This ECCV 2018 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ECCV 2018
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/eccv

ISNN: Impact Sound Neural Network for Audio-Visual Object Classification

Auston Sterling[1], Justin Wilson[1], Sam Lowe[1], and Ming C. Lin[1,2]

[1] Department of Computer Science, University of North Carolina at Chapel Hill
{austonst,wilson,samlowe,lin}@cs.unc.edu
[2] Department of Computer Science, University of Maryland, College Park
lin@cs.umd.edu

Abstract. 3D object geometry reconstruction remains a challenge when working with transparent, occluded, or highly reflective surfaces. While recent methods classify shape features using raw audio, we present a multimodal neural network optimized for estimating an object's geometry and material. Our networks use spectrograms of recorded and synthesized object impact sounds and voxelized shape estimates to extend the capabilities of vision-based reconstruction. We evaluate our method on multiple datasets of both recorded and synthesized sounds. We further present an interactive application for real-time scene reconstruction in which a user can strike objects, producing sound that can instantly classify and segment the struck object, even if the object is transparent or visually occluded.

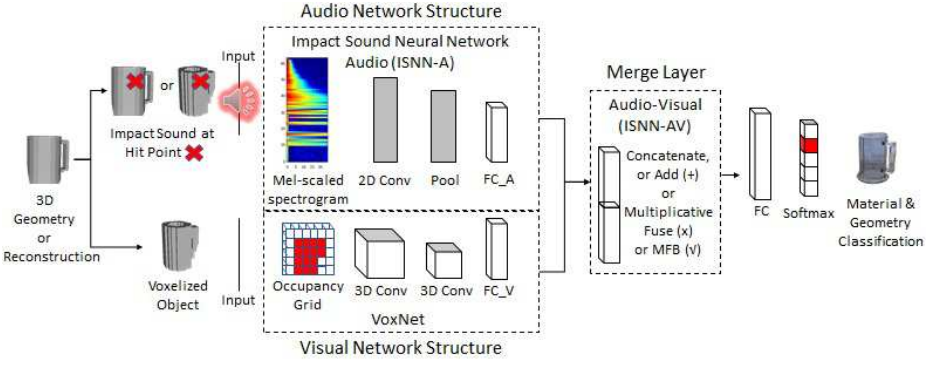


Fig. 1: Our Impact Sound Neural Network - Audio (ISNN-A) uses as input a spectrogram of sound created by a real or synthetic object being struck. Our audio-visual network (ISNN-AV) combines ISNN-A with VoxNet to produce state-of-the-art object classification accuracy.

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 [1,2]. Classification techniques have steadily improved, advancing our ability to accurately label an object by class given its depth image [3], voxelization [4], and/or RGB-D data [5]. Segmentation of objects from scenes provides contextual understanding of scenes [6,7]. 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 [8,9]. However, previous work has not focused on identifying complete object geometries. Identifying object geometry from a combined audiovisual approach expands the capabilities of scene understanding.

In this paper, 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 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: In this paper, 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 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.

2 Previous Work

3D Object Datasets Thanks to a plethora of 3D scene and object datasets such as BigBIRD[10] and RGB-D Object Dataset [11], neural network models have been trained to label objects based on their visual representation. 3D ShapeNets [3] also provides two sets of object categories for object classification referred to as ModelNet10 and ModelNet40, which are common benchmarks for evaluation [12]. Scene-based datasets have also been built from RGB-D reconstruction scans of entire spaces, allowing for semantic data such as object and room relationships. For instance, NYU Depth Dataset [13] and SUNCG [14] enable indoor segmentation and semantic scene completion from depth images.

