CHAPTER 1: ISNN: Impact Sound Neural Network for Audio-Visual Object Classification[1]

1.1 Introduction

The problem of object detection, classification, and segmentation are central to understanding complex scenes. Detection of objects is typically approached using visual cues (Girshick et al., 2014; Ren et al., 2015). Classification techniques have steadily improved, advancing our ability to accurately label an object by class given its depth image (Wu et al., 2015), voxelization (Maturana and Scherer, 2015), and/or RGB-D data (Socher et al., 2012). Segmentation of objects from scenes provides contextual understanding of scenes (Golodetz* et al., 2015; Valentin et al., 2015). While these state-of-the-art techniques often result in high accuracy for common scenes and environments, there is still room for improvement when accounting for different object materials, textures, lighting, and other variable conditions.

The challenges introduced by transparent and highly reflective objects remain open research areas in 3D object classification. Common vision-based approaches cannot gain information about the internal structure of objects, however audio-augmented techniques may contribute that missing information. Sound as a modality of input has the potential to close the audio-visual feedback loop and enhance object classification. It has been demonstrated that sound can augment visual information-gathering techniques, providing additional clues for classification of material and general shape features (Zhang et al., 2017; Arnab et al., 2015). However, previous work has not focused on identifying complete object geometries. Identifying object geometry from a combined audio-visual approach expands the capabilities of scene understanding.

In this chapter, we consider identification of rigid objects such as tableware, tools, and furniture that are common in indoor scenes. Each object is identified by its geometry and its material. A discriminative factor for object classification is the sound that these objects produce when struck, referred to as an *impact sound*. This sound depends on a combination of the object's material composition and geometric model. Impact

[1]This chapter previously appeared as a paper in the European Conference on Computer Vision (ECCV 2018). The original citation is as follows: Sterling, A., Wilson, J., Lowe, S., and Lin, M. C. (2018). ISNN: Impact sound neural network for audio-visual object classification. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 578–595, Cham. Springer International Publishing

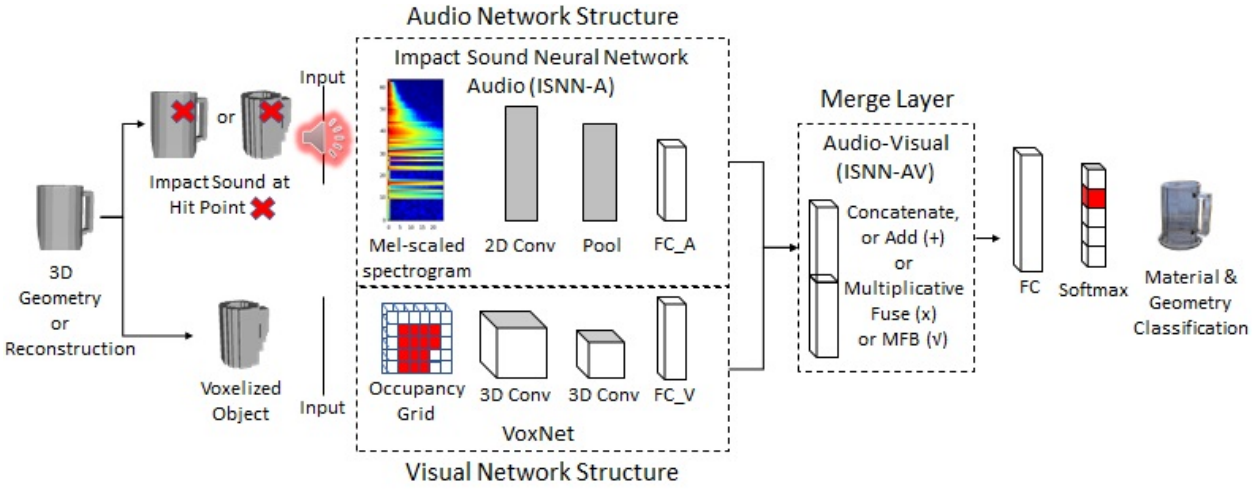


Figure 1.1: Our Impact Sound Neural Network - Audio (ISNN-A) uses as input a spectrogram of sound created by a real or synthetic object being struck. Our audio-visual network (ISNN-AV) combines ISNN-A with VoxNet to produce state-of-the-art object classification accuracy.

sounds are distinguished as object discriminators from video in that they reflect the internal structure of the object, providing clues about parts of an opaque or transparent object that cannot be seen visually. Impact sounds, therefore, complement video as an input to object recognition problems by addressing the some inherent limitations of incomplete or partial visual data.

Main Results: We introduce an audio-only Impact Sound Neural Network (ISNN-A) and a multimodal audio-visual neural network (ISNN-AV). These networks:

- Are the first networks to show high classification accuracy of both an object's geometry and material based on its impact sound;
- Use impact sound spectrograms as input to reduce overfitting and improve accuracy and generalizability;
- Merge multimodal inputs through bilinear models, which have not been previously applied to audio-visual networks yet result in higher accuracy as demonstrated in Table 1.4;
- Provide state-of-the-art results on geometry classification; and
- Enable real time, interactive scene reconstruction in which users can strike objects to automatically insert the appropriate object into the scene.

Figure 1.2: We use various datasets for training and testing: (1) our RSAudio dataset with real and synthesized impact sounds from objects of varying shapes and sizes and (2) voxelized ModelNet objects. (3) Audio inputs are formatted as spectrograms.

1.2 Audio and Visual Datasets

To perform multimodal classification of object geometries, we need datasets containing appropriate multimodal information. Visual object reconstruction can provide a rough approximation of object geometry, serving as one form of input. *Impact audio produced from real or simulated object vibrations provide information about internal and occluded object structure, making for an effective second input.* Figure 1.2 provides examples of object geometries, while the corresponding spectrograms model the sounds that provide another input modality.

