Generative Modeling of Audible Shapes for Object Perception

Zhoutong Zhang*[1] Jiajun Wu*[1] Qiujia Li[2] Zhengjia Huang[3] James Traer[1]
Josh H. McDermott[1] Joshua B. Tenenbaum[1] William T. Freeman[1,4]

[1]Massachusetts Institute of Technology [2]University of Cambridge
[3]ShanghaiTech University [4]Google Research

Abstract

Humans infer rich knowledge of objects from both auditory and visual cues. Building a machine of such competency, however, is very challenging, due to the great difficulty in capturing large-scale, clean data of objects with both their appearance and the sound they make. In this paper, we present a novel, open-source pipeline that generates audio-visual data, purely from 3D object shapes and their physical properties. Through comparison with audio recordings and human behavioral studies, we validate the accuracy of the sounds it generates. Using this generative model, we are able to construct a synthetic audio-visual dataset, namely Sound-20K, for object perception tasks. We demonstrate that auditory and visual information play complementary roles in object perception, and further, that the representation learned on synthetic audio-visual data can transfer to real-world scenarios.

1. Introduction

Humans perceive objects through both their visual appearance and the sounds they make. Given a short audio clip of objects interacting, humans, including young children, can recover rich information about the materials, surface smoothness, and the quantity of objects involved [46, 22, 35]. Although visual information provides cues for some of these questions, others can only be assessed with sound. Figure 1 shows an example: objects with different masses and Young's moduli may have almost identical appearance, but they make different sounds when impacted, and vice versa. This suggests the importance of using both modalities when building machines for object perception tasks.

Compared to visual data, collecting large-scale audio recordings with rich object-level annotations is time-consuming and technically challenging for multiple reasons.

∗ indicates equal contributions.

(A) paper cup (B) ceramic cup (C) ceramic plate

Figure 1: Audio and visual data provide complementary information: visual cues tell us that A and B are cups and C is a plate, but only auditory cues inform us that A is made of a different material (paper) than B and C are (ceramic).

First, labeling objects at a finer granularity requires strong domain knowledge: various types of wood may have different physical properties, and therefore sound distinct; however, labeling wood type itself is already a highly nontrivial task. Further, some core object properties such as Young's modulus and density greatly affects the sound of an object, but it is often expensive and even intractable to obtain ground truth values. This could possibly explain why recent large-scale audio datasets, like the AudioSet [16], provide labels only on audio events, but not on the objects that generate the sounds.

Secondly, sound recorded in real life is generally a mixture of multiple sound sources and background noise. This poses an additional challenge in disentangling the sound each object makes. For example, the The Greatest Hits dataset [32] contains videos of objects hit by a drumstick. Despite being object-centric, the dataset contains many audio clips where the sound from the object is overwhelmed by that of the drumstick.

We introduce an alternative approach to overcome such difficulties — synthesizing audio-visual data for object perception — inspired by recent attempts in using synthetic visual data [36]. Synthesized data is automatically labeled,

easy to scale up and to increase its variance. In addition, synthesized sounds are naturally disentangled, as one could always generate the sound of each object independently. Our data synthesis framework is composed of three core generative models: a physics engine, a graphics engine, and an audio engine. The physics engine takes objects' shapes, material properties, and initial conditions as input, and then simulates their subsequent motions and collisions. The graphics engine renders videos based on the simulated object motion. The audio engine is built upon on a line of sound simulation works in computer graphics [19]. It combines pre-computed object mode shapes and object collisions for accurate audio synthesis. Our physics-based generative model contrasts with recent neural audio synthesis methods [32].

The core challenge for data synthesis is to achieve authenticity. To ensure our synthetic audio is realistic, we validate our synthesized audio by comparing it with real recordings obtained under experimental settings. Through spectral analysis and human studies, we demonstrate that our pipeline for audio synthesis produces realistic sounds.

With our generative model, we built a new synthetic dataset with audio-visual information. Our dataset, Sound-20K, consists of 20,378 videos with corresponding audio of objects interacting in a set of scenarios. We show, on both Sound-20K and real-world datasets, that visual and auditory information contribute in a complementary manner to object perception tasks including shape attribute and material recognition. We further demonstrate that knowledge learned on our synthetic dataset can be transferred for object perception on two real-world video datasets, Physics 101 [40] and The Greatest Hits [32].

Our contributions are three-fold: first, we propose to use synthesized audio-visual data for physical object perception, which provides unique advantages over purely gathering and using real recordings; second, we develop a fully generative, open-source audio-visual synthesis engine, and have generated a new dataset Sound-20K for studying object perception in various scenarios; third, we demonstrate that auditory and visual information can jointly contribute to infer geometric and physical properties of objects, and that the knowledge learned from our synthetic dataset is transferable to constrained real world scenes.

2. Related Work

Human Visual and Auditory Perception In the past decades, there have been extensive studies on how visual data enables human perception [20]. In the field of auditory perception, or psychoacoustics, researchers have also explored how humans could infer object properties including shape, material, size from audio [46, 22, 34, 21, 35]. Recently, Mcdermott *et al*. [27] proposed compact sound representations that capture semantic information and are informative of human auditory perception.

Physical Object Modeling Our work studies physical object perception, and therefore relates to research in modeling object shape and physics. There have been abundant works in computer vision to recover a 3D shape representation from visual input [37, 38, 5, 43]. For large-scale 3D shape modeling, the recently introduced ShapeNet [6] contains a large number of 3D CAD models for shape modeling, some with physical attributes. Fouhey *et al*. proposed the concept of 3D Shape Attributes [14], which, instead of modeling 3D shapes directly, characterized them with distinct attributes.