2.1 3D Reconstruction

Structure from Motion (SFM) [15], Multi-View Stereo (MVS) [16], and Shape from Shading [17] are all techniques to estimate shape properties of a scanned scene. Although these methods alone do not achieve a segmented representation of the objects within the scene, they serve as a foundation for many algorithms. RGB-D depth-based, active reconstruction methods can also be used to generate 3D geometrical models of static [18,6] and dynamic [19,20] scenes using commodity sensors such as the Microsoft Kinect and GPU hardware in real-time. Techniques have also been developed to overcome some limitations of vision-based reconstruction techniques [21] such as scene lighting, occlusions, clutter, and overlapping transparent objects. When limited solely to visual input, these challenges remain. Additional modalities, such as the impact sounds we explore in this paper, have the potential to address these issues.

Alternate Modalities While image and depth-based techniques cover the majority of reconstruction use cases, edge cases have motivated research to explore alternative modalities that may procure the level of detail that vision-based techniques alone cannot. The dip transform for 3D shape reconstruction [22] uses fluid displacement of an object to obtain shape information. Time-of-Flight cameras introduce another modality to better classify materials and correct the depth of transparent objects [23]. This work uses both recorded and synthetic audio as additional modalities to complement vision-based reconstruction.

2.2 Environmental Sound Classification

Audio descriptors have been primarily explored in the context of environmental sound classification. Multiple datasets have been established for evaluating classification of various environmental sounds [24,25,26]. Traditional techniques use a variety of features extracted from sounds, such as Mel frequency spectral coefficients and spectral shape descriptors [27,28]. Similar approaches are used to classify an environment based on the sounds heard within it [29].

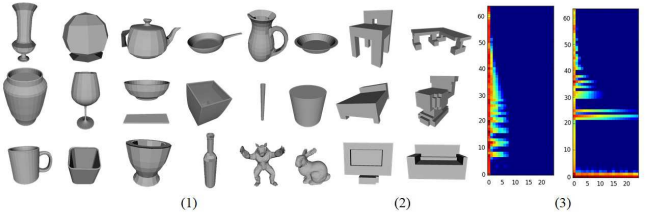


Fig. 2: We use various datasets for training and testing: (1) our RSAudio dataset with real and synthesized impact sounds from objects of varying shapes and sizes and (2) voxelized ModelNet objects. (3) Audio inputs are formatted as spectrograms.

Convolutional neural networks have also been applied to these problems, producing improved results [30,31]. Recently, some interest has been given to exploring the performance of different network structures [32,33]. Impact sounds are a specific category of environmental sounds, and in this paper we perform fine-grained object classification between perceptually similar sounds.

2.3 Object/Scene Understanding Through Sound

Sound can be used as a source of information for deeper understanding of 3D scenes and objects. Specifically considering impact sounds, sound can be used to estimate the material of objects using iterative optimization-based parameter identification techniques [34,35]. Sound has also been used to obtain information about the physical properties of objects involved in impact simulations [36]. Shape primitives were included as a part of these physical properties, but were not representative of real-world object geometries. However, it has been proven that any given impact sound may have come from one of multiple possible object geometries, and thus cannot be uniquely reconstructed [37]. Previous work has not attempted complete object reconstruction. In contrast, we constrain the outputs to known objects, making the problem tractable in this work.

Sound and video are intrinsically linked modalities for understanding the same scene, object, or event. Using visual and audio information, it is possible to predict the sound corresponding to a visual image or video [38,39]. Sound prediction from video has also been specifically explored for impact sounds [40].

Multimodal Fusion Other works have fused audio and visual cues to better understand objects and scenes. Sparse auditory clues can supplement the ability of random fields to obtain material labels and perform segmentation [9]. Neural networks have proven valuable in fusing audio-visual input to emulate the sensory interactions of human information processing [8]. While multimodal methods have succeeded in fusing input streams to capture material and low-level shape properties to aid segmentation, they have not attempted to identify specific object geometries.

Early attempts at multimodal fusion in neural networks focused on increasing classification specificity by combining the individual classification results of separate input streams [41]. Bilinear modeling can model the multiplicative interactions of differing input types, and has been applied as a method of pooling input streams in neural networks [42,43]. Bilinear methods have been further developed to reduce complexity and increase speed, while other approaches to modeling multiplicative interactions have also been explored [44,45,46]. Bilinear methods have not yet been applied to merging audio-visual networks, and our ISNN-AV network is the first to do so.