Appropriate audio can be found in some existing datasets, but the corresponding geometries are difficult to model. AudioSet contains impact sounds in its "Generic impact sounds" and "{Bell, Wood, Glass}" categories (Gemmeke et al., 2017), while ESC-50 has specific categories including "Door knock" and "Church bells" (Piczak, 2015b). The *Greatest Hits* sound dataset comes closest to our needs, containing impact sounds labeled according to the type of object (Owens et al., 2016). However, many of the categories do not contain rigid objects (*e.g.*, cloth, water, grass) or contain complex structures that cannot be represented with one geometric model of one material (*e.g.*, a stump with roots embedded in the ground).

We want to use an impact sound as one input to identify a specific geometric model that could have created that sound. A classifier for this purpose could be trained on a large number of recorded sounds produced from struck objects. However, it is difficult and time-consuming to obtain a representative sample of real-world objects of all shapes and sizes. It is much easier to create a large dataset of synthetic sounds using geometric shapes and materials which can be applied to the objects. We now describe our methodology for generating the data used for training, as visualized in Figure 1.3.

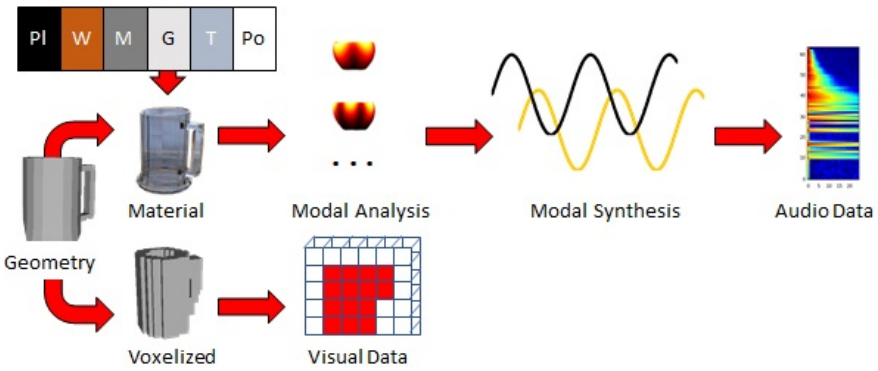


Figure 1.3: We build multimodal datasets through separate processing flows. Modal sound synthesis produces spectrograms used for audio input. Voxelization as another modality provides a first estimate of shape. Incorporating audio features improves classification accuracy through understanding of how objects vibrate.

1.2.1 Audio Data

We create a large amount of our training data by simulating the vibrations of rigid-body objects and the sounds that they produce. We use the established process of modal sound synthesis to create synthetic sound datasets from 3D models. The process of modal sound synthesis is described in detail in **??**.

1.2.2 Audio Augmentations

Modal sound synthesis produces the set of frequencies, damping rates, and initial amplitudes of an object's surface vibrations. However, since we are attempting to imitate real-world sounds, there are some additional auditory effects to take into account: acoustic radiance, room acoustics, background noise, and time variance.

Acoustic Radiance: Sound waves produced by the object must propagate through the air to reach a listener or microphone position. Even in an empty space, the resulting sound will change with different listener positions depending on the vibrational mode shapes; this is the acoustic radiance of the object (James et al., 2006). This effect has a high computational cost for each geometric model, and since we use datasets with relatively large numbers of models, we do not include it in our simulations.

Room Acoustics: In an enclosed space, sound waves bounce off walls to produce early echo-like reflections and noisy late reverberations; this is the effect of room acoustics. We created a set of room impulse responses in rooms of different sizes and materials using a real-time sound propagation simulator, GSound (Schissler and Manocha, 2011). Each modal sound is convolved with a randomly selected room impulse response.

Background Noise: In most real-world situations, background noise will also be present in any recording. We simulate background noise through addition of a random segment of environmental audio from the DEMAND database (Thiemann et al., 2013). These noise samples come from diverse indoor and outdoor environments and contain around 1.5 hours of recordings.

Time Variance: Finally, we slightly randomize the start time of each modal sound. This reflects the imperfect timing of any real-world recording process. Together, these augmentations make the synthesized sounds more accurately simulate recordings that would be taken in the real world.

1.2.3 Visual Data

Our visual data consists of datasets of geometric models of rigid objects, ranging from small to large and of varying complexity. Given these geometric models, we can simulate synthesized sounds for a set of possible materials. During evaluation, object classification results were tested using multiple scenarios of voxelization, scale, and material assignment (Section 1.4.2).

1.3 Impact Sound Neural Network (Audio & Audio-Visual)

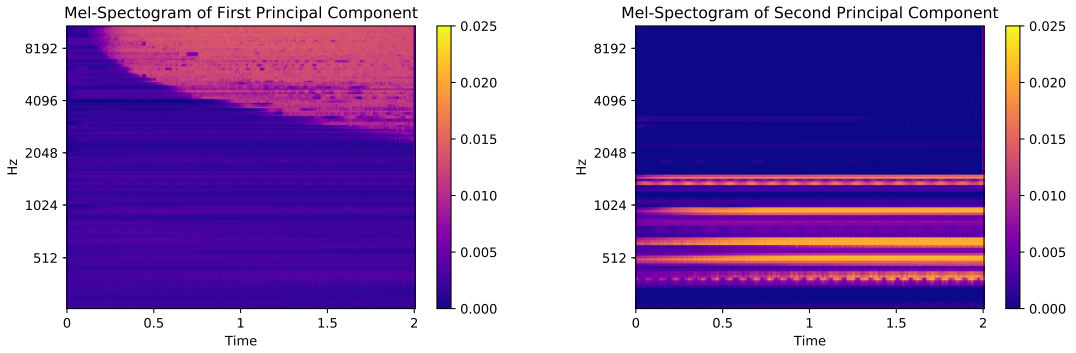
Given the impact sounds and representation described in Section 1.2, we now examine their ability to identify materials and geometric models. We begin with an analysis of the distributions of the features themselves as proper feature selection is a key component in classifier construction.

