GLAVNet: Global-Local Audio-Visual Cues for Fine-Grained Material Recognition

Fengmin Shi[1], Jie Guo[1,†], Haonan Zhang[1], Shan Yang[1], Xiying Wang[2], Yanwen Guo[1,†]
[1]State Key Lab for Novel Software Technology, Nanjing University [2]IQIYI Intelligence

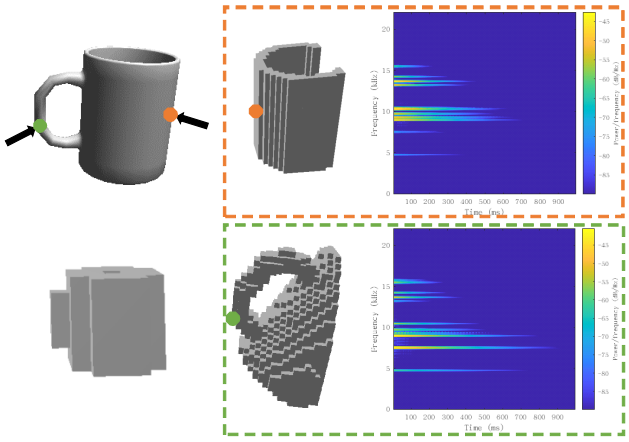
Abstract

In this paper, we aim to recognize materials with combined use of auditory and visual perception. To this end, we construct a new dataset named GLAudio *that consists of both the geometry of the object being struck and the sound captured from either modal sound synthesis (for virtual objects) or real measurements (for real objects). Besides global geometries, our dataset also takes local geometries around different hitpoints into consideration. This local information is less explored in existing datasets. We demonstrate that local geometry has a greater impact on the sound than the global geometry and offers more cues in material recognition. To extract features from different modalities and perform proper fusion, we propose a new deep neural network* GLAVNet *that comprises multiple branches and a well-designed fusion module. Once trained on GLAudio, our GLAVNet provides state-of-the-art performance on material identification and supports fine-grained material categorization.*

1. Introduction

A fundamental problem in computer vision is that of identifying material categories, *e.g.*, metal, glass or wood, from RGB images. In many applications, such as 3D scene understanding [18] and robot control [17], materials of objects provide useful hints for substantially improving the performance [16].

As an active area of research, there have been quite a few works on material recognition, ranging from traditional solutions relying on hand-designed image features [29, 23, 40] to convolutional neural networks (CNNs) [55, 53] combined with large-scale datasets, *e.g.*, Flickr Material Database (FMD) [44] and Materials in Context Database (MINC) [7]. Since these methods and most others in literature use solely the visual modality, rich surface textures or geometrical variations are required to make the recognition process robust.



(a) Global geometry (b) Local geometry (c) Spectrogram

Figure 1. Impact of the hitpoint on the generation of sound. Despite the same global geometry (a), different local geometries around the hitpoints (b) may lead to distinct auditory features (c).

However, there are many scenarios in which strong visual ambiguities exist, especially for smooth and textureless surfaces, as a given material can take on many different appearances depending on the viewpoint, illumination and shape [4]. Worse still, the visual appearance may be significantly altered by painting or coating (*e.g.*, a wooden sculpture covered with gold lacquer), making the visual perception unreliable. One way of tackling this problem is to involve additional sensory modalities to complement the visual perception. Recent research confirms that sound augments visual inputs and is an important modality for identifying materials [2, 36, 57, 48].

In this paper, we aim to identify materials of rigid objects that appear in our daily life. To this end, we construct a new dataset, called GLAudio, containing both auditory and visual perception of each object. In this dataset, every object (virtual or real) is encoded in a voxelized representation while every impact sound is generated either by modal sound synthesis or by striking a real object. Compared with some existing sound datasets such as Sound-20K [57] and RSAudio [48], the proposed GLAudio dataset offers two

† Corresponding Authors.

main improvements. Firstly, we incorporate the information of hitpoint producing the sound into our dataset. The hitpoint is encoded in the form of a local geometry around it and produces discriminative features for the sound, as demonstrated in Fig. 1. In some situations, this local geometry is more important than the global geometry of the whole object in generating the sound, *e.g.*, in knocking the handhold of a cup. Secondly, to our knowledge our dataset is the first one that contains fine-grained materials. For instance, our dataset has 10 different metal classes including aluminum, iron, gold, etc. This opens up new opportunities for solving more complex problems like testing gold for discerning the real from the fake one.

Based on our GLAudio dataset, we design a deep neural network, *i.e.*, GLAVNet, to infer material information from both visual and auditory cues. We show that the fusion of global and local shapes of the object outperforms previous methods which rely solely on the use of global information. We also demonstrate that GLAVNet enables fine-grained material recognition which is extremely challenging for these previous works.

