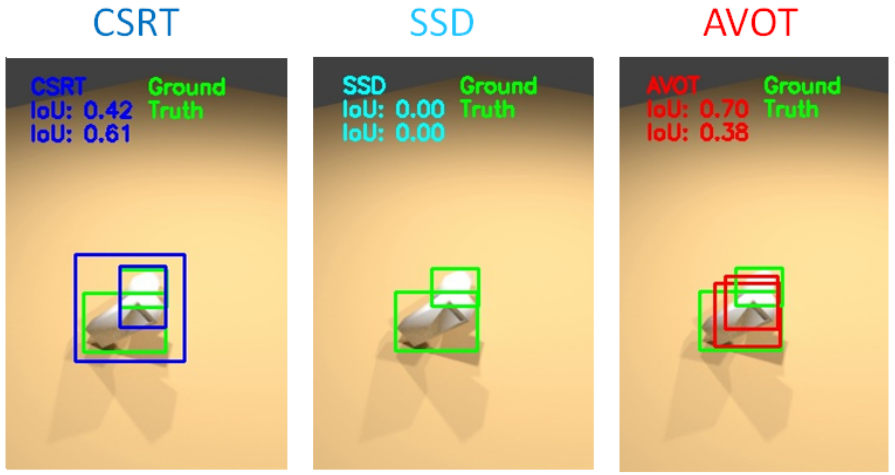


Predicted

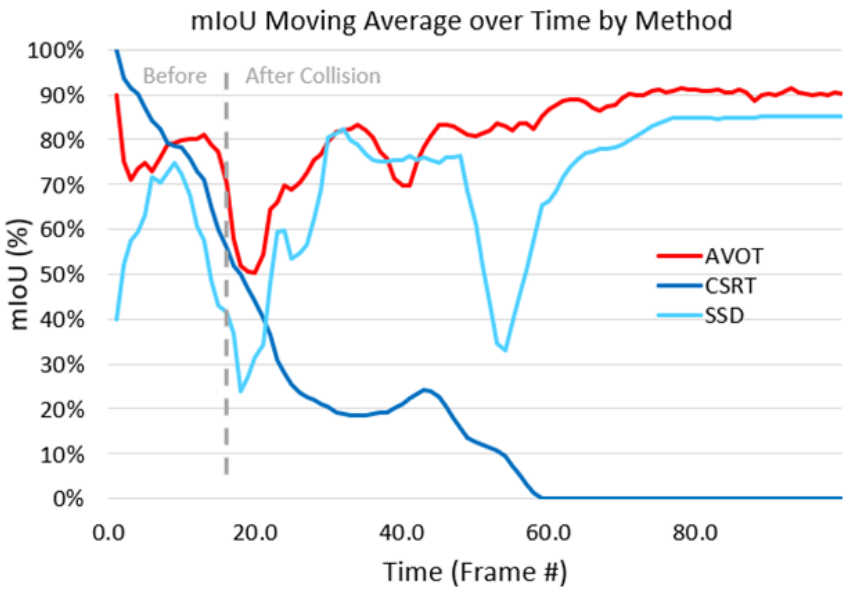
Ground Truth

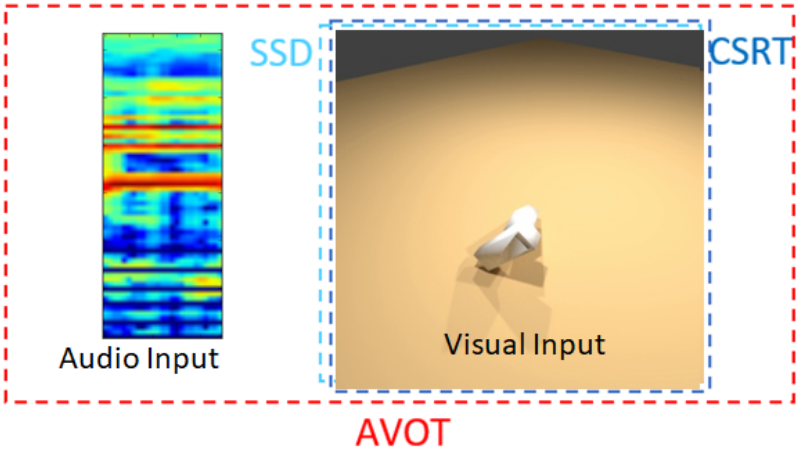
Example frame after collision

- In this ex., AVOT maintains tracking
  - CSRT does not recover
  - SSD temporarily loses tracking during collision