1.3.1 Input Features and Analysis

1.3.1.1 Audio Features

In environmental sound classification tasks, classification accuracy can be affected by the input sound's form of representation (Cowling and Sitte, 2003; Huzaifah, 2017). A one-dimensional time series of audio samples over time can be used as features (Aytar et al., 2016), but they do not capture the spectral properties of sound. A frequency dimension can be introduced to create a time-frequency representation and better represent the differentiating features of audio signals.

In this work, we use a mel-scaled spectrogram as input. Spectrograms have demonstrated high performance in CNNs for other tasks (Huzaifah, 2017). A given sound, originally represented as a waveform of audio samples over time, is first trimmed to one second in length since impact sounds are generally transient.



(a) (b)

Figure 1.4: The first two principal components of 420 synthesized sounds demonstrate that the key differentiating factors between sounds and models are the presence of high-frequency damping (first component) and the presence of specific frequency bins (second component).

The sound is resampled to 44.1 kHz, the Nyquist rate for the full range of audible frequencies up to 22.05 kHz. We compute the short-time Fourier transform of the sound, using a Hann window function with 2048 samples and an overlap of 25 %. The result is squared to produce a canonical "spectrogram", then the frequencies are mapped into mel-scaled bins to provide appropriate weights matching the logarithmic perception of frequency. Each spectrogram is individually normalized to reduce the effects of loudness and microphone distance. To create the final input features for the classifier, we downsample the mel-spectrogram to a size of 64 frequency bins by 25 time bins.

We performed principal component analysis on a small sample of synthesized impact sounds to demonstrate the advantage of mel-spectrograms as input features for audio of this type. We used 70 models and 6 materials with a single hit per combination to synthesize a total of 420 impact sounds for this analysis. Figure 1.4 displays the first two principal components as mel-spectrograms, describing important distinguishing factors in our dataset. The first component identifies damping in higher frequencies, while the second component identifies specific frequency bins.

Figure 1.5 contains a scatter plot of material classes on the axes of the first two principal components. The first principal component explains much of the variation between material classes, as there is clear horizontal delineation—albeit with overlap. This is consistent with the expectation of damping as a material-dependent property. The presence of specific frequency bins that comprise the second component likely delineates model more than material.

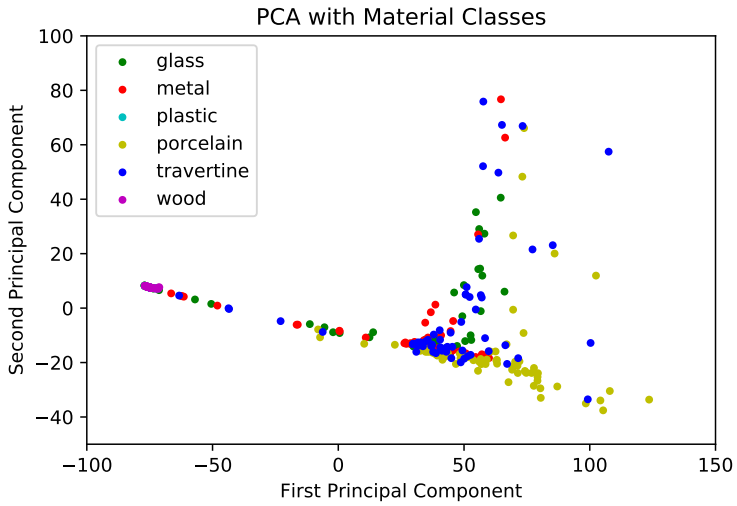


Figure 1.5: A scatter plot of material classes on the first two principle component axes. While the horizontal delineation of materials is useful in characterizing those sounds, a full understanding of the relationships between materials and models necessitates a deeper classification scheme.

1.3.1.2 Visual Features

As in VoxNet (Maturana and Scherer, 2015), visual data serves as an input into classification models based on a 30x30x30 voxelized representation of the object geometry. We voxelize models from our real and synthetic dataset and ShapeNets ModelNet10 and ModelNet40. All objects were voxelized using the same voxel and grid size. We generated audio and visual data for our dataset and up to 200 objects (train and test) per ModelNet class.

1.3.2 Model Architecture

Using our audio and visual features, our approach to performing object geometry classification uses convolutional neural networks (CNNs) due to their high accuracy in a wide variety of tasks, with the specific motivation that convolutional kernels should be able to capture the recurring patterns underlying the structure of our sounds.

1.3.2.1 Audio-Only Network (ISNN-A)

We first developed a network structure to perform object classification using audio only. Our audio Impact Sound Neural Network (ISNN-A) is based on optimization performed over a search space combining general network structure, such as the number of convolutional layers, and hyperparameter values. This

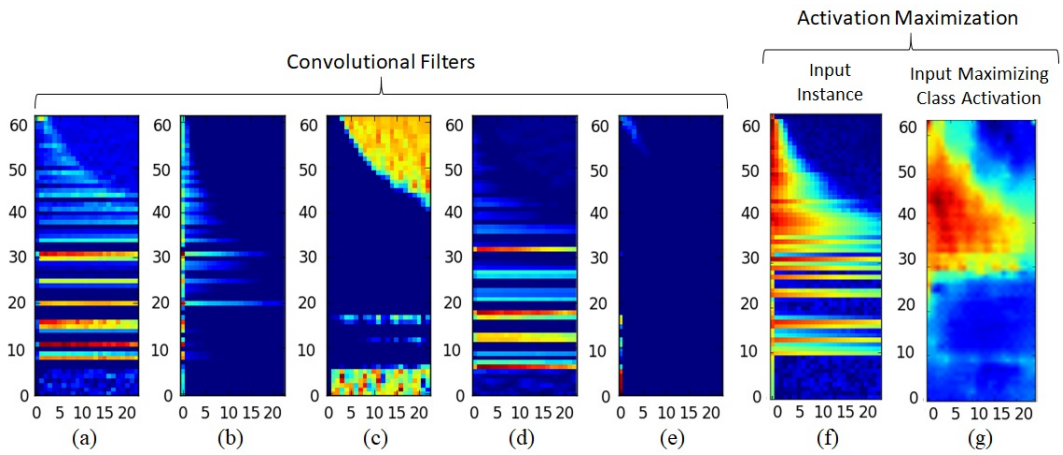


Figure 1.6: Sample activations (a-e) of ISNN convolution layer. Filters identify characteristic patterns in frequencies (a) (d), damping rates (b) (c), and high-frequency noise (e). The distinguishing characteristics in these activations match the expected factors discovered in the PCA analysis in Figure 1.4. An audio input spectrogram (f) and activation maximization (g) learned by the ISNN network for the toilet ModelNet10 class show correctly-learned patterns.

optimization was performed using the TPE algorithm (Bergstra et al., 2011). We found a single convolutional layer followed by two dense layers performs optimally on our classification tasks. This network structure utilizes a convolution kernel with increased frequency resolution to more effectively recognize spectral patterns across a range of frequencies (Piczak, 2015a). Our generally low number of filters and narrower layer sizes aim to reduce overfitting by encouraging the learning of generalizable geometric properties.