3 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 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 [24], while ESC-50 has specific categories including "Door knock" and "Church bells" [25]. The *Greatest Hits* sound dataset comes closest to our needs, containing impact sounds labeled according to the type of object [40]. 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 3.

3.1 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. Modal sound synthesis is an established method for synthesizing these sounds. We refer readers to the Supplemental Document (Section 1) for a mathematical overview and previous work for the full derivation of the algorithm [47,48,8]. Modal sound synthesis can be broken up into two steps: a preprocessing *modal analysis* step to process the inputs and a faster *modal synthesis* step to synthesize individual sounds.

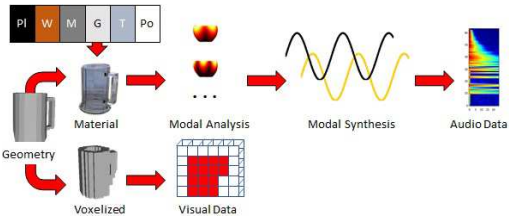


Fig. 3: We build multimodal datasets through separate processing flows. Modal sound synthesis produces spectrograms used for audio input. Voxelization as another modality provides a first estimate of shape. Incorporating audio features improves classification accuracy through understanding of how objects vibrate.

Modal Analysis Modal analysis is a process for modeling and understanding the vibrations of objects in response to external forces. Vibrations in an object can be modeled with the wave equation [49], but in order to handle arbitrary geometries with unknown analytical solutions, it is more common to perform finite element analysis on a discretized representation of the object [50,47].

Starting with a watertight triangle mesh representation of the object's surface, the interior volume of the object is filled with a tetrahedral volumetric mesh. A finite element model can then be constructed to represent the free vibrations of the object. Damping within the object causes vibrations to decay over time; we use Rayleigh damping to model this effect.

Given this representation of a vibrating object, we are interested in determining the frequencies at which it vibrates. This can be accomplished through a generalized eigen-decomposition of the finite element matrices. Finally, the system can be decoupled into linearly separable *modes of vibration*. Each mode has a solution in the form of a damped sinusoid, each having a different frequency and rate of decay. This modal analysis step is performed once per object, and is a computationally-intensive task. The resulting modes' frequencies of vibration and damping rates are saved to be used in modal synthesis.

Modal Synthesis Striking an object excites its modes of vibration, causing a change in pressure waves. For a simulated object, an impulse in object-space can be converted to mode impulses to determine initial amplitudes for the corresponding sinusoids. The sinusoids for the modes are then sampled through time and added to produce the final sound. This process can be repeated for different materials, geometries, and hit points to create a set of synthetic impact sounds.

3.2 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 [51]. 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 [52]. 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 [53]. 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.

3.3 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 5.2).

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

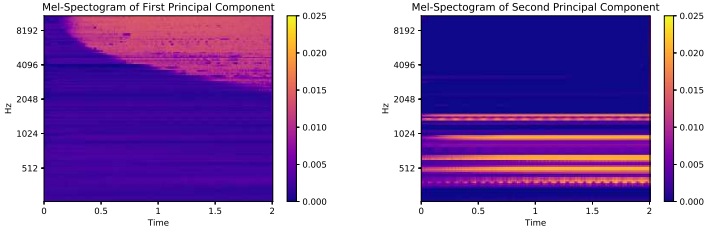
Given the impact sounds and representation described in Section 3, we now examine their ability to identify materials and geometric models. We begin with an analysis of the distributions of the features themselves as proper feature selection is a key component in classifier construction.

4.1 Input Features and Analysis

Audio Features In environmental sound classification tasks, classification accuracy can be affected by the input sound's form of representation [28,33]. A one-dimensional time series of audio samples over time can be used as features [39], 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 [33]. 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. 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 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.



(a) (b)

Fig. 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).