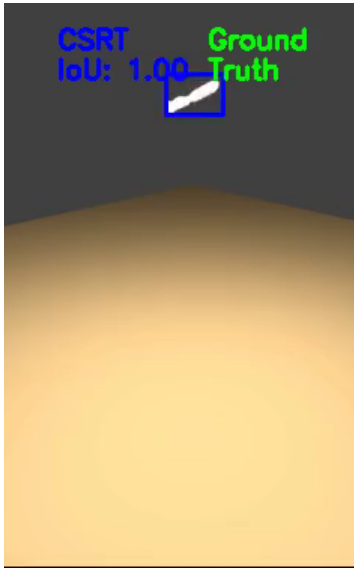
An Example Frame After Collision



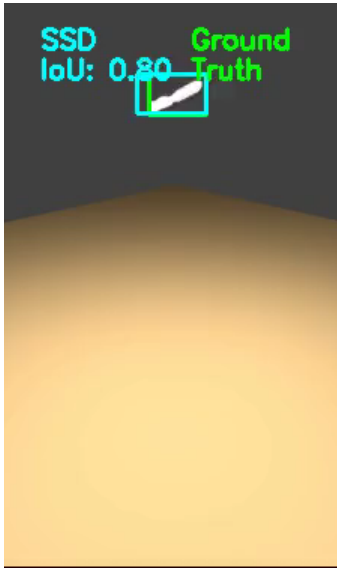


1 Object Free-Falling: Different Geometry & Materials

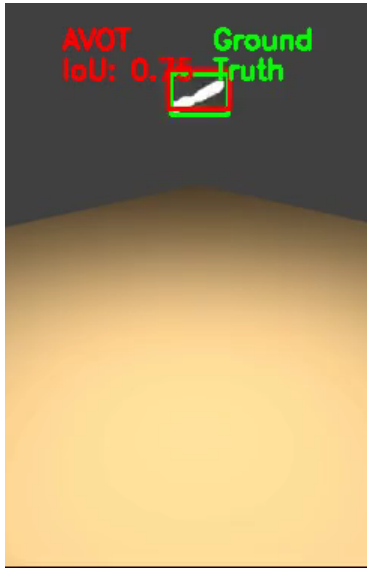
CSRT

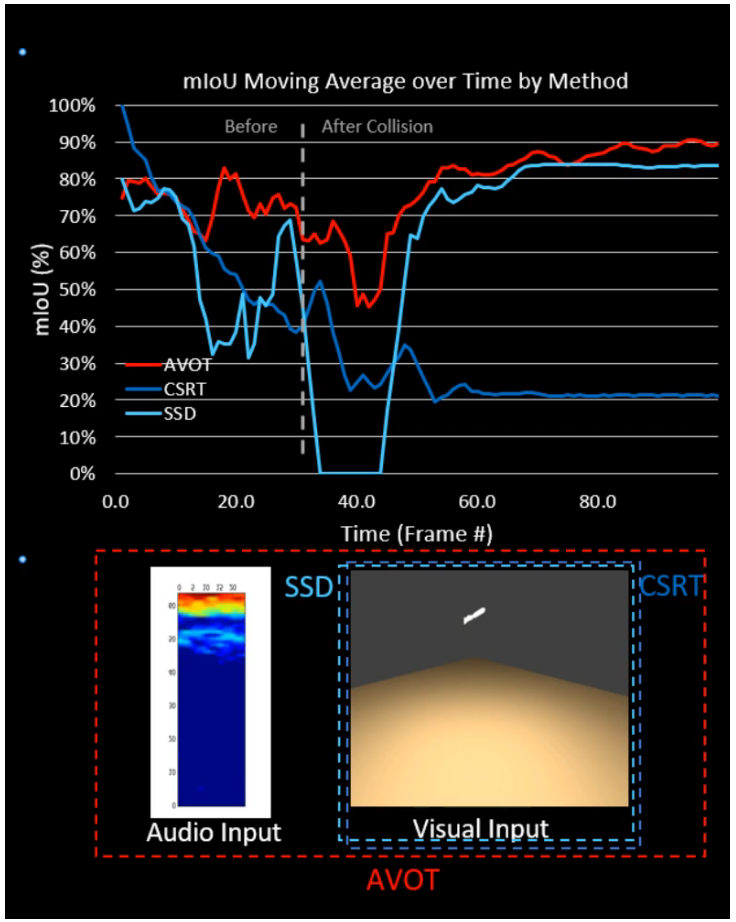


SSD



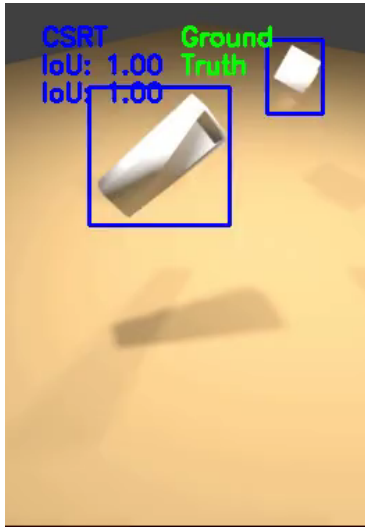
AVOT





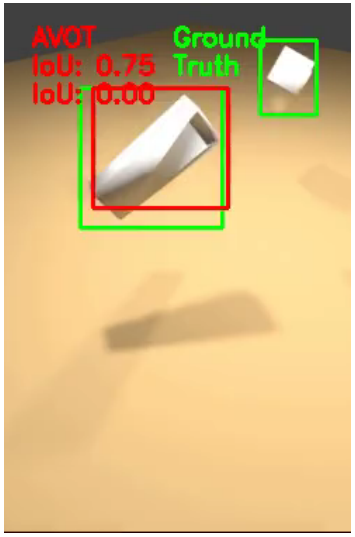
2 Objects Free-Falling: Same Geometry & Different Materials