Figure 1.6 shows sample activations of a convolutional layer of the ISNN-A network. Based on the PCA and modal analysis we performed, we expect that the differences between geometries primarily manifest as different sets of modal frequencies, as well as different sets of initial mode amplitudes and damping rates. These activations corroborate our expectations. In Figure 1.4(a), we see that damping is an important discriminating feature, which has been learned by filters (b) and (c) in Figure 1.6. Similarly, the frequency patterns that we expected because of Figure 1.4(b) can be seen in filters (a) and (d). This demonstrates that our model is learning statistically optimal kernels with high discriminatory power.

1.3.2.2 Multimodal Audio-Visual Network (ISNN-AV)

Our audio-visual network, as shown in Figure 1.1, consists of our audio-only network combined with a visual network based on VoxNet (Maturana and Scherer, 2015) using either a concatenation, addition, multiplicative fusion, or bilinear pooling operation. Concatenation and addition serve as our baseline

operations, in which the outputs of the first dense layers are concatenated or added before performing final classification. These operations are not ideal because they fail to emulate the interactions that occur between multiple forms of input. On the other hand, multiplicative interactions allow the input streams to modulate each other, providing a more accurate model.

We evaluate two multiplicative merging techniques to better model such interactions. Multiplicative fusion calculates element-wise products between inputs, while projecting the interactions into a lower-dimensional space to reduce dimensionality (Park et al., 2016). Multimodal factorized bilinear pooling takes advantage of optimizations in size and complexity, and is our final merged model (Yu et al., 2017). This method builds on the basic idea of multiplicative fusion by performing a sequence of pooling and regularization steps after the initial element-wise multiplication.

1.4 Results

We now present our training and evaluation methodology along with final results. For each of the datasets, we evaluate the network architectures described in Section 1.3.2. We compare against several baselines: a K Nearest Neighbor classifier, a linear SVM trained through SGD (Bottou, 2010), VoxNet (Maturana and Scherer, 2015), and SoundNet (Aytar et al., 2016). Our multimodal networks combined VoxNet with either ISNN-A or SoundNet8 and were merged through either concatenation (MergeCat), element-wise addition (MergeAdd), multiplicative fusion (MergeMultFuse) (Park et al., 2016), or multimodal factorized bilinear pooling (MergeMFB) (Yu et al., 2017). Training was performed using an Adam optimizer (Kingma and Ba, 2015) and run with a batch size of 64, with remaining hyperparameters hand-tuned on a validation set before final evaluation on a test set.

1.4.1 RSAudio Evaluation

Our "RSAudio" dataset was constructed from real and synthesized sounds. When performing geometry classification, each geometric model is its own class; given a query sound, the network returns the geometric model that would produce the most similar sound. RSAudio combines real and synthetic sounds to increase dataset size and improve accuracy.

Geometry Classification Accuracy: RSAudio and Related Work Datasets (ISNN-A Ours)

Method	Input	RSA S	RSA R	RSA Merged	Sound-20K*	Arnab A	ImageNet
Nearest Neighbor	A	**96.92%**	68.63%	97.59%	95.54%	87.50%	N/A
Linear SVM (Bottou, 2010)	A	2.31%	2.30%	3.20%	82.07%	7.14%	N/A
SoundNet5 (Aytar et al., 2016)	A	94.74%	16.10%	**97.70%**	58.81%	23.21%	N/A
SoundNet8 (Aytar et al., 2016)	A	83.83%	4.24%	89.62%	71.43%	58.93%	N/A
ISNN-A	A	96.74%	**92.37%**	97.07%	**99.52%**	**89.29%**	N/A
Pre-Trained VGG16	V	N/A	N/A	N/A	N/A	N/A	73.27%

Table 1.1: For real sounds, ISNN-A significantly outperforms all other methods, with an accuracy up to 92.37 %. For some synthetic datasets, ISNN-A produces results competitive with the top-performing methods. *Based on a subset of Sound-20K.

62 400 synthesized sounds come from a set of 59 geometric models and 11 sets of material parameters categorized into 6 classes of materials. For each model and material pairing (with a few exceptions), 100 sounds with random hit points were synthesized.

1183 real impact sounds come from a set of 24 struck rigid objects. These objects are each made of one homogeneous material and primarily consist of dining dishes, utensils, tools, and material samples used for building construction. A majority of the sounds were recorded in a padded sound booth using a Zoom H4 microphone to reduce background noise and room acoustics. The remaining sounds were recorded in a wider set of environments ranging from small offices to large outdoor areas. Each recording contains one impact in isolation from other impacts.

Objects were either struck with a small metal wrench or a rubber-headed drumstick, and in most cases, both. In either case, the striking tool was tightly gripped in a hand while striking in order to minimize its vibrations while the main struck object could vibrate freely. No post-processing was performed to attempt to remove the remaining sound from the striking tool.