In summary, we make the following contributions:

- We introduce the first audio-visual dataset, GLAudio, that contains both global geometry of a 3D object and local geometry around the hitpoint.
- We propose a new deep neural network, GLAVNet, that supports fine-grained material recognition with the help of GLAudio.
- Once GLAVNet is trained, our method achieves state-of-the-art performance on material recognition and enables fast inference.

2. Related Work

2.1. Material Datasets

Large-scale datasets combined with deep neural networks prevail in tackling many vision tasks. For materials, an early dataset that has been widely used in both computer vision and graphics is the CUReT dataset [13], which only consists of 61 material samples captured in a lab under 205 different lighting and viewing directions. Later, Hayman *et al.* [21] released KTH-TIPS by imaging 10 categories from the CUReT dataset at different scales. Caputo *et al.* [8] released KTH-TIPS2 that adds images from 4 physical samples per category. To push material recognition into the real-world, Sharan *et al.* [44] created the Flickr Materials Database (FMD) containing 10 material categories carefully selected from Flickr photos. The OpenSurfaces dataset [6] introduces 105,000 material segmentations from real-world consumer photographs, which is significantly larger than FMD. A more diverse material dataset, *i.e.*, MINC [7], is recently contributed by Bell *et al.* which has 3 million material samples. To address the inherent limitations of these visual data, alternate modalities, such as sound [2, 36, 57, 48] and haptic features [16], have been introduced and collected from different sensors. However, compared with the above visual datasets, auditory or haptic datasets are much smaller in size due to the difficulty in collection. In this paper, we release a new multimodal dataset combining both visual and auditory cues for material recognition.

2.2. Material Recognition

In vision and graphics, material recognition is a fundamental problem that has been studied from several perspectives. One line of work seeks to estimate complete BRDF or BTF [13] from real-world measures. As this topic is out of the scope of this paper, we refer readers to a recent survey with a focus on deep-learning-based methods [15]. Another series of work, which is closely related to ours, aims to categorize materials using hand-designed or learned features. Early methods usually leverage filter bank responses to extract features from rich textures of materials [49, 28, 50]. Subsequently, Varma and Zisserman [51] adopted more direct clustering and statistics of intensities of small pixel neighborhoods to classify textures, achieving better performance than filter responses. Since then, there are many follow-up work [52, 10, 45, 47, 32]. Recent studies [11, 12, 55, 53, 31] show that features learned from CNNs outperforms traditional methods for texture classification. Liu *et al.* [29] suggested using a variety of color, texture, gradient and curvature features for classifying general materials.

Although image plays important roles in material recognition, it has some intrinsic limitations due to the diversity in material appearances. There are many scenarios in which it is very challenging to visually identify an object's material. Recently, there have been some attempts in learning jointly from visual and auditory data. Arnab *et al.* [2] estimate dense object and material labels with both dense visual cues and sparse auditory cues. Liu *et al.* [30] investigated the fusion of sound and acceleration measurements in surface material categorization. Sterling *et al.* [48] used spectrograms of impact sounds and voxelized shape estimates for improving the classification and reconstruction of 3D objects.

2.3. Sound for Scene Understanding

Besides material recognition, sound is increasingly becoming an important modality for other tasks in scene understanding. Owens *et al.* [37] demonstrated that ambient sounds contain significant information about objects and scenes and can be used as a supervisory signal for learning visual models. Aytar *et al.* [3] proposed SoundNet to jointly

Table 1. Statistical comparison of GLAudio against other datasets. Here, #Mat., #Geo. and #Snd. represent the numbers of materials, geometries and sounds, respectively. ?H., ?R. and ?S. indicate whether the dataset contains hitpoints, real sounds or synthesized sounds.

Dataset	#Mat.	#Geo.	#Snd.	?H.	?R.	?S.
GreatestHits	17	—	46,577	×	√	×
Sound-20K	7	39	20,378	×	×	√
RSAudio	11	59	63,583	×	√	√
GLAudio	17	41	40,299	√	√	√

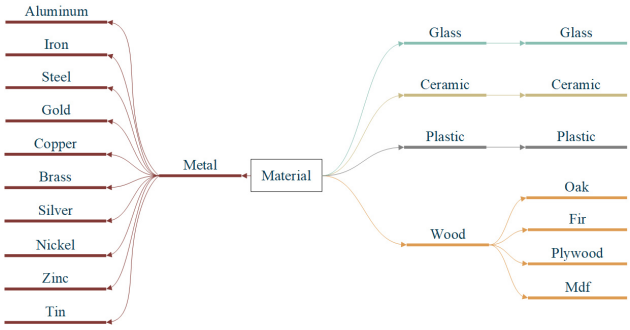


Figure 2. Two-scale material categories used in GLAudio.

learn from audio and video for scene classification. Arandjelovic and Zisserman [1] trained vision and sound models jointly to perform fine-grained recognition tasks. The use of audio data has also been beneficial in other contexts of scene understanding such as environment classification [43, 20, 39, 42], object tracking [5, 19] and reconstruction [56].