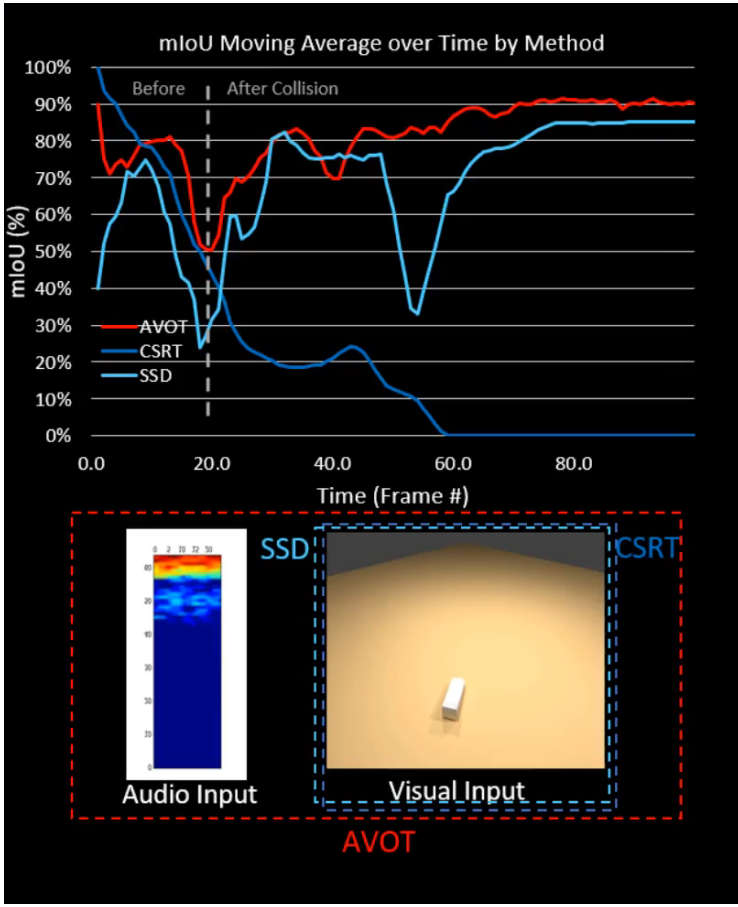
CSRT



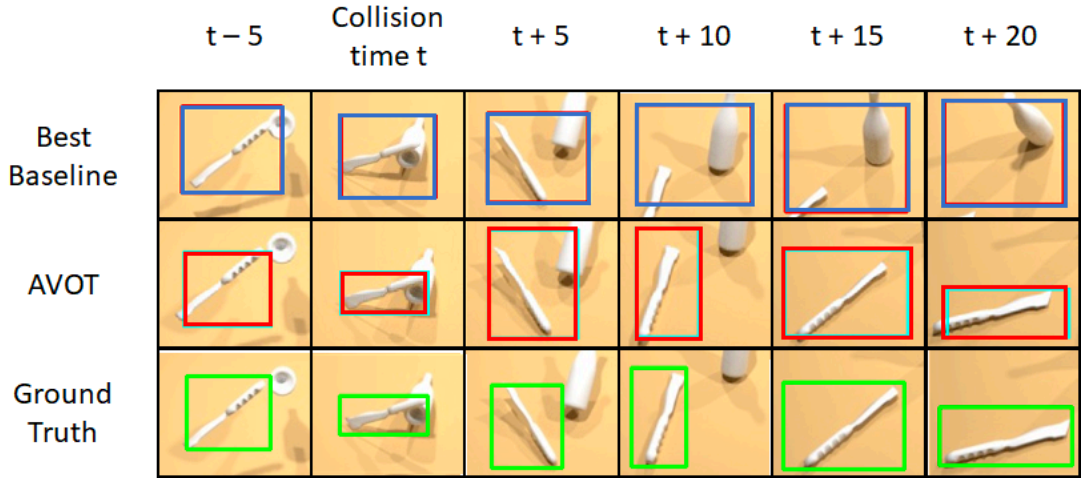
SSD

AVOT



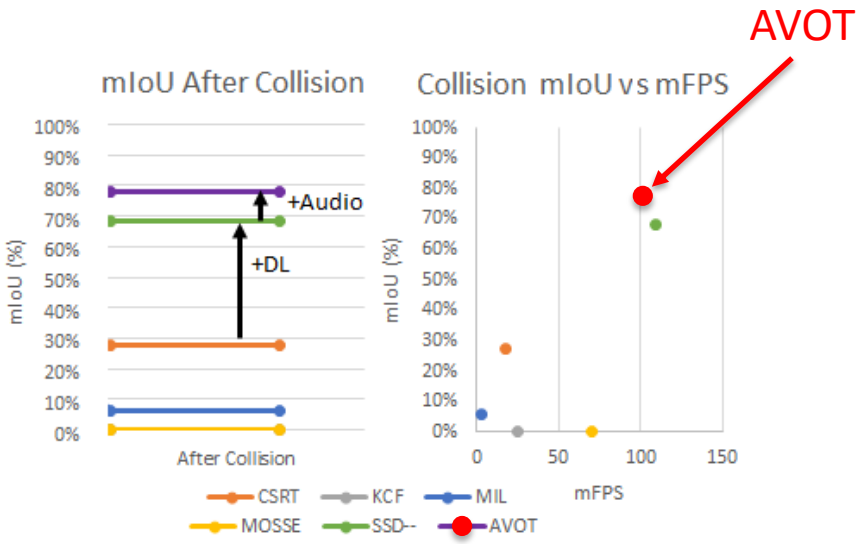


AVOT Neural Network Results



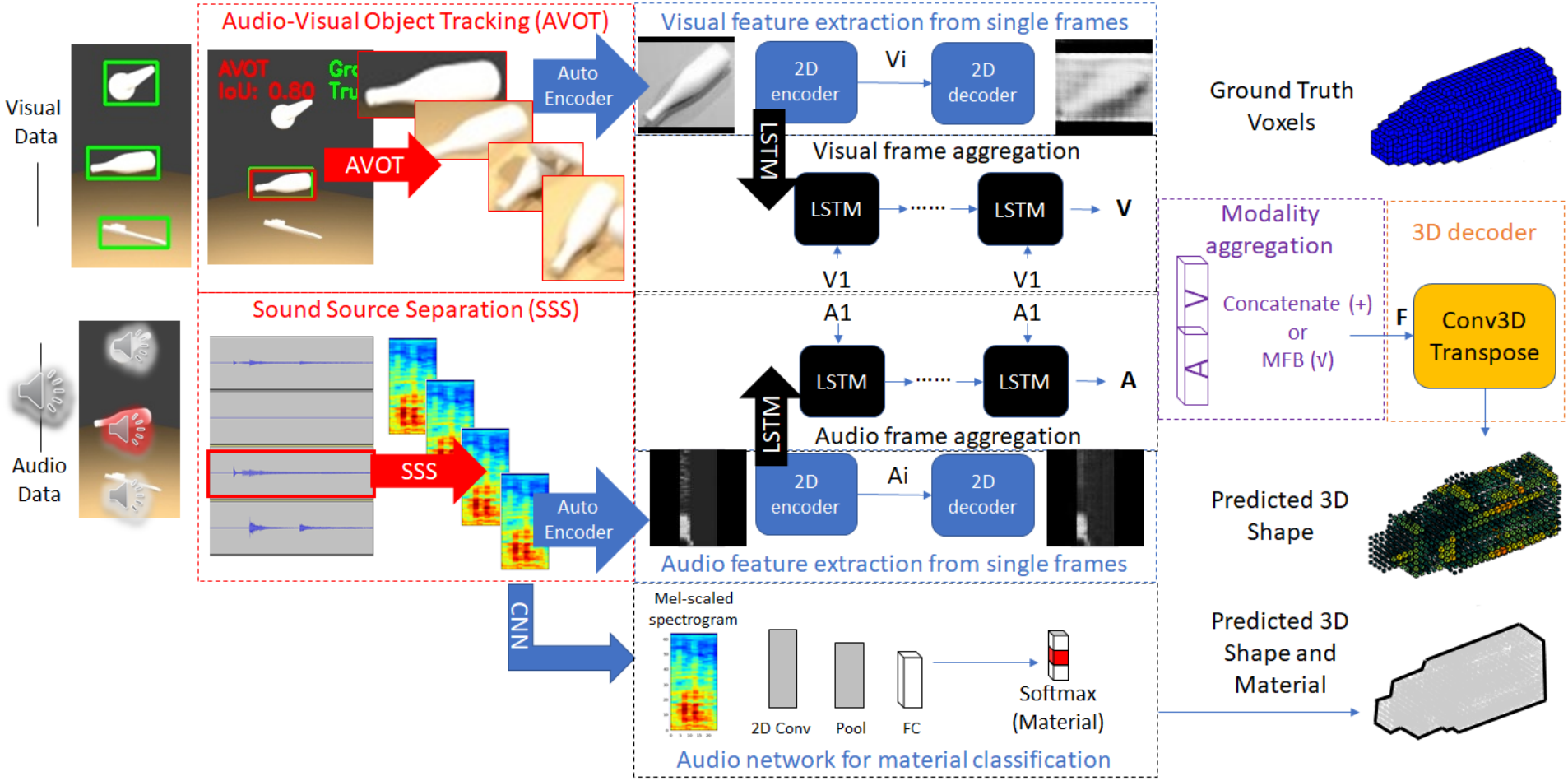
mIoU / mFPS Object Tracking Accuracy by Method

Method	2 Objects	3 Objects
AVOT (Ours)	**58.3%** / 101.6	**66.1%** / 101.0
CSRT [40]	46.9% / 17.1	30.1% / 4.7
KCF [26]	13.5% / 24.9	1.7% / 38.6
MIL [6]	43.0% / 2.5	21.6% / 1.6
MOSSE [8]	7.6% / 70.4	1.0% / **74.5**
SSD– [39]	55.5% / **108.7**	65.9% / **103.8**



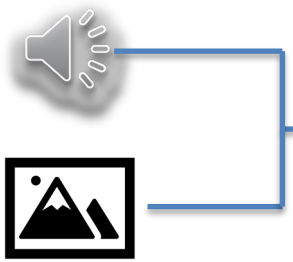