The results for geometry classification are presented in Tables 1.1 to 1.4. For RSAudio synthetic (S) and real (R), ISNN-A provides competitive results with all other tested algorithms. For real sounds, where issues of recordings are most problematic, ISNN-A significantly outperforms all other algorithms, with an accuracy of 92.37 %. On the merged RSAudio dataset of real and synthetic sounds, all models actually produce *higher* accuracy than on either synthetic or real alone, indicating that training on both sets improves generalizability. As an additional baseline, we classified 100 ImageNet RGB transparent object images based on the VGG16 pre-trained model and obtained 73.27% accuracy with top 5 labels and an average confidence of 46.64%.

Geometry Classification Accuracy: Audio Methods (ISNN-A Ours), ModelNet

Method	Input	MN10o	MN10os	MN10om	MN10osm	MN10	MN40o	MN40osm
Nearest Neighbor	A	40.73%	32.42%	62.81%	67.97%	—	26.55%	54.41%
Linear SVM	A	16.67%	7.81%	28.85%	15.63%	11.73%	3.97%	12.18%
SoundNet5	A	16.96%	10.00%	10.70%	11.00%	—	4.10%	10.95%
SoundNet8	A	10.64%	19.50%	20.74%	29.67%	—	5.73%	49.27%
ISNN-A	A	**43.35%**	**56.50%**	**68.00%**	**71.50%**	**42.90%**	**32.51%**	**65.07%**

Table 1.2: Our audio-only ISNN-A outperforms other audio-only baselines.

Geometry Classification Accuracy: Visual Methods (All Baselines), ModelNet

Method	Input	MN10o	MN10os	MN10om	MN10osm	MN10	MN40o	MN40osm
Nearest Neighbor	V	83.11%	72.57%	82.62%	72.96%	—	65.72%	67.23%
Linear SVM	V	74.06%	66.80%	68.65%	77.34%	35.39%	51.15%	12.06%
VoxNet (Maturana and Scherer, 2015)	V			**89.47%**			**80.17%**	

Table 1.3: VoxNet achieves the highest level of accuracy compared to other alternative methods for geometry classification with visual input only.

While the accuracy is not directly comparable with ModelNet and RSAudio results, it provides a preliminary suggestion that a second modality could further improve results.

1.4.2 ModelNet Evaluation

In Tables 1.2 to 1.4, ModelNet results are categorized by input: audio (A), voxel (V), or both (AV). The "MN10" dataset consists of 119.620 total synthetic sounds: multiple sounds at different hit points for each geometry and material combination. The "o" suffix (*e.g.*, "MN10o") indicates that only one sound per model was produced, and all models were assigned one identical material. The "s" suffix (*e.g.*, "MN10os") indicates that each ModelNet class was assigned a realistic and normally distributed scale before synthesizing sounds. The "m" suffix (*e.g.*, "MN10om") indicates that each ModelNet class was assigned a realistic material.

By assigning a material and scale to each ModelNet10 class (MN10osm), classification performance achieved 71.50% for ISNN-A. Real-world objects within a class will tend to be made of a similar material and scale, so MN10osm is likely more reflective of performance in real-world settings where these factors provide increased potential for classification. However, for the multimodal ISNN-AV, material and scale assignments do not improve accuracy. In MN10o, larger geometric features will correspond to lower-pitched sounds (*i.e.*, a large object will produce a deeper sound than a small object), and the multimodal fusion of those cues produces higher accuracy. However, when models are given materials and scales in MN10o{s,m,sm}, the

Geometry Classification Accuracy: Audio-Visual Methods (ISNN-AV Ours), ModelNet

Method	Input	MN10o	MN10os	MN10om	MN10osm	MN10	MN40o	MN40osm
Nearest Neighbor	AV	82.91%	72.57%	83.40%	74.05%	—	65.84%	71.25%
Linear SVM	AV	80.63%	73.44%	82.50%	81.64%	36.70%	54.93%	66.15%
MergeCat (ISNN-AV)	AV	86.25%	78.50%	88.96%	88.50%	87.40%	79.93%	92.30%
MergeCat (SoundNet8)	AV	88.14%	52.50%	72.80%	54.50%	—	79.56%	56.39%
MergeAdd (ISNN-AV)	AV	88.91%	80.00%	88.52%	86.00%	88.27%	79.40%	90.43%
MergeAdd (SoundNet8)	AV	88.58%	50.50%	72.91%	64.33%	—	79.89%	24.43%
MergeMultFuse (ISNN-AV)	AV	89.14%	84.00%	89.41%	86.24%	87.51%	81.35%	**93.24%**
MergeMultFuse (SoundNet8)	AV	83.48%	66.00%	71.79%	51.67%	—	61.44%	38.97%
MergeMFB (ISNN-AV)	AV	**91.80%**	**84.50%**	**89.97%**	**90.12%**	**89.16%**	**82.04%**	92.51%
MergeMFB (SoundNet8)	AV	88.69%	76.50%	73.02%	42.00%	—	80.90%	91.33%

Table 1.4: Our merged networks produce accuracy upto 90.12 % on MN10osm and upto 93.24 % on MN40osm. Please visit ModelNet for more information on other methods and results.

voxel inputs remain unchanged, weakening the relationship between voxel and audio inputs. Scaling the voxel representation as well as the model used for sound synthesis may reduce this issue.

Assigning scale and material improve ModelNet40 accuracy (MN40osm) because its object classes differ more in size and material than ModelNet10. The merged audio-visual networks outperform the separate audio or visual networks in every case except for MN10os, as discussed above. Across all ModelNet10 datasets, ISNN-AV with multimodal factorized bilinear pooling produces the highest accuracy on MN10o, at 91.80 %. Similarly, ModelNet40 produces optimal results using ISNN-AV with multiplicative fusion on MN40osm, at 93.24 %. Entries with a "—" were not completed due to prohibitive time or memory costs when using the large MN10 dataset.

1.4.3 Additional Evaluations

We evaluated on additional audio-only datasets such as Arnab et. al (Arnab et al., 2015) and Sound-20K (Zhang et al., 2017), with results displayed in Table 1.1. The Arnab dataset consists of audio of tabletop objects being struck, with ground-truth object labels provided. ISNN-A produces 89.29 % geometry classification accuracy, the highest of all evaluated algorithms. This accuracy is slightly lower than ISNN-A's accuracy on RSAudio's real sounds, likely due to the loosened constraints on the recording environment and striking methodology.