An important topic in understanding physical object properties is material recognition, which has been another long-standing research problem in computer vision [24, 25, 3]. Recently, Owens *et al*. [32] attempted to infer material properties from audio, focusing on the scenario of hitting objects with a drumstick. There have also been some recent works on understanding physical object properties like masses and frictions from visual input [42, 40, 7], and on modeling object or scene dynamics with explicit or learned physical laws [15, 28, 29, 45].

Synthesizing Data for Learning Most representation learning algorithms requires large amounts of data to achieve good performance, but for many tasks labeled data are scarce. Therefore, researchers have explored using synthetic visual data for tasks like viewpoint estimation and 3D reconstruction [36, 41]. Compared to these works, we explore synthesizing both visual and auditory data for learning.

Our sound synthesis pipeline builds upon several sound simulation works [31, 12, 19, 4, 39] in computer graphics. Early works [39] simulated object vibration using Finite Element Method and approximated the vibrating object as a single point source. For better synthesis quality, O'Brien *et al*. [12, 31] used the Rayleigh method to approximate wave equation solutions. James *et al*. [19] proposed to solve the Helmholtz equation using Boundary Element Method, where each object's vibration mode is approximated by a set of vibrating points. Bonneel *et al*. [4] proposed a frequency-domain sound synthesize method that could achieve near real-time performance. In contrast with the above work, we first construct an open-source pipeline that is able to synthesize audio-visual data at a large scale. Then, we propose to investigate how such synthetic date could help with physical object perception tasks.

Learning from Visual and Auditory Data Our framework enables generating large-scale audio-visual data for learning systems. Many multi-modal learning algorithms focused on learning jointly from visual and textual data, but there have been also some attempts in learning jointly from video and audio [30, 2]. Especially, Owens *et al*. [32, 33] explored using audio as supervision for sound synthesis and visual representation learning, and Aytar *et al*. [2] discussed how to jointly learn from audio and video for scene classification.

3. A Physical, Audio-Visual Generative Model

3.1. Overview

Our design of the generative model originates from the underlying mechanism of the physical world: when objects move in a scene, physical laws apply and physical events like collisions may occur. What we see is a video rendering of the scene with respect to object appearance, external lighting, *etc*., and what we hear is the vibrations of object shapes caused by physical events.

Our generative model therefore consists of a physics engine at its core, an associated graphics engine and an audio engine, as shown in Figure 2.

Physics Engine The physics engine serves as a core component for generating physical events. For audio and visual data, the important events are motions and collisions. Motion determines the position of objects at any given time, and collisions determine how vibrations in objects are excited. To this end, we used the open-source physical simulation engine Bullet [10], which feeds the physical events to graphical engine and audio engine. For accuracy purposes, the time step for physical simulation is set to 1/300 s. The outputs of the physics engine are motion information, which consists of an object's center position plus rotation, and collision information, which consists of collision position, direction and amplitude.

Graphics Engine We used Blender* and its Cycles ray tracer as our rendering tool. The rendering pipeline takes the motion information produced by the physics engine as input, *i.e*., object's center positions and its rotations at given time, then rigidly transform each object accordingly. The graphics engine is configured to render 30 frames per second for video generation.

3.2. Audio Engine

The audio engine, parallel to the graphics engine, takes object collision information as input and renders corresponding sound heard at the camera position. We generally followed the pipeline introduced by James *et al*. [19] to construct our audio engine. The audio engine first converts the object's collision information into its vibration using Finite Element Methods (FEM), then uses the vibration as a boundary condition to solve the wave propagation equation, which gives the air pressure, *i.e*. sound, at the camera position. We explain each step in detail below.

Collision to Vibration We adopt Finite Element Methods for converting collisions to object vibration. We first convert the surface mesh of an object into a volumetric tetrahedral mesh using isosurface stuffing [23], which represents the original shape with N tetrahedrons. Then, we apply FEM to solve for modal shapes of an object in the range of k audible frequencies. Specifically, the vibration equation can be written as

$$M\frac{\partial^2 \mathbf{u}}{\partial t^2} + K\mathbf{u} = \mathbf{f}, \tag{1}$$

where $\mathbf{u} \in \mathbb{R}^{3N}$ denotes the displacement of elemental nodes in 3D; $M, K \in \mathbb{R}^{3N\times 3N}$ denotes the mass and stiffness matrix of the system respectively, and $f \in \mathbb{R}^{3N}$ stands for the applied external force. We then find the modal matrix $\Phi \in \mathbb{R}^{3N\times k}$ by solving the generalized eigenvalue problem:

$$\Phi^T M\Phi = I, \qquad \Phi^T K\Phi = D, \tag{2}$$

where $D \in \mathbb{R}^{k\times k}$ is a diagonal matrix.

Modal shapes are then defined as the columns of matrix Φ. According to modal analysis procedures in FEM [18], $\mathbf{u}$ can be decomposed into a set of modal coefficients $\mathbf{c} \in \mathbb{R}^k$ under modal basis Φ, *i.e*.:

$$\Phi\mathbf{c} = \mathbf{u}. \tag{3}$$

The induced vibration is then a linear combination of modal shapes with modal coefficients $\mathbf{c}$. To see this, substitute Equation 1 into Equation 3, which gives

$$\frac{\partial^2 \mathbf{c}}{\partial t^2} + K\Phi\mathbf{c} = \mathbf{f} \Rightarrow \frac{\partial^2 \mathbf{c}}{\partial t^2} + D\mathbf{c} = \Phi^T\mathbf{f}. \tag{4}$$

Note that since D is diagonal, the system can be decoupled into independent sinusoid solutions, and the actual vibration is given by $\Phi\mathbf{c}$.