mIoU = mean Intersection over Union
mFPS = mean frames per second

Audio-Visual Object Reconstruction



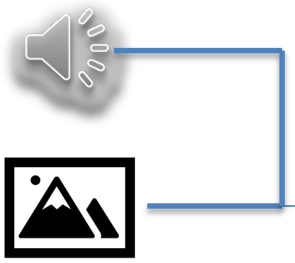
Part V

Audio-Augmented 3D Reconstruction



Main publication: Under review

Related publication: “ISNN: Impact Sound Neural Network for Audio-Visual Object Classification” (ECCV 2018)

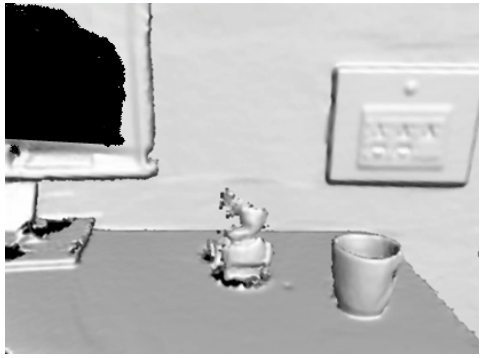


Outline

- I: Introduction
- II: Fluid-Structure Sound Synthesis
- III: Liquid Pouring Sequences
- IV: Audio-Visual Object Tracking
- V: Audio-Augmented 3D Reconstruction
- VI: Conclusions