Material Classification Accuracy

Model	RSAudio S	RSAudio R	Arnab Audio (Arnab et al., 2015)
ISNN-A	98.69%	**95.76%**	**71.86%**
SoundNet5 (Aytar et al., 2016)	**99.97%**	29.66%	43.11%
SoundNet8 (Aytar et al., 2016)	92.66%	30.51%	43.11%

Table 1.5: Material classification accuracy on subsets of the RSAudio dataset. Our ISNN-A network produces the highest accuracy on the two real-world datasets, with competitive accuracy on the synthetic RSAudio dataset

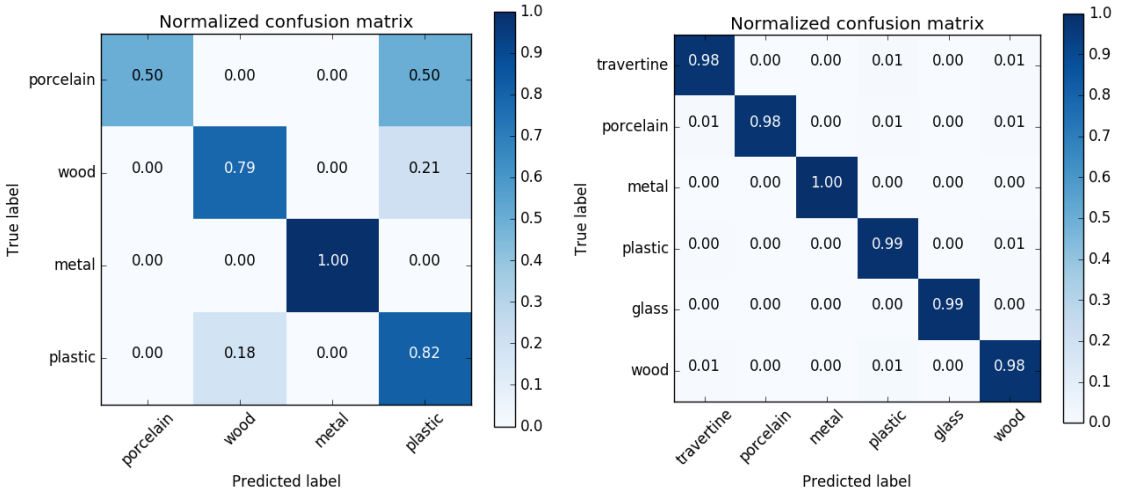
The Sound-20K dataset consists of impact sounds produced from a physics-based simulation, which may produce multiple impacts spread over time. ISNN-A produces 99.52 % geometry classification accuracy, again the highest accuracy of all evaluated algorithms.

1.4.3.1 Material Classification

The ISNN networks can also be trained for the task of material classification. That is, given an input impact sound, ISNN trained in this way will produce an estimate of the material class of the object. This is a task that has been more thoroughly evaluated by previous work, but we are still interested in the performance of ISNN on this same task.

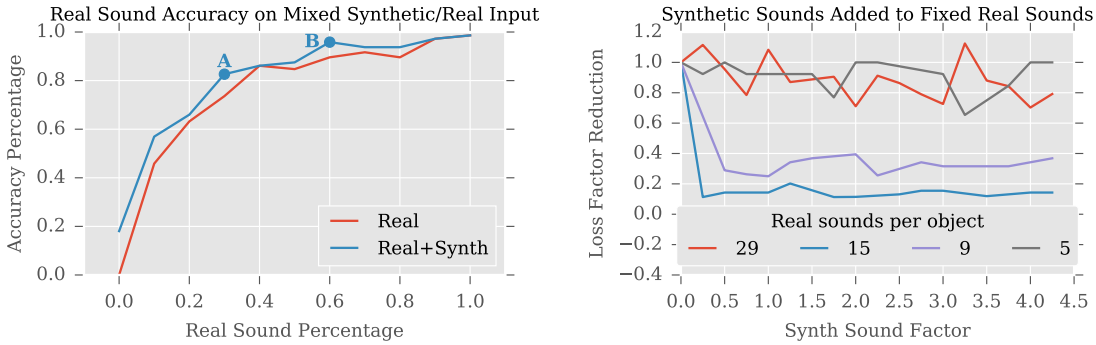
In Table 1.5, we compare the material classification accuracy of various classification models on multiple datasets. ISNN-A produces consistently high accuracy, up to 98.69 %, and is either competitive with or or outperforms SoundNet. The material labels provided with the Arnab dataset are not consistent with those listed in their publication, but we selected a subset of those labels with clearly-distinct material names for this test. In comparison to geometry classification (Table 1.2), material classification accuracies are a few percent higher on the RSA datasets, but somewhat lower on the Arnab dataset, likely due to the labeling discrepancies.

In Figure 1.7, we look at a breakdown of the classifications performed by ISNN on RSAudio's synthetic sounds and Arnab sounds. While RSAudio produces consistently accurate classifications with only minor error, the majority of misclassifications on the Arnab dataset come from porcelain classified as plastic.



(a) (b)

Figure 1.7: Material classification confusion matrices produced by ISNN-A on (a) the Arnab audio dataset and (b) the synthetic subset of RSAudio. In both cases, there is high accuracy with only a minimal amount of confusion.



(a) (b)

Figure 1.8: Classification accuracy on a test set of real sounds using ISNN trained on a combination of real and synthetic sounds. (a) When trained on combined real and synthetic sounds (Real+Synth), classification accuracy is upto 11% higher than when trained on the real sounds alone (Real). (b) When insufficient real sounds are provided, synthetic sounds further reduce loss.