Solving the Wave Equation Then, given an object's vibration, we need to solve for the actual air pressure profile at the camera position. The underlying equation to describe this process is the wave equation,

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right) p(\mathbf{x}, t) = 0, \tag{5}$$

where $p(\mathbf{x}, t)$ denotes the air pressure at time t and position $\mathbf{x}$. v denotes the speed of sound in air. Suppose $p_i(x, t)$ is the solution to the Neumann boundary condition composed by object vibration:

$$\frac{\partial}{\partial \mathbf{n}} p_i(\mathbf{x}, 0) = \Phi_i, \quad \mathbf{x} \in S, \tag{6}$$

where Φ_i is the i-th column of Φ, S is the object's surface, and $\mathbf{n}$ is the surface normal. The solution to Equation 6 is simply $\sum_i c_i p_i(\mathbf{x}, t)$. Since each modal shape has its natural frequency ω_i, $p_i(\mathbf{x}, t)$ can be written as $q_i(\mathbf{x})e^{-j\omega_i t}$,where $j = \sqrt{-1}$. Then, the wave equation can be turned into the Helmholtz equation for each modal shape,

$$\left(\nabla^2 + k^2\right) q_i(\mathbf{x}) = 0, \text{s.t.} \quad \frac{\partial}{\partial n} q_i(\mathbf{x}) = \Phi_i \;\; \mathbf{x} \in S. \tag{7}$$

Our audio engine solves the above Helmholtz equation using Boundary Element Method (BEM) [8]. We built two solvers for Equation 7: a direct integration solver using open-source library NiHu [13] and an iterative solver with Fast Multipole Method [26] acceleration using the FMM3D library [44]. A more detailed description can be found in the

*https://www.blender.org

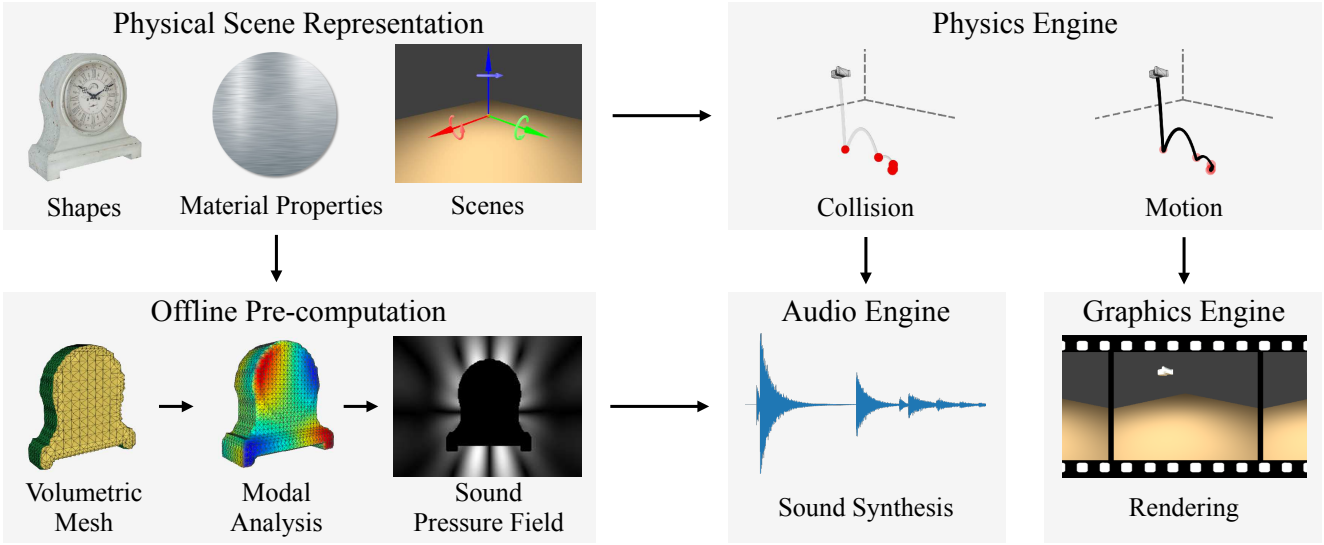


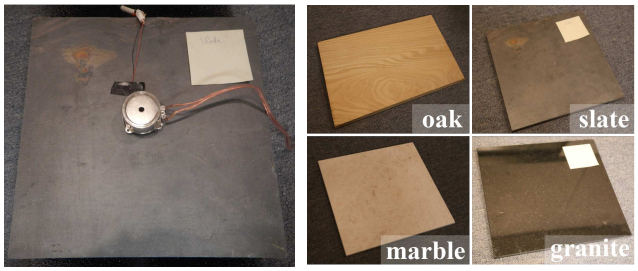
Figure 2: Our generative model for audio-visual scenes. Given object shapes, material properties, and a scene configuration, a physics engine simulates both object collisions and motion. An audio engine then takes the collision data and pre-computed object mode shapes for sound synthesis. There is also a graphics engine that renders accompanying video based on object motion.

supplementary material.

Offline-Online Decomposition A straightforward implementation of the above framework would be inefficient. James *et al*. [19] proposed to accelerate it by decomposing the audio engine into an on-line phase and an off-line phase. The offline part computes the modal shapes of an object and their corresponding air pressure on its surface, using FMM-accelerated BEM. Then, the surface air pressure for each mode is approximated by the pressure generated by a set of vibrating points inside the object, whose location and vibration amplitude are pre-computed and stored. The points' position and amplitude are computed so that the pressure they produce approximates the original surface pressure in least square sense. The on-line step first decomposes the collision information into modal coefficients. Then, for each mode, we simply calculate the sound pressure field generated by the pre-computed points sets, which is far more efficient than solving the Helmholtz equation. Finally, we linearly combine the sound pressure to generate the audio. For more details, please refer to the original paper [19].