Visual Features As in VoxNet [4], 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.

4.2 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.

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 optimization was performed using the TPE algorithm [54]. 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 [30]. Our generally low number of filters and narrower layer sizes aim to reduce overfitting by encouraging the learning of generalizable geometric properties.

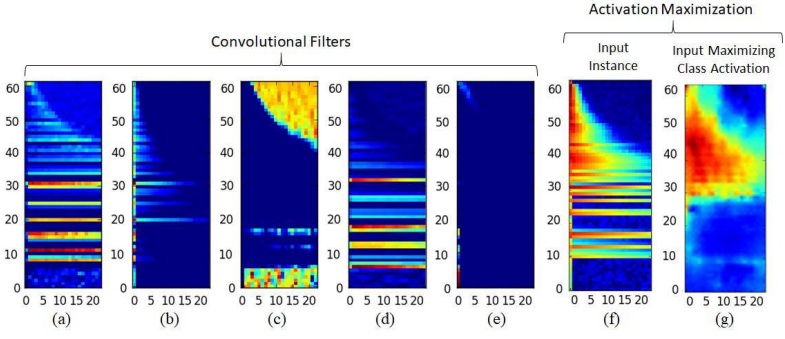


Fig. 5: Sample activations (a-e) of ISNN convolution layer. Filters identify characteristic patterns in frequencies (a) (d), damping rates (b) (c), and high-frequency noise (e). The distinguishing characteristics in these activations match the expected factors discovered in the PCA analysis in Figure 4. An audio input spectrogram (f) and activation maximization (g) learned by the ISNN network for the toilet ModelNet10 class show correctly-learned patterns.

Figure 5 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 4a, we see that damping is an important discriminating feature, which has been learned by filters

(b) and (c) in Figure 5. Similarly, the frequency patterns that we expected because of Figure 4b can be seen in filters (a) and (d). This demonstrates that our model is learning statistically optimal kernels with high discriminatory power.

Multimodal Audio-Visual Network (ISNN-AV) Our audio-visual network, as shown in Figure 1, consists of our audio-only network combined with a visual network based on VoxNet [4] 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 [46]. Multimodal factorized bilinear pooling takes advantage of optimizations in size and complexity, and is our final merged model [45]. 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.

5 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 4.2. We compare against several baselines: a K Nearest Neighbor classifier, a linear SVM trained through SGD [55], VoxNet [4], and SoundNet [39]. 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) [46], or multimodal factorized bilinear pooling (MergeMFB) [45]. Training was performed using an Adam optimizer [56] and run with a batch size of 64, with remaining hyperparameters hand-tuned on a validation set before final evaluation on a test set.

5.1 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. Specific details about the recording process for the real sounds can be found in Supplemental Document (Section 4). The dataset is publicly available at http://gamma.cs.unc.edu/ISNN/.

The results for geometry classification are presented in Tables 1, 2, 3, and 4. For RSAudio synthetic (S) and real (R), ISNN-A provides competitive results

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 [55]	A	2.31%	2.30%	3.20%	82.07%	7.14%	N/A
SoundNet5 [39]	A	94.74%	16.10%	**97.70%**	58.81%	23.21%	N/A
SoundNet8 [39]	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: For real sounds, ISNN-A significantly outperforms all other algorithms, with an accuracy upto 92.37 %. *Based on a subset of Sound-20K.

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%. 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.

5.2 ModelNet Evaluation

In Tables 2, 3, and 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.

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 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 [4]	V			**89.47%**			**80.17%**	

Table 3: VoxNet [4] achieves the highest level of accuracy compared to other alternative methods for geometry classification with visual input only.

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 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.