Related Work: 3D Reconstruction

RGB-D

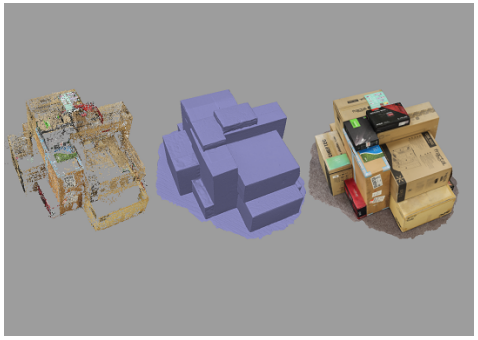


Source: InfiniTAM [Kahler et al. 2015], KinectFusion [Izadi et al. 2011]

RGB

Source: Live Metric 3D Reconstruction on Mobile Phones [Tanskanen et al. 2013]

Photogrammetry

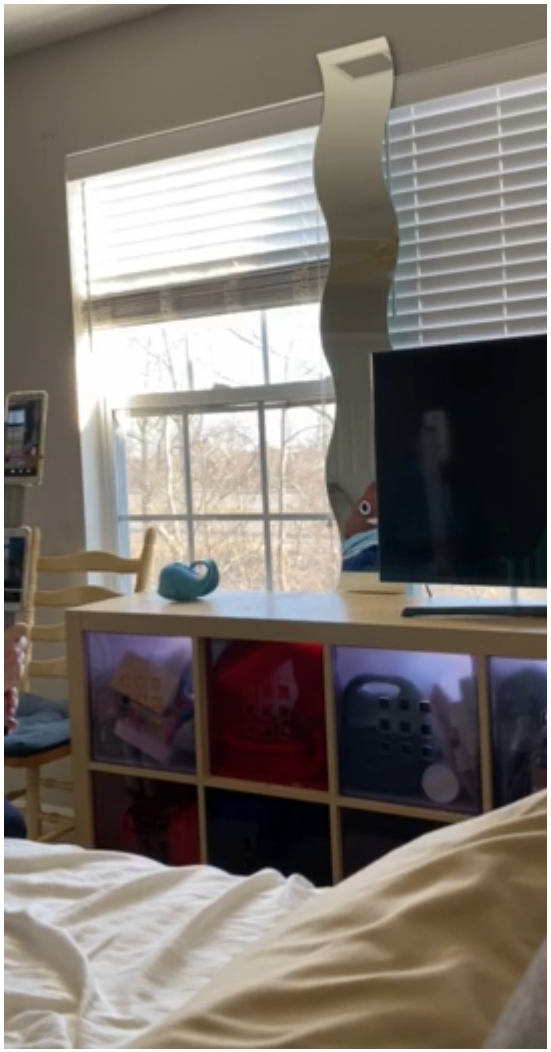


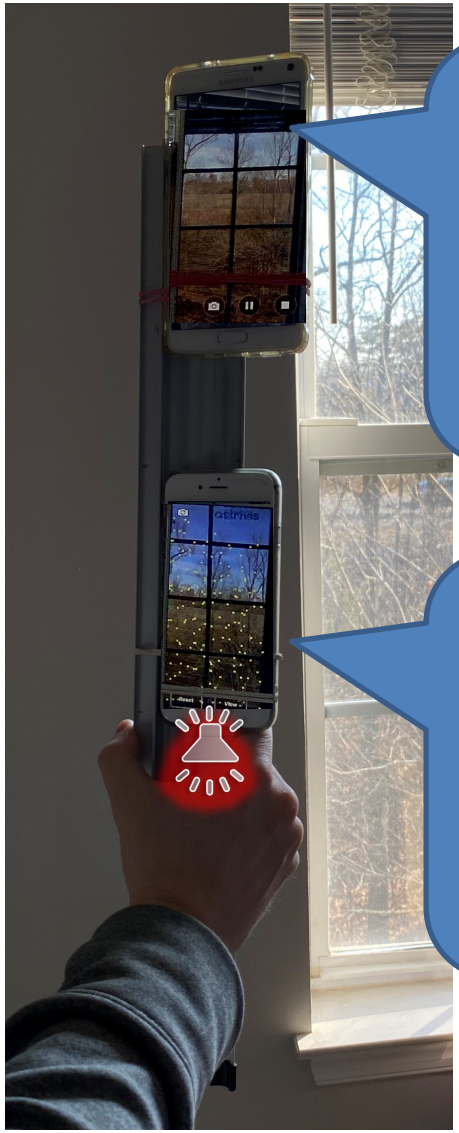
Source: Agisoft Metashape

Audio-Visual

Source: 3D Room Geometry Reconstruction Using Audio-Visual Sensors

System Demo and Mobile Prototype

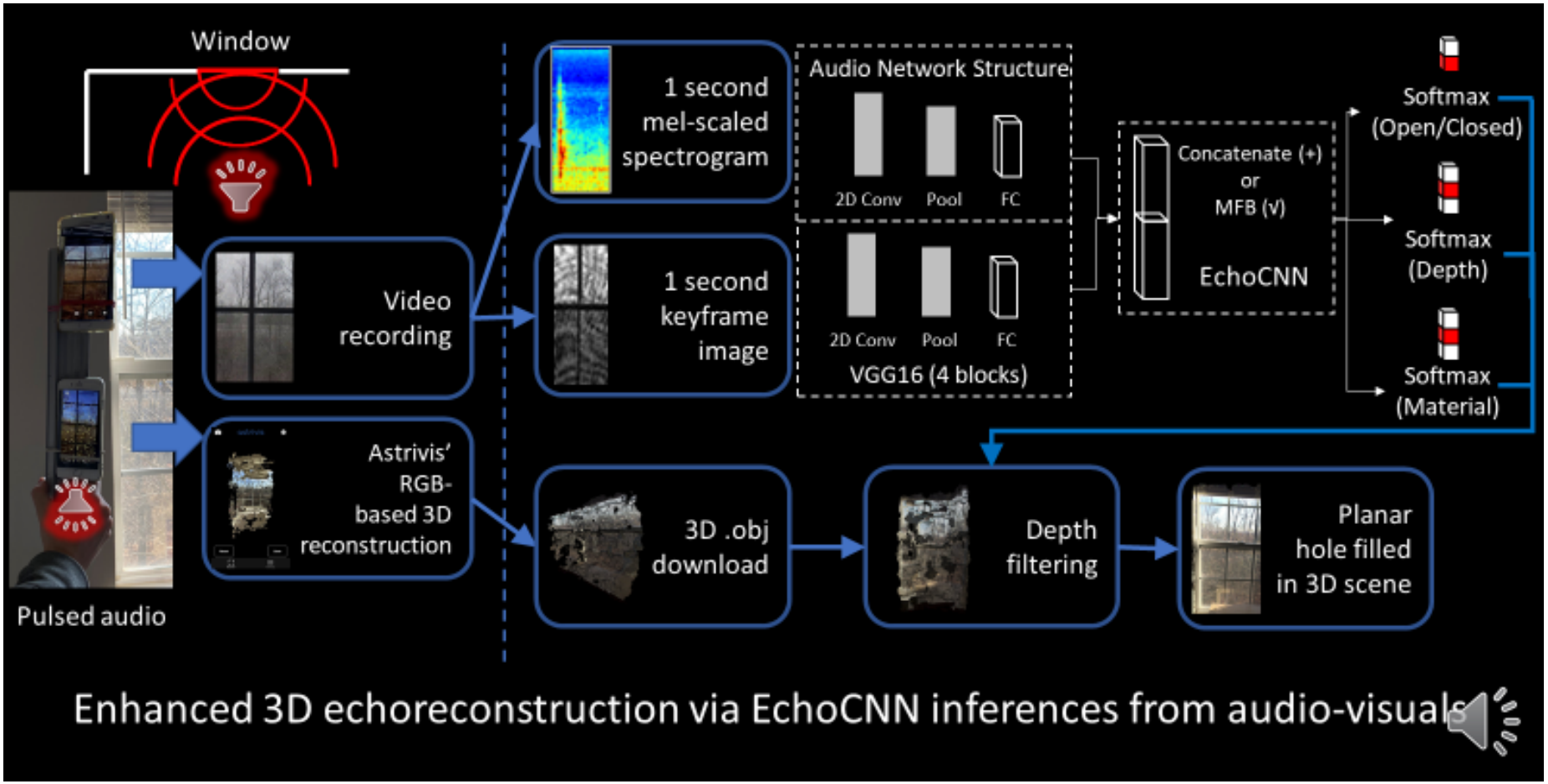




Top smartphone (e.g. Samsung Galaxy Note 4) records video for audio and visual input to enhance bottom reconstruction via EchoCNN classifiers

Bottom smartphone (e.g. iPhone 6) performs a live, RGB-based 3D reconstruction [Astrivis] and emits pulsed audio

System Overview and EchoCNN



EchoCNN Classifiers

Ground Truth

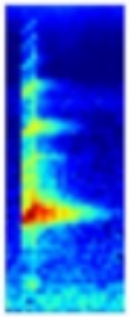
Closed
Glass
2.6 ft

Open/Closed
Material
Depth

Predicted

Closed
Glass
3 ft

Audio



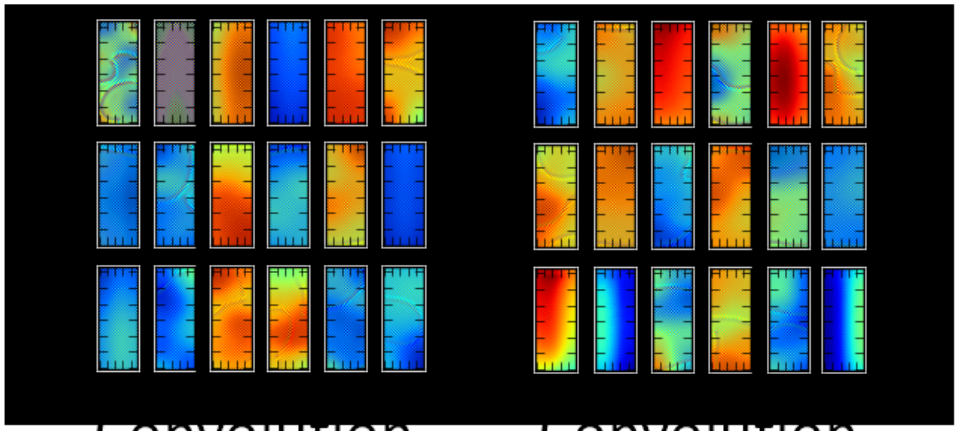
Visual

EchoCNN

Audio-Visual EchoCNN performs surface detection (open/closed), depth estimation, and material classification