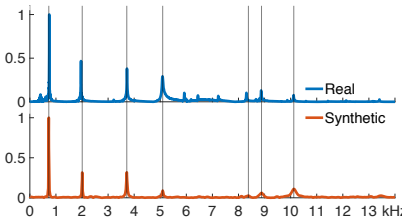
4. Audio Synthesis Validation

We validated the accuracy of our audio synthesis by comparing it with real world recordings. We recorded the sounds made by striking four plates of different materials (granite, slate, oak and marble) as shown in Figure 3b. The audio was measured by exciting the center of the plates with a contact speaker and measuring the resulting vibrations with a piezo-electric contact microphone placed adjacent to the speaker (shown in Figure 3a). All measurements were made in a sound-proof booth to minimize background noise in the recording.



(a) The recording setup

(b) The four tiles we used



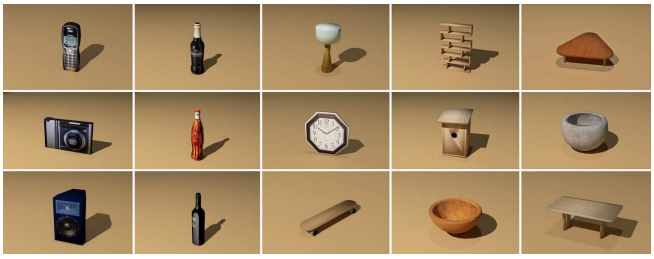
(c) Real *vs*. synthetic spectra of granite

Tiles	Labeled *real* (%)
Oak	45.3
Slate	50.5
Marble	52.6
Granite	46.3

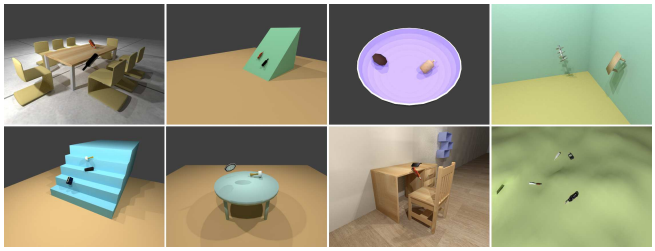
(d) Behaviorial results

Figure 3: We validate our audio synthesis pipeline through carefully-designed physics experiments. We record the sound of four tiles of different materials (a-b), and compare its spectrum with our synthesized audio (c) with corresponding physical properties. We also conducted behavioral studies, asking humans which of the two sounds match the image better. We show results in (d).

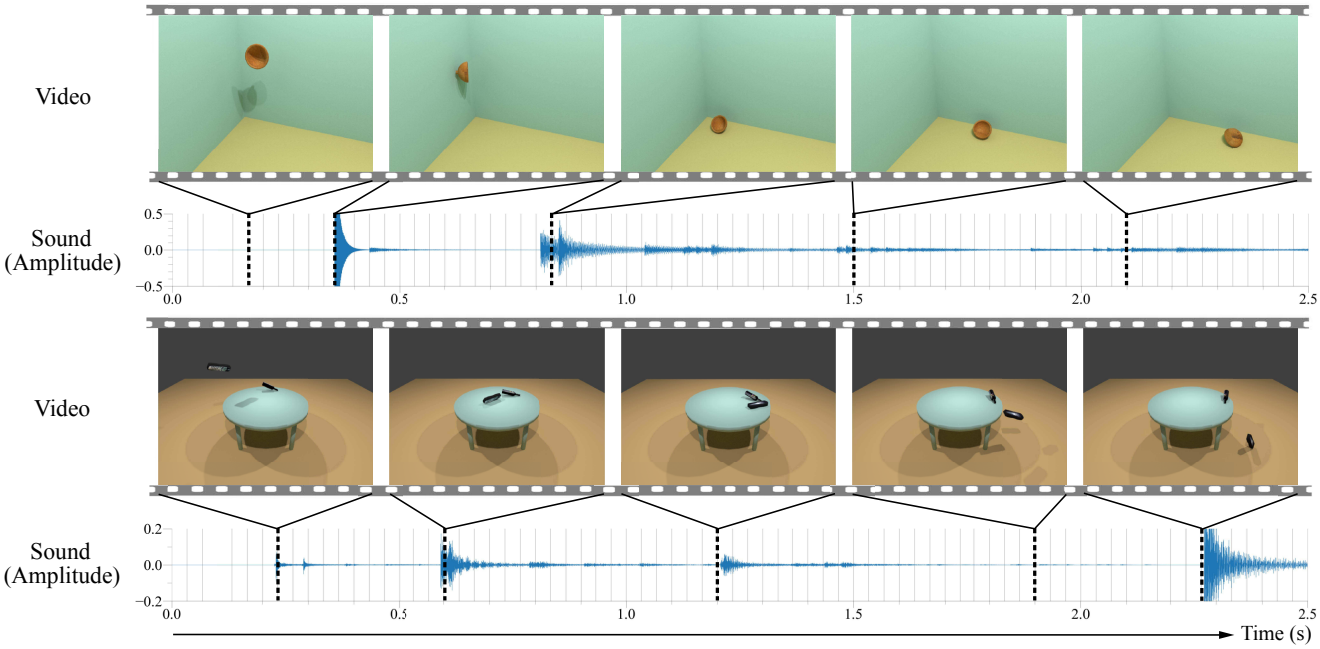
To generate synthetic sounds for the four plates, we used FEM on object meshes, matching the shape and material properties of the plates. We obtained ranges of material properties (Young's modulus, density, Poisson's ratio and damping coefficients) from engineering tables and performed grid search to select optimal values. The most similar synthetic sound was selected by comparing the power spectra of the synthetic sound and the measured impact sound of the