3. GLAudio Dataset

In this section, we describe the construction of our GLAudio dataset and analyze its statistics. This dataset has the following characteristics that make it different from others. Specifically, each sample in the dataset is equipped with

- a clip of audio with 44,100Hz sample rate and 3 seconds length,
- a global geometry of the object and a local geometry around the hitpoint generating the sound,
- a set of material parameters and a multi-scale material category,

Currently, our GLAudio dataset consists of 40,256 synthesized samples produced by 629 3D objects. These objects differ either in geometry (37 items) or in material (17 items). It also contains 43 real examples captured from 4 real objects. Table 1 lists statistical comparison of the proposed GLAudio dataset against others containing auditory data.

3.1. Shapes and Materials

The synthesized sounds are captured from 37 geometry shapes selected from public available ShapeNet [9] and TurboSquid[‡] models. All models are voxelized with a fixed spatial resolution of $32 \times 32 \times 32$, representing the global geometries of the objects. We also selected 17 commonly used materials from online material property tables. These materials are firstly divided into five basic categories: *metal, wood, glass, plastic and ceramics*. Then, *metal* and *wood* are further divided into several fine-grained categories, as shown in Fig. 2. These materials are distinguished by a set of parameters including Young's modulus E, Poisson ratio ν, density ρ and loss factor η. The distributions of these parameters in our dataset are plotted in Fig. 3. From these diagrams we see that fine-grained material categories can be distinguished by some basic material parameters. For instance, different metals have a wide range of Young's modulus/density, resulting in different auditory features. We use these parameters to generate impact sound based on *modal sound synthesis*.

For the real examples, we model the global geometries of 4 real objects using Microsoft Kinect V2, and then capture sounds by hitting different positions on the objects, resulting in 43 audio clips in total. These objects have 4 different material categories (glass, ceramics, wood and metal). More details about the construction of these real examples are provided in the supplemental document.

3.2. Modal Sound Synthesis

Modal sound synthesis has been widely used for physically-based audio content generation [14, 35, 41, 25]. Firstly, surface mesh of an object is converted into volumetric data of n cells by tetrahedralization. Here, we generate these tetrahedral cells using TetGen [46]. Secondly, mass matrix $\mathbf{M}$ and stiffness matrix $\mathbf{K}$ are calculated using finite element methods from volumetric mesh and material's Young's modulus, Poisson ratio and density. Assuming that $\mathbf{C}$ is the damping matrix, the displacement $\mathbf{x} \in \mathbb{R}^{3n}$ of nodes in volumetric mesh under external forces $\mathbf{f} \in \mathbb{R}^{3n}$ fulfills the following differential equation:

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{C}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f}. \tag{1}$$

Here, the Rayleigh damping model is applied to our modal analysis. The model damping matrix $\mathbf{C}$ is given by $\mathbf{C} = \alpha\mathbf{M} + \beta\mathbf{K}$, in which α and β are constant Rayleigh damping coefficients that can be calculated from loss factor η.

[‡] https://www.turbosquid.com/

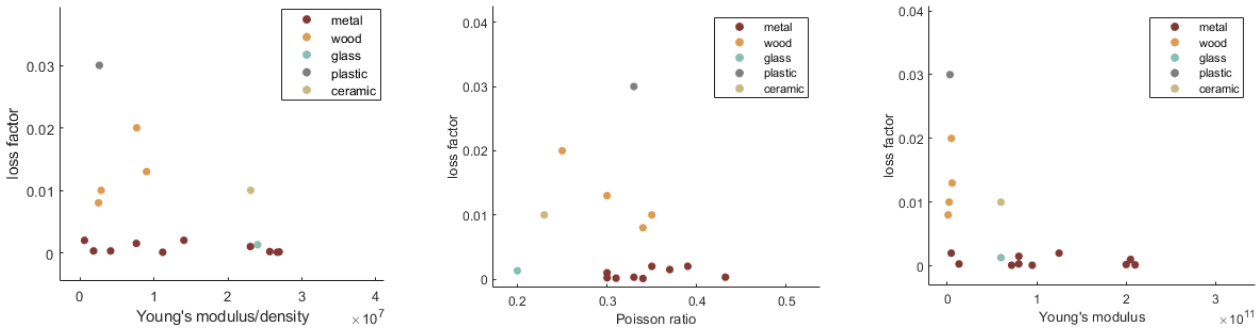


Figure 3. Scatter plots of Young's modulus/density, Poisson ratio and Young's modulus vs. loss factor for different material categories in our GLAudio dataset. Every dot represents a fine-grained material category in the dataset.