Filter Visualizations and Activation Maximization

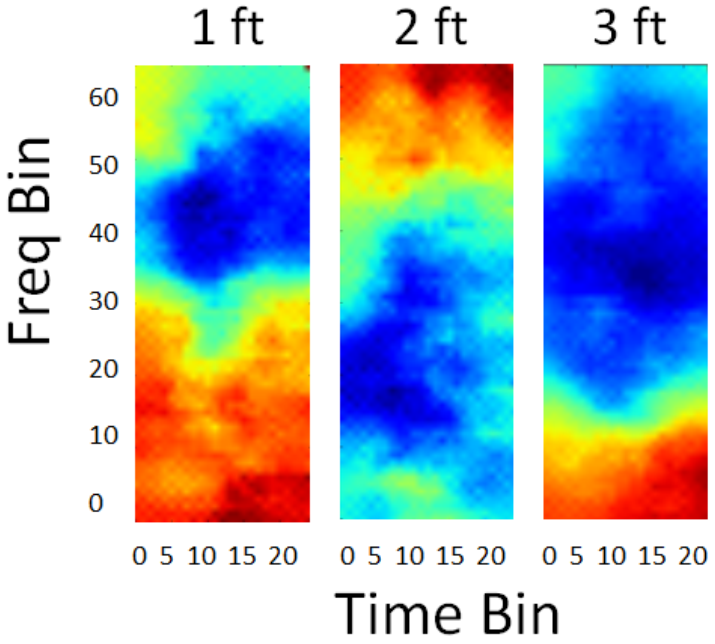
Sample of Filters



Convolution Layer 1

Convolution Layer 2

Depth Activation Maximization

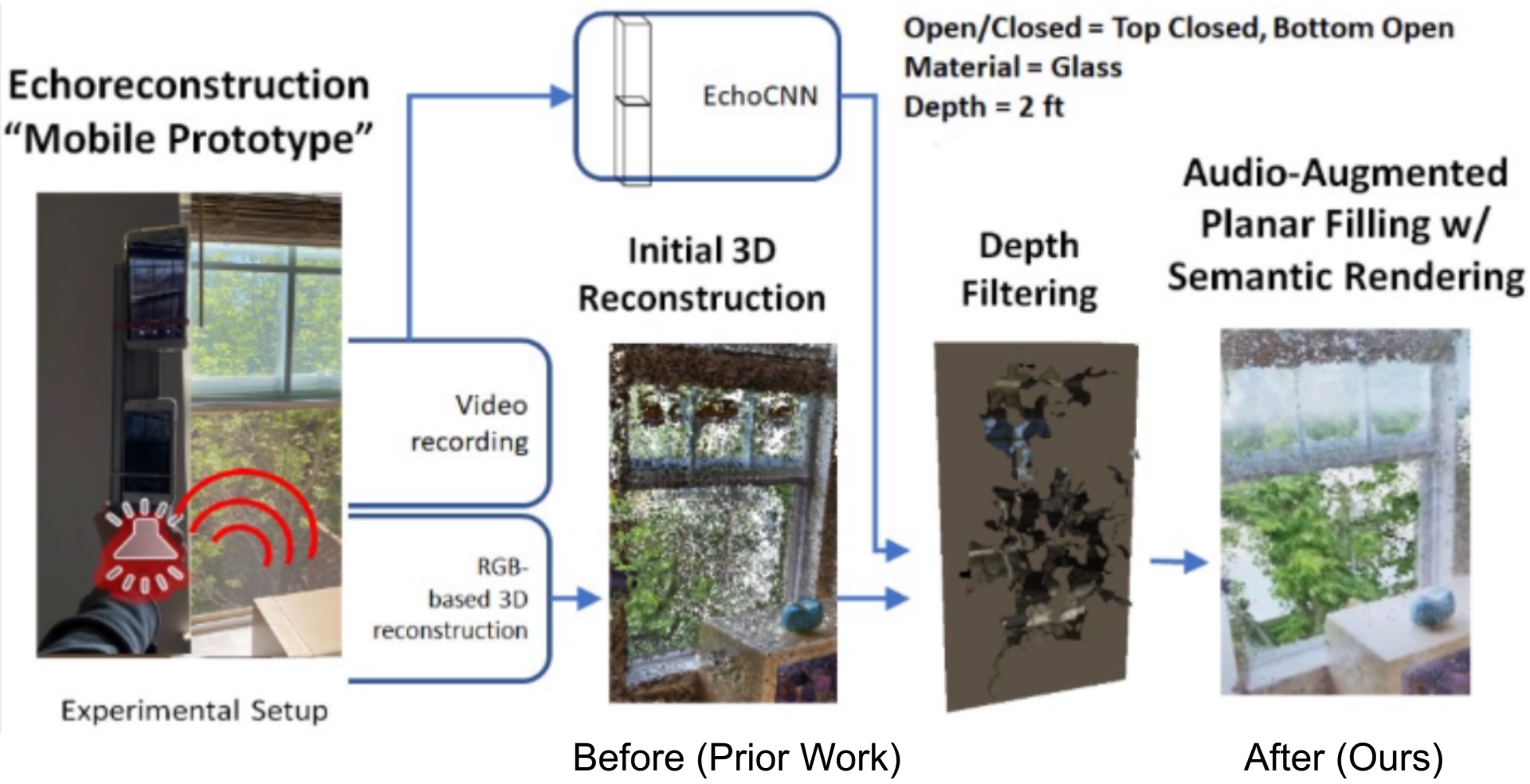


EchoCNN learning longer reverberation times tend to occur at lower frequencies (3 ft) than at higher frequencies (1 and 2 ft) due to typical high frequency damping and absorption

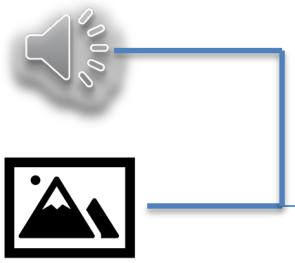
Extensions to Existing Work

- Our approach uses commodity hardware (smartphones)
- Outputs for EchoCNN neural network are
  - (1) surface and sound source detection;
  - (2) depth estimation,
  - (3) and material classification
- We enhance state-of-the-art visual reconstructions by detecting discontinuities using open/closed inferences from our pre-trained EchoCNN

EchoReconstruction Results



VR Demo of Audio-Augmented Scene and Audio Reconstructions

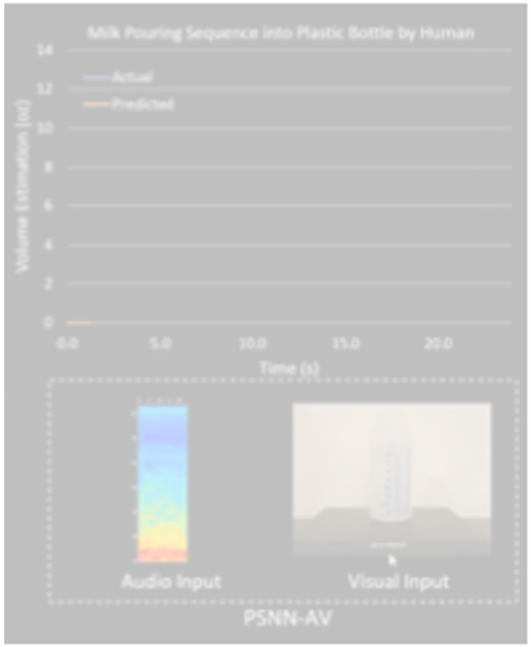


Outline

- I: Introduction
- II: Fluid-Structure Sound Synthesis
- III: Liquid Pouring Sequences
- IV: Audio-Visual Object Tracking
- V: Audio-Augmented 3D Reconstruction
- VI: Conclusions

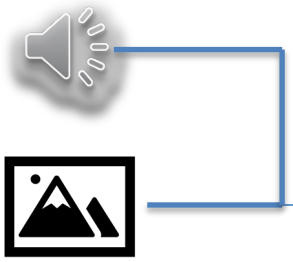
Thesis Statement Revisited

- Coupling multimodal information enhances task performance and processing of audio-visual learning based methods for **fluid-structure sound synthesis**, **liquid pouring sequences**, **object tracking**, and **3D reconstructions** while also allowing for single mode application for special cases.



Conclusions / Contributions
II: Fluid-Structure Sound Synthesis

- Transforming the problem into a single fluid-structure system using the added mass operator
- Enhancing the rigid-body sound synthesis pipeline with pre-processing steps for objects containing a liquid
- Demonstrating on real-time VR applications

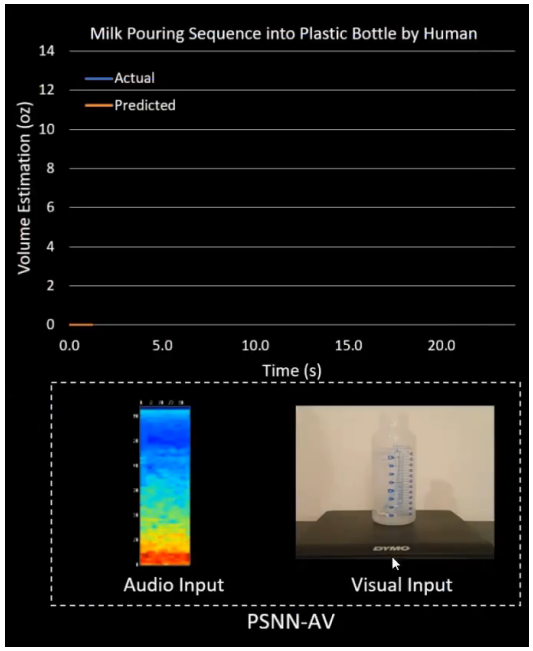


II: Fluid-Structure Sound Synthesis

- Limitations
  - Assumes that the liquids are inviscid, remain steady, and are not mixed
  - Assumes a non-moving domain; that is, the structural vibration must move the liquid along with the structure
  - Limited by the granularity of the solid mesh discretization since modifications to the mass matrix occur at the mesh nodes
- Future Work
  - Investigation of acoustic transfer
  - User evaluation on auditory perception
  - Acquire model of the object using 3D reconstruction