(a) Sample shapes in our Sound-20K dataset



(b) Sample scenarios in our Sound-20K dataset



(c) Sample videos in our Sound-20K dataset

Figure 4: An overview of our Sound-20K dataset. We show objects (a), scenarios (b), and sample audio-visual data (c) in our dataset.

corresponding plate. Under synthetic settings, an impulse force is applied at the center of plate's surface. The microphone position is set close to the center of the tile, in order to match the measurement configuration.

We validated the accuracy of our synthetic sounds by comparing the spectrum of synthetic audio with real recordings. Figure 3c shows the spectrum comparison between the synthetic sound and the real recording of the granite tile. We also designed a human perceptual study in which 95 people were asked to judge whether the recording or the synthetic was more realistic. Table 3d shows the percentage of people who labeled synthetic sounds as real.

5. The Sound-20K Dataset

We built a new dataset, named Sound-20K, consisting of 20,378 audio-video pairs and other related data required to generate Sound-20K.

Shapes We carefully selected 39 3D shapes from ShapeNet [6]. They are all watertight, manifold meshes with consistent outward facing normals. Objects are classified into six super categories and 21 categories. All shapes in the dataset have been normalized with unit diagonal length and been repositioned to their geometric center.

Recently, Fouhey *et al*. [14] introduced the concepts of 3D shape attributes, where they defined 12 attributes for 3D shapes. We decide to also annotate the 3D attributes of the objects in our dataset. Three of the attributes are specifically for sculptures (*e.g*., multiple pieces), and thus we chose to exclude them. The attributes we use include *has planarity, no planarity, has cylindrical, has roughness, mainly empty, has hole, thin structures, mirror symmetry*, and *cubic aspect ratio*. We refer readers to [14] for more details.

We also label objects' size as small, medium, or large, which are taken into consideration for dataset generation to ensure realism. Third, we specify possible materials associated with the shape. The most common material is labeled and the corresponding texture is applied. Sample shapes are shown in Figure 4a.

Materials Seven frequently used materials and their physical property parameters for sound generation and physical simulation are provided. Seven materials include ceramic,

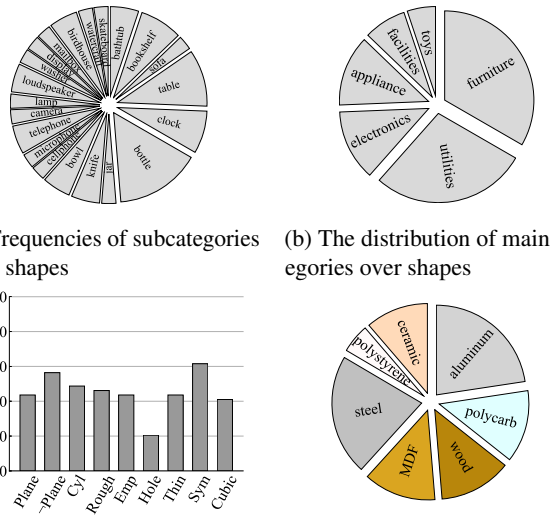
polystyrene, steel, medium-density fiberboard (MDF), wood, polycarbonate and aluminum. Each material has its mechanical properties and acoustic property: Young's modulus (E), Poisson ratio (ν), density (ρ), friction coefficient (μ), coefficient of restitution (e) and Rayleigh damping coefficients (α, β) [1]. ρ, μ, e of each object are fed into the physics engine for rigid body simulation; $E, \nu, \rho, \alpha, \beta$ are used for pre-computation and online audio synthesis.

Scenarios We have created 22 scenarios with different levels of complexity. Each scenario is independent of shape or material designations. These scenarios are derived from 9 basic frameworks in three levels of complexity. As shown in Figure 4b, preliminary ones contain a few simple geometries, including flat ground, a ramp, staircases, a corner between walls and hemispherical shells. More advanced scenes have a table on the ground and an uneven floor. Furthermore, a study room and a dining table are also set up to meet high level of conceptual richness. For each framework, we appointed multiple initialization parameters with different number of objects, linear velocities and angular velocities, which control the potential interactions among objects in the same scene during simulations.

Dataset We synthesize 20,378 videos with audios via sampling from numerous possible combinations of shape-material pairs in a customized manner. Explicitly, we impose two constraints as follows. Only shapes of similar actual dimensions are allowed to jointly appear in a certain scenario; and additionally, shapes are chosen only if their relative sizes with respect to the scene setup are realistic. For example, a chair and a table, which are considered to be large in size, are reasonable to drop onto the ground together, but falling down on a dining table would be unrealistic. Given these restrictions, we are able to automatically generate scene entries in batches as the input to our generative model. Sample videos are shown in Figure 4c. We include more videos in the supplementary material.

5.1. Dataset Analysis

Figure 5a shows that although the original ShapeNet dataset has a strong skewness towards chairs, bottles and some other particular categories, our selection tries to even this distribution across most of the subcategories. After we concatenated some subcategories with similar functionality into a main category, the distribution is demonstrated in Figure 5b. As illustrated in Figure 5c, the distribution of objects over all attributes is relatively even, which indicates a wide coverage of shape characteristics. We plot the statistics for the number of objects assigned to each material in Figure 5d. The prior distribution across materials does not differ significantly such that no material is marginalized due to its minimal appearance in the dataset.