1.4.3.2 Combined Real and Synthetic Training

We also evaluate the ability of synthetic sounds to supplement a smaller number of real sounds for training, which would reduce necessary human effort in obtaining sounds. Figure 1.8 shows classification accuracy on a real subset of our RSAudio dataset for ISNN-A trained on a combination of real and synthetic sounds. The training sets have identical total sizes but are created with specific percentages of real and synthetic sounds, then networks are trained on either the combined dataset or the real sounds independently. We find that the addition of synthetic sounds to the dataset improves accuracy by up to 11 %. With only 30 % real sounds (Point A), accuracy begins to plateau, reaching over 90 % accuracy with only 60 % real sounds (Point B). These indicate that synthetic audio can supplement a smaller amount of recorded audio to improve accuracy.

Augmentations in Section 1.2.2 were designed to enhance the realism of synthetic audio for improved transfer learning from synthetic to real sounds. However, we were unable to find an instance when these augmentations significantly improved test accuracy of RSAudio real when trained on RSAudio synthetic. This indicates that *modal* components of sounds (frequencies, amplitudes) are sufficient and most critical in object classification, and that acoustic radiance, noise, and propagation effects produce little, if any, impact on accuracy.

1.4.3.3 Activation Maximization

We additionally use activation maximization to visualize the spectrogram inputs that would produce the highest activation for a given ModelNet class. Figure 1.9 shows how the result of activation maximization changes as different modifications to ModelNet sounds are performed. When no scale or material are applied, the maximized spectrogram demonstrates a need for robustness to variance in frequency and damping. When scale is fixed, so is the fundamental frequency, as can be seen by the single active region and lower overall activation weights. When material is fixed, so are the damping rates, which become recognizable identifiers for this particular class of object.

1.4.4 Application: Audio-Guided 3D Reconstruction

A primary use case of the ISNN networks is to improve reconstruction of transparent, occluded, or reflective objects. Existing methods have become very effective at 3D scene reconstruction from RGB-D

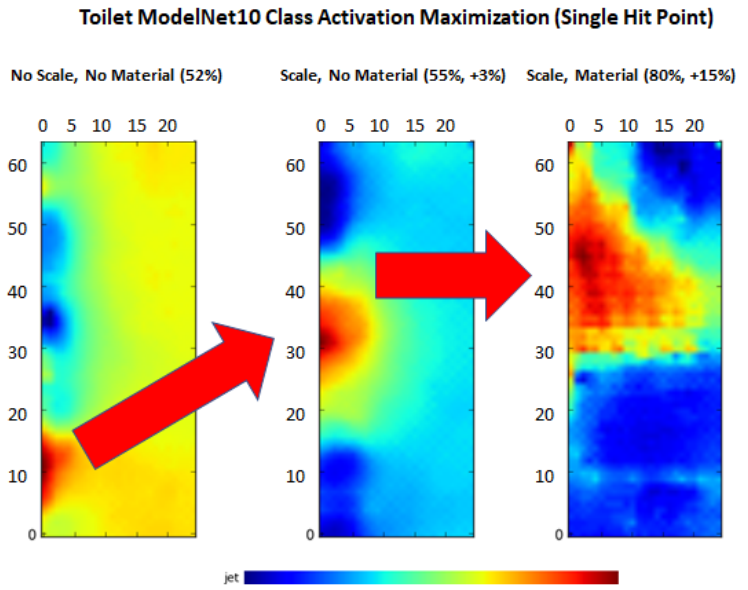


Figure 1.9: Activation maximization results for the Toilet class of ModelNet10 as different modifications are made to the model for sound synthesis. When both scale and material are not fixed as distinguishing factors, the network must be general and robust to differences (left). When both are fixed, the network clearly identifies a recognizable pattern (right).

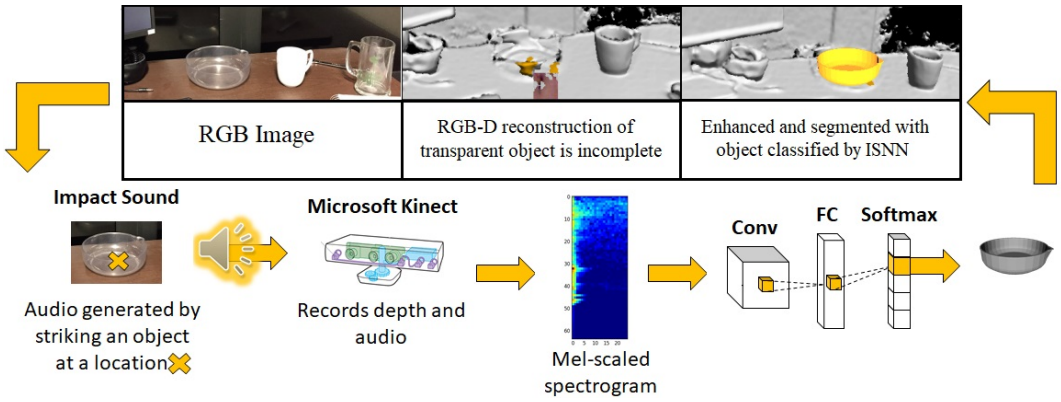


Figure 1.10: A user strikes a real-world object to generate sound as an input into our ISNN network which returns material and object classification. Based on these, the real-time 3D reconstruction is enhanced and segmented.

video, even in real-time. However, due to limitations of vision-based methods, transparent and occluded objects are still a challenge for these methods. We have constructed a demo utility in which our method enables real-time scene reconstruction and augmentation. Figure 1.10 illustrates the application pipeline.

1.4.4.1 Algorithm

The utility at its simplest provides a system for real-time scene reconstruction, based on previous real-time RGB-D work (Golodetz* et al., 2015; Valentin et al., 2015). Using the RGB-D camera of a Kinect, a user scans the scene from multiple angles until estimations have sufficiently converged. At this point, transparent objects may be incomplete or missing. The user interacts with the application to select one of these objects, then physically reaches into the scene to strike the corresponding object.

The Kinect's microphone array records the impact sound, identifies the time of impact, and extracts a 1-second clip containing the sound and its decay. The recorded audio waveform is converted to the form of input to the ISNN-A network: a downsampled mel-scaled spectrogram. This spectrogram is passed through ISNN-A trained on the full RSAudio dataset. One network trained to perform geometric model classification identifies the closest matching geometry to the recorded sound, while another network trained to perform material classification identifies the closest matching material class.