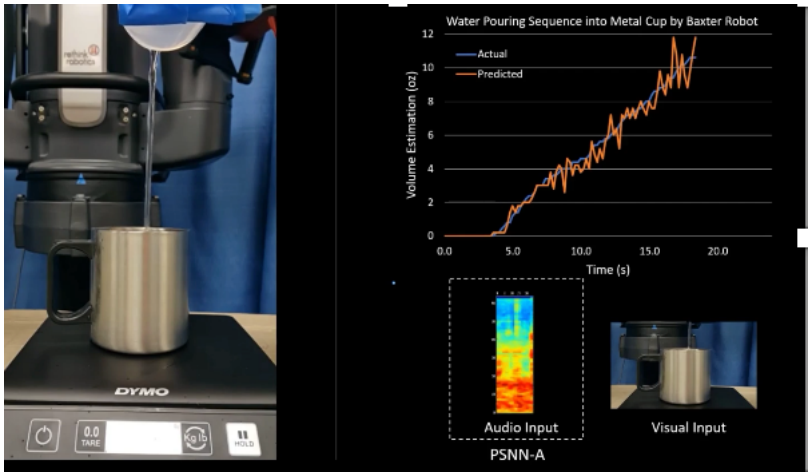
Thesis Statement Revisited

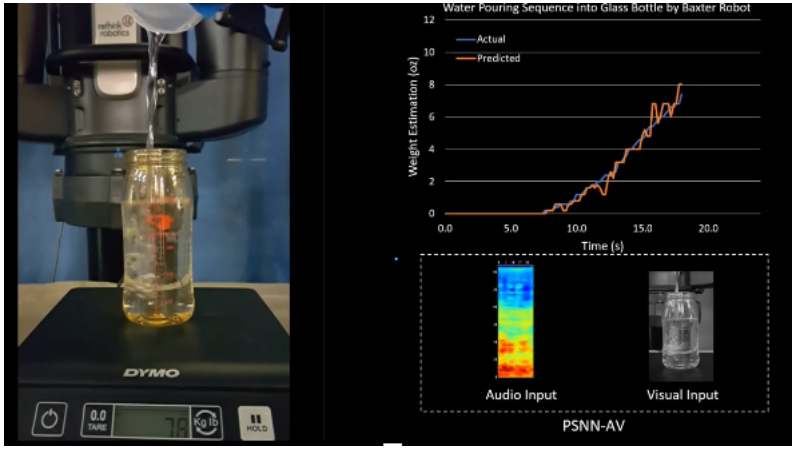
- Coupling multimodal information enhances task performance and processing of audio-visual learning based methods for **fluid-structure sound synthesis**, **liquid pouring sequences**, **object tracking**, and **3D reconstructions** while also allowing for single mode application for special cases.



Conclusions / Contributions
III: Liquid Pouring Sequences

- Audio-visual pouring dataset
- Audio PSNN-A for multiclass weight estimation and binary overflow detection
- Audio-augmented PSNN-AV neural network enhancing the audio based method with fused visual inputs
- Pouring content and container classification



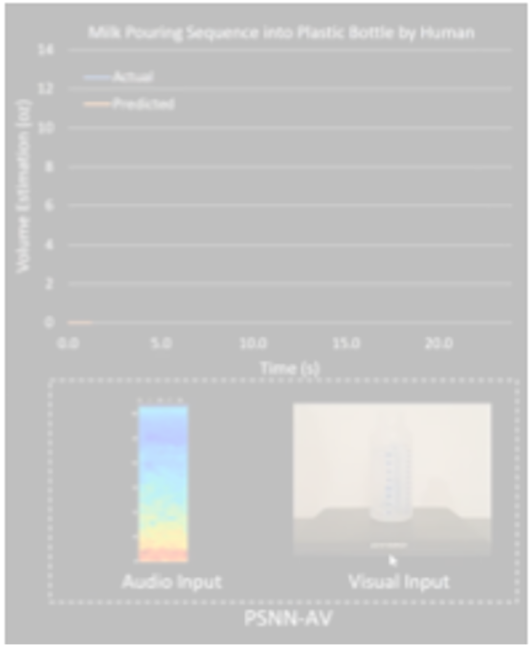


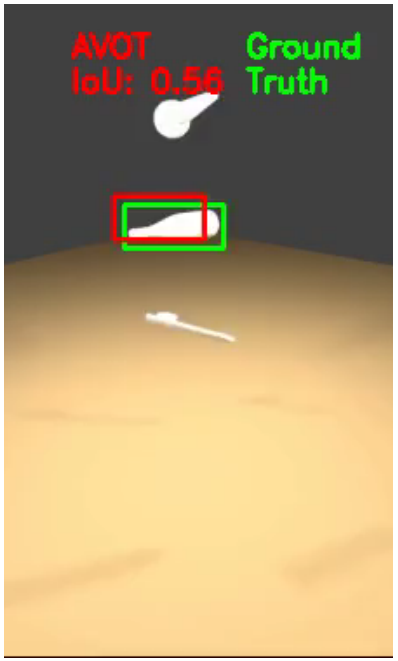
III: Liquid Pouring Sequences

- Limitations
  - Liquids are not mixed
- Future Work
  - Evaluate effectiveness of augmenting with synthetic data
  - Explore if approach can be applied to other containers, liquids, and granular materials

Thesis Statement Revisited

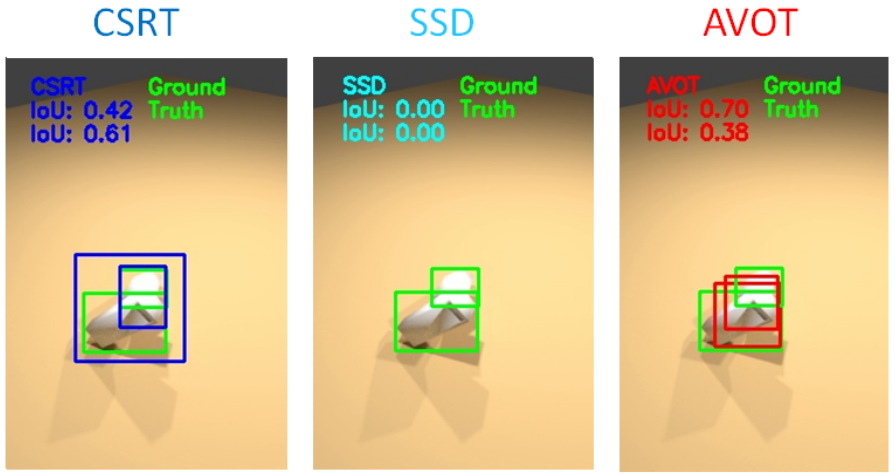
- Coupling multimodal information enhances task performance and processing of audio-visual learning based methods for **fluid-structure sound synthesis**, **liquid pouring sequences**, **object tracking**, and **3D reconstructions** while also allowing for single mode application for special cases.