(a) Frequencies of subcategories over shapes

(b) The distribution of main categories over shapes

(c) The distribution of shape attributes over shapes

(d) The distribution of materials over shapes

Figure 5: Statistics of Sound-20K. Objects in the dataset distribute across a diverse set of categories (a) and super-categories (b). They also have an even distribution of shape attributes (c) and span across a set of materials (d).

6. Object Perception with Audio-Visual Data

Our model generates data in two modalities: audio and video. In this section, we verify that both auditory and visual data contribute to physical object perception, but in complementary ways. We look into two tasks of object perception: material recognition and 3D shape attribute recognition. We observe that auditory data contain richer information of object material, while visual data help to recognize 3D shape attributes better.

6.1. Data

In addition to our synthetic Sound-20K, we consider two real-world video datasets with audio.

Physics 101 [40] is a dataset containing 15,190 videos of 101 objects behaving in five scenarios. The scenarios include objects falling onto a surface of a certain material, objects splashing into water, *etc*. Each object is annotated with its ground truth material, mass, and volume.

The Greatest Hits [32] is a dataset containing 977 videos of a person probing environments with a drumstick. Each video, on average, contains 48 actions (hits and scratches), and there are 46,577 actions in total. The dataset also contains the material label of the target object in each video.

6.2. Methods

We use standard convolutional networks for our study. As we are learning jointly from audio and video, our network has an audio stream and a visual stream. We show the

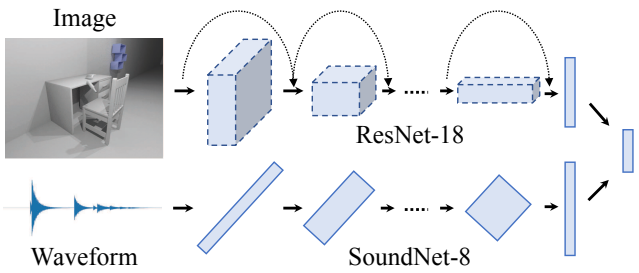


Figure 6: We used a two-stream convolutional network, where the structure of the visual stream is an 18-layer ResNet [17], and that of the auditory stream is an 8-layer SoundNet [2].

network structure in Figure 6.

Our visual stream employs a ResNet-18 [17], taking an image as input and producing a 512-dim vector. Our audio stream uses the same structure as SoundNet-8 [2], taking a raw audio sequence as input and producing a 1,024-dim vector. For multi-modal learning, we simply concatenate the two vectors, leading to a latent vector of dimension 1,536.

The final output of the networks and the loss function depend on the task. For material prediction, it would be n k-dim vectors with cross entropy loss, where n is the number of objects in the scene and k is the number of material types in the dataset. For shape attribute recognition, it would be n 9-dim vectors with binary cross entropy loss, as we consider 9 shape attributes in our experiments.

Training Details We used stochastic gradient descent for training, with a learning rate of 0.001, a momentum of 0.9, and a batch size of 16. We implemented our framework in Torch7 [9]. We trained all models from scratch, without any pre-training.

6.3. Material Recognition

We first start with the task of material recognition. We conduct experiments on Sound-20K and Physics 101, as they all have material labels, with visual and/or auditory data as input. For Sound-20K, we choose to use texture-less videos to exclude the correlations between object appearance and its material. For both datasets, we used 95% of the videos for training, and 5% for testing.

Figure 7 shows the confusion matrices of material classification on the Physics 101 dataset, where we observe that auditory data also contain rich and complementary information for inferring object material, compared to visual data. The results on Sound-20K are expected, because the visual appearance of texture-less objects contains little information about its material. The results on Physics 101 verified that these effects exist in real-world videos. Specifically, Figures 7b and 7c show that objects which are hard to discriminate from visual data (a small metal coin *vs*. a piece of wooden block), are identified by the model from the distinct sounds they make.

Dataset	Chance	Auditory	Visual	A + V
Sound-20K	14.3	65.6	46.2	68.5
Physics 101	6.6	78.1	94.5	99.4

(a) Material classification on Sound-20K and Physics 101 [40].

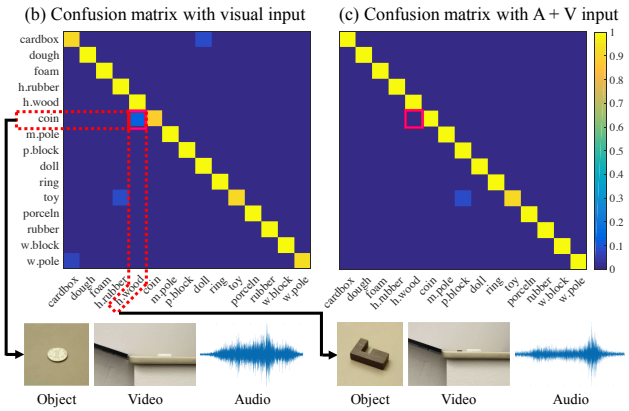


Figure 7: Auditory and visual data are complementary on material recognition. We show classification accuracies in (a), where learning jointly from two modalities outperforms learning from either one. In (b) and (c), we show confusion matrices on Physics 101 of models using either visual or audio-visual data. Though visual data is very informative, it makes mistakes for objects that are hardly visible (coin *vs*. block). Using auditory data resolves the ambiguity.