The full object can then be inserted into the reconstructed scene. The object is inserted at the position earlier selected, using the classified geometric model. In our reconstruction utility, the object is textured with a different color than that of the original geometry, indicating the segmentation of the object from the rest of the scene. Alternatively, the material classification could correspond to a texture which could be applied to the object. As a result of this process, the transparent object that had previously been incomplete or missing, has been both completed and segmented.

1.4.4.2 Utility Limitations

ISNN's geometric model classification cannot interpolate or extrapolate geometry given new sounds. When ISNN is trained on the RSAudio dataset, each individual geometric model is considered to be its own class, and classification of a test sound is selection of the closest *training* geometry to that sound. For the utility, this means that the inserted geometric model may be similar to the ground truth object, but not match exactly. Shape optimization from sound is still an open area of research. We have also tested pose estimation methods based on RGB (Brachmann et al., 2016) and RGB-D (Lysenkov and Rabaud, 2013; Lysenkov et al.,

2013); however, future work is needed to extend these to accept asymmetric transparent objects as input and integrate into our application.

1.5 Summary

We presented a novel approach for improving the reconstruction of 3D objects using audio-visual data. Given impact sound as an additional input, ISNN-A and ISNN-AV have been optimized to achieve high accuracy on object classification tasks. The use of spectrogram representations of input reduce overfitting by directly inputting spectral information to the networks. ISNN has further shown higher performance when using a dataset with combined synthetic and real audio. Sound provides additional cues, allowing us to estimate the object's material class, provide segmentation, and enhance scene reconstruction.

Limitations and Future Work: While VoxNet serves as a strong baseline for the visual component of ISNN-AV, different visual networks in its place could identify more optimal network pairings. As with existing learning methods, VoxNet is limited to performing classifications of known geometries. However, impact sounds hold potential of identifying correct geometry, even when a model database is not provided, allowing for accurate 3D reconstructions or hole-filling.

BIBLIOGRAPHY

Arnab, A., Sapienza, M., Golodetz, S., Valentin, J., Miksik, O., Izadi, S., and Torr, P. H. S. (2015). Joint object-material category segmentation from audio-visual cues. In Xianghua Xie, M. W. J. and Tam, G. K. L., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 40.1–40.12. BMVA Press.

Aytar, Y., Vondrick, C., and Torralba, A. (2016). SoundNet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900.

Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2546–2554. Curran Associates, Inc.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Lechevallier, Y. and Saporta, G., editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France. Springer.

Brachmann, E., Michel, F., Krull, A., Yang, M. Y., Gumhold, S., and Rother, C. (2016). Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3364–3372.

Cowling, M. and Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, 24(15):2895 – 2907.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 580–587, Washington, DC, USA. IEEE Computer Society.

Golodetz*, S., Sapienza*, M., Valentin, J. P. C., Vineet, V., Cheng, M.-M., Arnab, A., Prisacariu, V. A., Kähler, O., Ren, C. Y., Murray, D. W., Izadi, S., and Torr, P. H. S. (2015). SemanticPaint: A Framework for the Interactive Segmentation of 3D Scenes. Technical Report TVG-2015-1, Department of Engineering Science, University of Oxford. Released as arXiv e-print 1510.03727.

Huzaifah, M. (2017). Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. *CoRR*, abs/1706.07156.

James, D. L., Barbič, J., and Pai, D. K. (2006). Precomputed acoustic transfer: output-sensitive, accurate sound generation for geometrically complex vibration sources. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 987–995. ACM.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Lysenkov, I., Eruhimov, V., and Bradski, G. (2013). Recognition and pose estimation of rigid transparent objects with a kinect sensor. In *Robotics: Science and Systems Conference (RSS)*, volume 273.

Lysenkov, I. and Rabaud, V. (2013). Pose estimation of rigid transparent objects in transparent clutter. In *2013 IEEE International Conference on Robotics and Automation*, pages 162–169.

Maturana, D. and Scherer, S. (2015). VoxNet: A 3D convolutional neural network for real-time object recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, page 922 – 928.

Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., and Freeman, W. T. (2016). Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2405–2413.

Park, E., Han, X., Berg, T. L., and Berg, A. C. (2016). Combining multiple sources of knowledge in deep CNNs for action recognition. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE.

Piczak, K. J. (2015a). Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

Piczak, K. J. (2015b). ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 1015–1018, New York, NY, USA. ACM.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc.

Schissler, C. and Manocha, D. (2011). GSound: Interactive sound propagation for games. In *Audio Engineering Society Conference: 41st International Conference: Audio for Games*.

Socher, R., Huval, B., Bhat, B., Manning, C. D., and Ng, A. Y. (2012). Convolutional-recursive deep learning for 3D object classification. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 656–664, USA. Curran Associates Inc.

Sterling, A., Wilson, J., Lowe, S., and Lin, M. C. (2018). ISNN: Impact sound neural network for audio-visual object classification. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 578–595, Cham. Springer International Publishing.

Thiemann, J., Ito, N., and Vincent, E. (2013). The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings. *Proceedings of Meetings on Acoustics*, 19(1):035081.

Valentin, J., Vineet, V., Cheng, M.-M., Kim, D., Shotton, J., Kohli, P., Niessner, M., Criminisi, A., Izadi, S., and Torr, P. H. S. (2015). SemanticPaint: Interactive 3D Labeling and Learning at your Fingertips. *ACM Transactions on Graphics*, 34(5).

Wu, Z., Song, S., Khosla, A., and Xiao, J. (2015). 3D ShapeNets: A deep representation for volumetric shapes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920.

Yu, Z., Yu, J., Fan, J., and Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *IEEE International Conference on Computer Vision (ICCV)*, pages 1839–1848.

Zhang, Z., Wu, J., Li, Q., Huang, Z., Traer, J., McDermott, J. H., Tenenbaum, J. B., and Freeman, W. T. (2017). Generative modeling of audible shapes for object perception. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1260–1269.