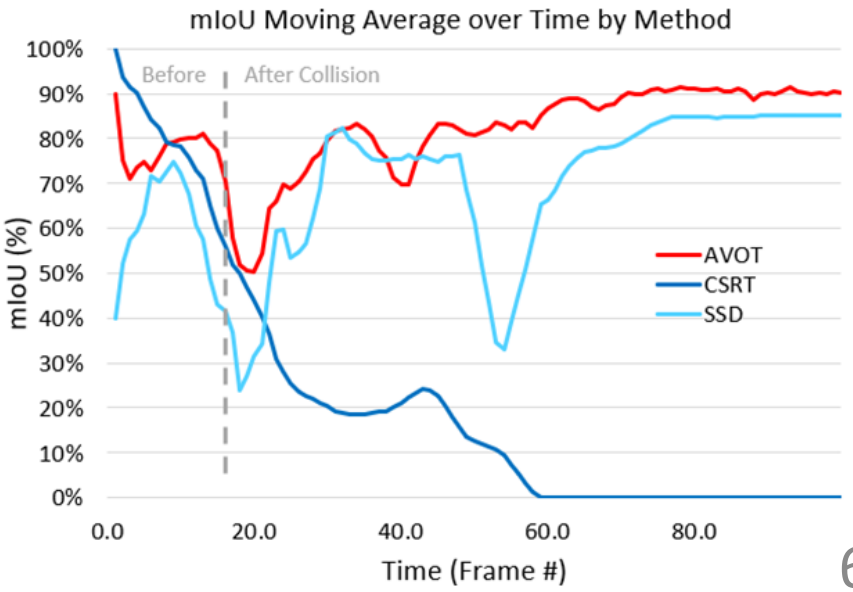


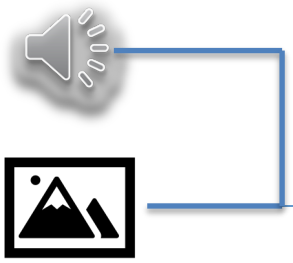
Conclusions / Contributions
IV: Audio-Visual Object Tracking

- An end-to-end, jointly trained audio-visual object tracker (AVOT) to enhance visual object tracking
- Ground truth bounding box annotations for Sound-20K audio-visual dataset with 1, 2, and 3 object scenes
- Experimental results against baselines for mIoU and mFPS



An Example Frame After Collision



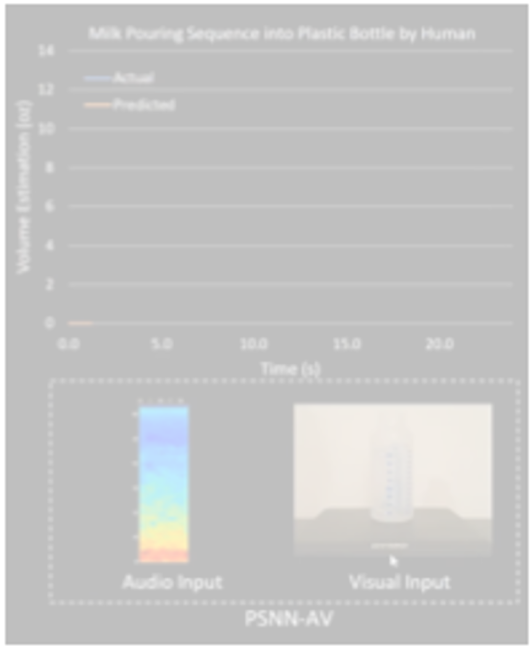


IV: Audio-Visual Object Tracking

- Limitations
  - Few audio-visual object tracking datasets
  - Dataset contains 18 objects varying geometry & materials
- Future Work
  - Evaluate alternative audio-visual fusion methods
  - Augment audio data and test audio-only object tracker
  - Research on generative models and classifying material per voxel using both audio and visual data

Thesis Statement Revisited

- Coupling multimodal information enhances task performance and processing of audio-visual learning based methods for **fluid-structure sound synthesis**, **liquid pouring sequences**, **object tracking**, and **3D reconstructions** while also allowing for single mode application for special cases.

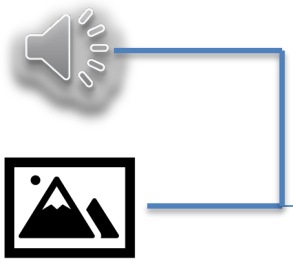


Conclusions / Contributions
V: Audio-Augmented Reconstruction

- EchoCNN, a fused audio-visual CNN architecture for classifying open/closed surfaces, depth, and material
- EchoReconstruction, a staged audio-visual 3D reconstruction pipeline using mobile phones to enhance scene geometry containing windows, mirrors, and open surfaces with depth filtering and inpainting

Scene Reconstruction

Audio Reconstruction



V: Audio-Augmented Reconstruction

- Limitations
  - EchoCNN's depth estimation inference in increments of 6 and 12 inches
  - Staged approach for audio-augmentation instead of an integrated pipeline
- Future Work
  - Perform audio emission, receiving, and 3D reconstruction simultaneously and in real-time
  - Investigate the impact of live audio for training and testing

Acknowledgements

I would like to thank my advisors and committee:

- Henry Fuchs
- Ming Lin
- Gary Bishop
- Dinesh Manocha
- Shahriar Nirjon

Co-authors and colleagues:

- Auston Sterling
- Nick Rewkowski
- Graphics & Virtual Reality research group
- GAMMA research group
- MITRE

Funding agencies:

- U.S. National Science Foundation
- University of Maryland Elizabeth Stevinson Iribe Chair Professorship

Publications

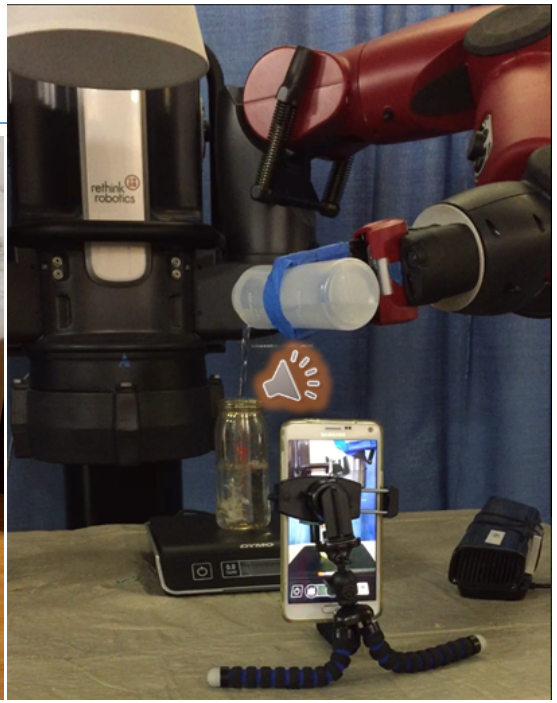
Main publications

- **J. Wilson**, A. Sterling, N. Rewkowski, and M. Lin, "Glass Half Full: Sound Synthesis for Fluid-Structure Coupling using Added Mass Operator" (CGI 2017)
- **J. Wilson**, A. Sterling, and M. Lin, "Analyzing Liquid Pouring Sequences via Audio-Visual Neural Networks" (IROS 2019)
- **J. Wilson** and M. Lin, "AVOT: Audio-Visual Object Tracking of Multiple Objects for Robotics" (ICRA 2020)
- **J. Wilson**, N. Rewkowski, M. Lin, and H. Fuchs, "Echo-Reconstruction: Audio-Augmented Scene Reconstruction with Mobile Devices" (Under Review)

Related publications

- A. Sterling, **J. Wilson**, S. Lowe, and M. Lin, "ISNN: Impact Sound Neural Network for Audio-Visual Object Classification" (ECCV 2018)

Thank You



AVOT