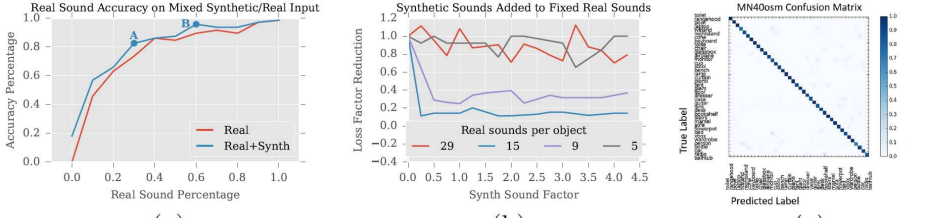
By assigning a material and scale to each ModelNet10 class (MN10osm), classification performance achieved 71.50% for ISNN-A. Real-world objects within a class will tend to be made of a similar material and scale, so MN10osm is likely more reflective of performance in real-world settings where these factors 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 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.

5.3 Additional Evaluations

We evaluated on additional datasets such as Arnab et. al [9]. This dataset consists of audio of tabletop objects being struck, with ground-truth object labels provided. ISNN-A produces 89.29 % 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. The same networks were also considered for material classification, and results can be seen in Supplemental Document (Section 6).

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 6 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.



(a) (b) (c)

Fig. 6: 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. (c) Our method has been able to correctly classify impact sounds with voxel data across ModelNet40 classes, as displayed by the MN40osm confusion matrix, for instance.

Augmentations in subsection 3.2 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.

5.4 Application: Audio-Guided 3D Reconstruction

A primary use case of the method described in this paper is to improve reconstruction of transparent, occluded, or reflective objects. We have constructed a demo application in which our method enables real-time scene reconstruction and augmentation. We enhance open-source 3D reconstruction software [6,7] by adding an audio-based selector function. Figure 7 illustrates the application pipeline. Further details are in the Supplemental Document (Section 2) and demo video at http://gamma.cs.unc.edu/ISNN/.

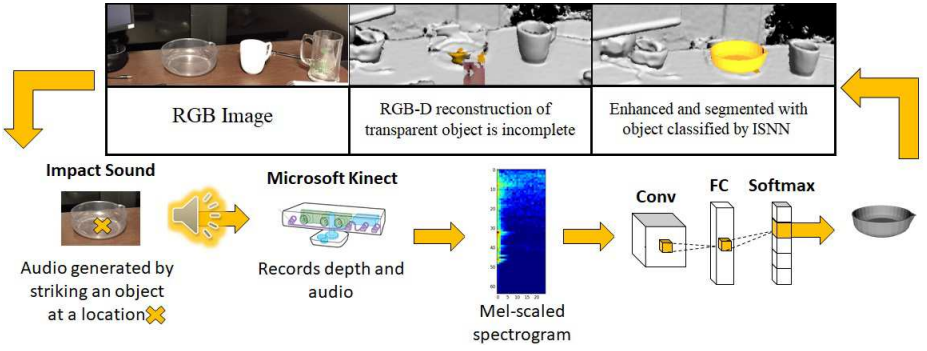


Fig. 7: A user strikes a real-world object to generate sound as an input into our ISNN network which returns material and object classification. Based on these, the real-time 3D reconstruction [6,7] is enhanced and segmented.

6 Conclusion

In this paper, we have 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.