6.4. Shape Attribute Recognition

Inferring full 3D shape from purely auditory data would be challenging, if not intractable. However, it would be interesting to investigate what information about 3D shape we may recover from audio. As discussed in Section 5, we labeled all objects in Sound-20K and Physics 101 with nine attributes. We therefore would like to study what attributes we may infer from audio-visual data. Our experiment setup is the same as that in Section 6.3.

Table 1 shows results on 3D shape attribute recognition. As expected, visual data contains rich information of most shape attributes. At the same time, it is intriguing to find that auditory data is also informative of many shape attributes.

7. Transferring from Synthetic to Real Data

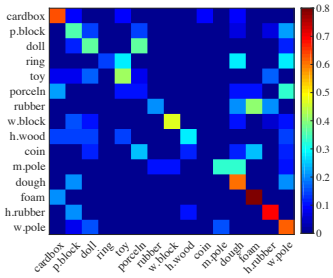
We now demonstrate how knowledge learned on our synthetic audio-visual dataset may transfer to real-world data. As discussed earlier, auditory and visual data contain complementary information of physical objects; we therefore study two tasks: inferring object materials from auditory data, and inferring shape attributes from audio-visual data.

7.1. Material Recognition from Auditory Data

Both Physics 101 and The Greatest Hits have object material labels. However, their label sets are different from the label set we used in Sound-20K. Therefore, we chose to

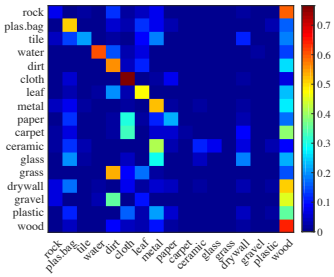
Dataset	Input	Curvature				Occupancy					Avg
		Plane	¬Plane	Cyl	Rough	Emp	Hole	Thin	Sym	Cubic	
Sound-20K	Auditory	78.1	77.9	69.9	72.2	78.8	86.9	77.8	69.7	73.2	76.1
	Visual	86.5	86.3	78.1	71.4	78.7	87.1	75.3	72.2	78.9	79.4
	A + V	86.9	86.9	81.1	80.6	84.0	90.8	84.7	79.0	81.4	83.9
Physics 101	Auditory	85.2	85.9	83.1	88.0	76.1	97.2	92.3	88.7	82.4	86.5
	Visual	97.2	96.5	95.1	95.8	94.4	98.6	97.2	97.2	97.9	96.6
	A + V	97.2	97.2	97.2	97.9	97.2	97.9	96.5	96.5	95.8	97.0

Table 1: Auditory and visual data are complementary on shape attribute recognition. On both Sound-20K and Physics 101 [40], learning jointly from audio-visual data achieves the highest performance. Though for this task in specific, visual data plays the major role, we still observe that audio also contains rich information of object shape attributes.



Methods	Acc. (%)
Real data	58.5
Ours (pool5)	29.6
Ours (conv6)	35.9
Ours (conv7)	41.5
MFCC [11]	40.7
Chance	6.7

Figure 8: Material classification on Physics 101 [40] with various features learned on the synthetic Sound-20K and a linear SVM. Left: confusion matrix with our conv7 features. Right: accuracies.



Methods	Acc. (%)
Real data	44.2
Ours (pool5)	33.7
Ours (conv6)	37.9
Ours (conv7)	39.2
MFCC [11]	38.1
Chance	5.6

Figure 9: Material classification on The Greatest Hits [32] with various features learned on the synthetic Sound-20K and a linear SVM. Left: confusion matrix with our conv7 features. Right: classification accuracies.

evaluate how well the features learned on Sound-20K can transfer to these real datasets. In specific, we extract features from middle layers of the network, and apply a linear SVM on these features for material classification.

In Figure 8 and Figure 9 we show results on material classification, where we compare features from various layers of the audio stream of our network, trained on the synthetic Sound-20K, with other features that are either hand-designed, or learned directly on these real datasets. We also show the confusion matrices of our conv7 features. We see that our features, learned on synthetic data, achieve comparable results with the classic MFCC features [11].

7.2. Shape Attribute Recognition

We also study, on Physics 101, how well the model trained on Sound-20K can perform on shape attribute recognition.

Input	Plane	¬Plane	Cyl	Rough	Emp	Hole	Thin	Sym	Cubic	Avg
A	48.1	50.0	53.9	86.5	78.9	83.7	31.7	41.8	79.8	61.6
V	70.2	69.2	58.7	43.3	38.5	61.5	42.3	71.2	63.5	57.6

Table 2: Shape attribute recognition on the object falling scenario in Physics 101, using models trained on a subset of audios and object silhouettes from Sound-20K

For this task, we use 785 videos in Sound-20K with one object in the scene for training; we use one scenario from Physics 101— object falling — for evaluation. This is because these scenes share similar visual layout. As videos in these two datasets have very different appearances, we use object silhouettes instead of RGB images as visual input. We obtain the silhouettes using background subtraction. For this task, we directly evaluate the learned model on real data, without any fine-tuning. As shown in Table 2, representations learned on synthetic data can directly be transferred to real-world input, and audio and visual data perform well on different sets of attributes.

8. Conclusion

In this paper, we proposed using synthetic audio-visual data for physical object perception. Using synthetic data has unique advantages: they are relatively easy to collect and expand, and they are fully annotated. Secondly, we constructed a physics-based open-source pipeline that synthesizes authentic audio-visual data at a large scale. Such a pipeline provides easy access and flexibility for future researchers to investigate how auditory and visual information could help in various perception tasks.

9. Acknowledgement

The authors would like to thank Changxi Zheng, Eitan Grinspun, and Maddie Cusimano for helpful discussions. This work is supported by Toyota Research Institute, Samsung Research, and the Center for Brain, Minds and Machines (NSF STC award CCF-1231216).

References

[1] S. Adhikari. *Damping models for structural vibration*. PhD thesis, Cambridge University, 2000. 6

[2] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016. 2, 7

[3] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Material recognition in the wild with the materials in context database. In *CVPR*, 2015. 2

[4] N. Bonneel, G. Drettakis, N. Tsingos, I. Viaud-Delmon, and D. James. Fast modal sounds with scalable frequency-domain synthesis. *ACM TOG*, 27(3):24, 2008. 2

[5] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, 2000. 2

[6] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 5

[7] M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum. A compositional object-based approach to learning physical dynamics. In *ICLR*, 2017. 2

[8] R. D. Ciskowski and C. A. Brebbia. *Boundary element methods in acoustics*. Springer, 1991. 3

[9] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011. 7

[10] E. Coumans. Bullet physics engine. *Open Source Software: http://bulletphysics. org*, 1, 2010. 3

[11] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE TASSP*, 28(4):357–366, 1980. 8

[12] J. F. Director-O'Brien. Synthesizing sounds from physically based motion. In *ACM SIGGRAPH 2001 video review on Animation theater program*, page 59. ACM, 2001. 2

[13] P. Fiala and P. Rucz. Nihu: An open source c++ bem library. *Advances in Engineering Software*, 75:101–112, 2014. 3

[14] D. F. Fouhey, A. Gupta, and A. Zisserman. 3d shape attributes. In *CVPR*, 2016. 2, 5

[15] K. Fragkiadaki, P. Agrawal, S. Levine, and J. Malik. Learning visual predictive models of physics for playing billiards. In *ICLR*, 2016. 2

[16] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dartaset for audio events. In *ICASSP*, 2017. 1

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[18] T. J. Hughes. *The finite element method: linear static and dynamic finite element analysis*. Courier Corporation, 2012. 3

[19] D. L. James, J. Barbič, and D. K. Pai. Precomputed acoustic transfer: output-sensitive, accurate sound generation for geometrically complex vibration sources. *ACM TOG*, 25(3):987–995, 2006. 2, 3, 4

[20] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2):201–211, 1973. 2

[21] R. L. Klatzky, D. K. Pai, and E. P. Krotkov. Perception of material from contact sounds. *Presence: Teleoperators and Virtual Environments*, 9(4):399–410, 2000. 2

[22] A. J. Kunkler-Peck and M. Turvey. Hearing shape. *Journal of Experimental psychology: human perception and performance*, 26(1):279, 2000. 1, 2

[23] F. Labelle and J. R. Shewchuk. Isosurface stuffing: fast tetrahedral meshes with good dihedral angles. *ACM TOG*, 26(3):57, 2007. 3

[24] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001. 2

[25] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz. Exploring features in a bayesian framework for material recognition. In *CVPR*, 2010. 2

[26] Y. Liu. *Fast multipole boundary element method: theory and applications in engineering*. Cambridge university press, 2009. 3

[27] J. H. McDermott, M. Schemitsch, and E. P. Simoncelli. Summary statistics in auditory perception. *Nature neuroscience*, 16(4):493–498, 2013. 2

[28] R. Mottaghi, H. Bagherinezhad, M. Rastegari, and A. Farhadi. Newtonian scene understanding: Unfolding the dynamics of objects in static images. In *CVPR*, 2016. 2

[29] R. Mottaghi, M. Rastegari, A. Gupta, and A. Farhadi. what happens if... learning to predict the effect of forces in images. In *ECCV*, 2016. 2

[30] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011. 2

[31] J. F. O'Brien, C. Shen, and C. M. Gatchalian. Synthesizing sounds from rigid-body simulations. In *SCA*, 2002. 2

[32] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *CVPR*, 2016. 1, 2, 6, 8

[33] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016. 2

[34] D. Rocchesso and F. Fontana. *The sounding object*. Mondo estremo, 2003. 2

[35] M. Siegel, R. Magid, J. B. Tenenbaum, and L. Schulz. Black boxes: Hypothesis testing via indirect perceptual evidence. In *CogSci*, 2014. 1, 2

[36] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *ICCV*, 2015. 1, 2

[37] R. Szeliski and S. B. Kang. Recovering 3d shape and motion from image streams using non-linear least squares. In *CVPR*, 1993. 2

[38] D. Terzopoulos, A. Witkin, and M. Kass. Constraints on deformable models: Recovering 3d shape and nonrigid motion. *Artificial intelligence*, 36(1):91–123, 1988. 2

[39] K. Van den Doel and D. K. Pai. The sounds of physical shapes. *Presence: Teleoperators and Virtual Environments*, 7(4):382–395, 1998. 2

[40] J. Wu, J. J. Lim, H. Zhang, J. B. Tenenbaum, and W. T. Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, 2016. 2, 6, 7, 8

[41] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. In *ECCV*, 2016. 2

[42] J. Wu, I. Yildirim, J. J. Lim, W. T. Freeman, and J. B. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *NIPS*, 2015. 2

[43] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*, 2016. 2

[44] L. G. Z. Gimbutas. Fmmlib3d. *Fortran libraries for fast multiple method in three dimensions, http://www.cims.nyu.edu/cmcl/fmm3dlib/fmm3dlib.html.*, 1, 2011. 3

[45] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, and S.-C. Zhu. Beyond point clouds: Scene understanding by reasoning geometry and physics. In *CVPR*, pages 3127–3134, 2013. 2

[46] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media, 2013. 1, 2