References

1. Girshick, R., Donahue, J., Darrell, T., Malik, J.: 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, Washington, DC, USA, IEEE Computer Society (2014) 580–587 2
2. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS). (2015) 2
3. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D ShapeNets: A Deep Representation for Volumetric Shape Modeling. Proceedings of 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 2, 3
4. Maturana, D., Scherer, S.: VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition. In: IROS. (2015) 2, 8, 10, 12
5. Socher, R., Huval, B., Bhat, B., Manning, C.D., Ng, A.Y.: Convolutional-Recursive Deep Learning for 3D Object Classification. Conference on Neural Information Processing Systems (NIPS) (2012) 2
6. 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., Torr, P.H.S.: SemanticPaint: A Framework for the Interactive Segmentation of 3D Scenes. Technical Report TVG-2015-1, Department of Engineering Science, University of Oxford (October 2015) Released as arXiv e-print 1510.03727. 2, 3, 14
7. Valentin, J., Vineet, V., Cheng, M.M., Kim, D., Shotton, J., Kohli, P., Niessner, M., Criminisi, A., Izadi, S., Torr, P.H.S.: SemanticPaint: Interactive 3D Labeling and Learning at your Fingertips. ACM Transactions on Graphics **34**(5) (2015) 2, 14
8. Zhang, Z., Wu, J., Li, Q., Huang, Z., Traer, J., McDermott, J.H., Tenenbaum, J.B., Freeman, W.T.: Generative modeling of audible shapes for object perception. In: The IEEE International Conference on Computer Vision (ICCV). (2017) 2, 4, 5
9. Arnab, A., Sapienza, M., Golodetz, S., Valentin, J., Miksik, O., Izadi, S., Torr, P.H.S.: Joint object-material category segmentation from audio-visual cues. In: Proceedings of the British Machine Vision Conference (BMVC). (2015) 2, 4, 13
10. Singh, A., Sha, J., Narayan, K.S., Achim, T., Abbeel, P.: BigBIRD: A Large-Scale 3D Database of Object Instances. IEEE International Conference on Robotics and Automation (ICRA) (2014) 3
11. Lai, K., Bo, L., Ren, X., Fox, D.: A Large-Scale Hierarchical Multi-View RGB-D Object Dataset. IEEE International Conference on Robotics and Automation (ICRA) (2011) 3
12. Kanezaki, A., Matsushita, Y., Nishida, Y.: Rotationnet: Learning object classification using unsupervised viewpoint estimation. CoRR **abs/1603.06208** (2016) 3
13. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. (2012) 3
14. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition (2017) 3
15. Westoby, M., Brasington, J., Glasser, N., Hambrey, M., Reynolds, J.: structure-from-motion photogrammetry: A low-cost, effective tool for geoscience applications. Geomorphology **179** (2012) 300 – 314 3

16. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2006) 3
17. Zhang, R., Tsai, P.S., Cryer, J.E., Shah, M.: Shape-from-shading: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(8) (Aug 1999) 690–706 3
18. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: KinectFusion: Real-Time Dense Surface Mapping and Tracking. International Symposium on Mixed and Augmented Reality (ISMAR) (2011) 3
19. Newcombe, R., Fox, D., Seitz, S.: DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time. Proceedings of 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 3
20. Dai, A., Niessner, M., Zollhofer, M., Izadi, S., Theobalt, C.: BundleFusion: Real-time Globally Consistent 3D Reconstruction using On-the-fly Surface Reintegration. SIGGRAPH (2017) 3
21. Lysenkov, I., Eruhimov, V., Bradski, G.: Recognition and pose estimation of rigid transparent objects with a kinect sensor. In: Robotics: Science and Systems Conference (RSS). (2013) 3
22. Aberman, K., Katzir, O., Zhou, Q., Luo, Z., Sharf, A., Greif, C., Chen, B., Cohen-Or, D.: Dip Transform for 3D Shape Reconstruction. SIGGRAPH (2017) 3
23. Tanaka, K., Mukaigawa, Y., Funatomi, T., Kubo, H., Matsushita, Y., Yagi, Y.: Material Classification using Frequency- and Depth-dependent Time-of-Flight Distortion. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. (July 2017) 79–88 3
24. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: Proc. IEEE ICASSP 2017, New Orleans, LA (2017) 3, 5
25. Piczak, K.J.: Esc: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM International Conference on Multimedia. MM '15, New York, NY, USA, ACM (2015) 1015–1018 3, 5
26. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: Proceedings of the 22Nd ACM International Conference on Multimedia. MM '14, New York, NY, USA, ACM (2014) 1041–1044 3
27. Büchler, M., Allegro, S., Launer, S., Dillier, N.: Sound classification in hearing aids inspired by auditory scene analysis. EURASIP Journal on Advances in Signal Processing **2005**(18) (Nov 2005) 387845 3
28. Cowling, M., Sitte, R.: Comparison of techniques for environmental sound recognition. Pattern Recognition Letters **24**(15) (2003) 2895 – 2907 3, 7
29. Barchiesi, D., Giannoulis, D., Stowell, D., Plumbley, M.D.: Acoustic scene classification: Classifying environments from the sounds they produce. IEEE Signal Processing Magazine **32**(3) (May 2015) 16–34 3
30. Piczak, K.J.: Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). (Sept 2015) 1–6 4, 9
31. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters **24**(3) (March 2017) 279–283 4
32. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson,

K.: Cnn architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (March 2017) 131–135 4
33. Huzaifah, M.: Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. CoRR **abs/1706.07156** (2017) 4, 7, 8
34. Ren, Z., Yeh, H., Lin, M.C.: Example-guided physically based modal sound synthesis. ACM Trans. Graph. **32**(1) (February 2013) 1:1–1:16 4
35. Sterling, A., Lin, M.C.: Interactive modal sound synthesis using generalized proportional damping. In: Proceedings of the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games. I3D '16, New York, NY, USA, ACM (2016) 79–86 4
36. Zhang, Z., Li, Q., Huang, Z., Wu, J., Tenenbaum, J., Freeman, B.: Shape and material from sound. In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., eds.: Advances in Neural Information Processing Systems 30. Curran Associates, Inc. (2017) 1278–1288 4
37. Kac, M.: Can one hear the shape of a drum? The American Mathematical Monthly **73**(4) (1966) 1–23 4
38. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A. In: Ambient Sound Provides Supervision for Visual Learning. Springer International Publishing, Cham (2016) 801–816 4
39. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems. (2016) 892–900 4, 7, 10, 11
40. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2405–2413 4, 5
41. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems 27. Curran Associates, Inc. (2014) 568–576 5
42. Tenenbaum, J.B., Freeman, W.T.: Separating style and content with bilinear models. Neural Comput. **12**(6) (June 2000) 1247–1283 5
43. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: International Conference on Computer Vision (ICCV). (2015) 5
44. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 317–326 5
45. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. IEEE International Conference on Computer Vision (ICCV) (2017) 1839–1848 5, 10
46. Park, E., Han, X., Berg, T.L., Berg, A.C.: Combining multiple sources of knowledge in deep cnns for action recognition. In: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, IEEE (2016) 1–8 5, 10
47. O'Brien, J.F., Shen, C., Gatchalian, C.M.: Synthesizing sounds from rigid-body simulations. In: Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. SCA '02, New York, NY, USA, ACM (2002) 175–181 5, 6
48. Raghuvanshi, N., Lin, M.C.: Interactive sound synthesis for large scale environments. In: Proceedings of the 2006 Symposium on Interactive 3D Graphics and Games. I3D '06, New York, NY, USA, ACM (2006) 101–108 5

49. van den Doel, K., Pai, D.K.: The sounds of physical shapes. Presence **7** (1996) 382–395 6
50. Morrison, J.D., Adrien, J.M.: Mosaic: A framework for modal synthesis. Computer Music Journal **17**(1) (1993) 45–56 6
51. James, D.L., Barbič, J., Pai, D.K.: Precomputed acoustic transfer: output-sensitive, accurate sound generation for geometrically complex vibration sources. In: ACM Transactions on Graphics (TOG). Volume 25., ACM (2006) 987–995 7
52. Schissler, C., Manocha, D.: Gsound: Interactive sound propagation for games. In: Audio Engineering Society Conference: 41st International Conference: Audio for Games. (Feb 2011) 7
53. Thiemann, J., Ito, N., Vincent, E.: The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. Proceedings of Meetings on Acoustics **19**(1) (2013) 035081 7
54. Bergstra, J.S., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems 24. Curran Associates, Inc. (2011) 2546–2554 9
55. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In Lechevallier, Y., Saporta, G., eds.: Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010), Paris, France, Springer (August 2010) 177–187 10, 11
56. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR). (2015) 10