To find modal shapes and modal frequencies, simultaneous diagonalization is performed on both $\mathbf{K}$ and $\mathbf{M}$:

$$\mathbf{S}^T\mathbf{K}\mathbf{S} = \mathbf{D} \qquad \mathbf{S}^T\mathbf{M}\mathbf{S} = \mathbf{I} \tag{2}$$

where $\mathbf{D}$ denotes a diagonal matrix of generalized eigenvalue, $\mathbf{I}$ denotes the identity matrix and $\mathbf{S}$ is a nonsingular matrix. Using a coordinate transformation $\mathbf{S}\mathbf{y} = \mathbf{x}$, Eq. (1) can be simplified to

$$\mathbf{I}\ddot{\mathbf{y}} + (\alpha\mathbf{I} + \beta\mathbf{D})\dot{\mathbf{y}} + \mathbf{D}\mathbf{y} = \mathbf{S}^T\mathbf{f}. \tag{3}$$

Since both $\mathbf{I}$ and $\mathbf{D}$ are diagonal, the system is decoupled into $3n$ independent damped harmonic oscillation (modal oscillation) that can be solved easily. To generate high-quality sound samples, we adopt wave-based method for solving the sound radiation problem [26, 25]. For each modal frequency, we solve the corresponding Helmholtz equation with FMMLib3d[§]. Similar to that in [48], we add some background noise to the generated sound to make it more realistic.

3.3. Hitpoint and Local Geometry

One of the main characteristics of our datatset is that it explicitly considers the influence of the hitpoint. When generating a sound, the hitpoint on the surface plays an important role. This can be explained as follows. Let us focus on the i-th modal oscillation in Eq. (3). By defining $\mathbf{\Phi} = \mathbf{S}^T$ and λ_i to be i-th diagonal element of matrix $\mathbf{D}$, the i-th modal oscillation can be expressed as

$$\ddot{y_i} + (\alpha + \beta\lambda_i)\dot{y_i} + \lambda_i y_i = \mathbf{\Phi}_i\mathbf{f} \tag{4}$$

in which the modal shape $\mathbf{\Phi}_i$ is the i-th row of $\mathbf{\Phi}$. Assuming that an impulse represented by the Dirac's delta function $A_h\delta(t)$ is applied on node h, then the solution of this oscillation is given by

$$y_i = A_h\mathbf{\Phi}_{ih}\frac{1}{\omega_i}e^{-\xi t}\sin(\omega_i t) \tag{5}$$

[§] https://cims.nyu.edu/cmcl/fmm3dlib/fmm3dlib.html

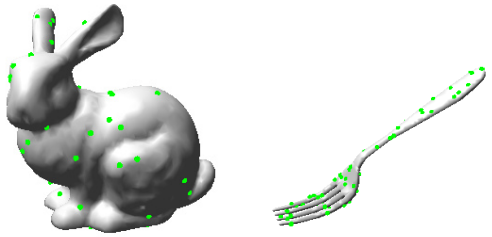


Figure 4. Illustration of hitpoints. The sampled hitpoints are marked by green points on the surface.

in which $\xi = \frac{1}{2}(\alpha + \beta\lambda_i)$ is the modal damping factor and $\omega_i = \sqrt{\lambda_i(1-\xi^2)}$ is the i-th damped natural frequency. Here, $A_h\mathbf{\Phi}_{ih}$ reflects the excitation of modal oscillation of the i-th mode and it clearly depends on the hitpoint h. This indicates that the hitpoint h is closely related to the generated sound.

We also notice that there are strong correlations between local geometry and modal shape. As demonstrated in Fig. 5, some regions of high amplitude in a modal shape may appear together, *e.g.*, near the bunny ear. According to our analysis above, when we strike these regions the related modal oscillation will be more evident than striking other regions, yielding distinct sounds. This is further illustrated in Fig. 6. We see that when striking an object at different positions, the excited amplitudes of modes, *i.e.*, $A_h\mathbf{\Phi}_{ih}$, change wildly, leading to quite different sounds. Moreover, in Fig. 7 we show that when striking a cup with or without the handhold, the generated sounds are quite similar. This indicates that the handhold has less influence on the auditory features when striking the cup body. Therefore, we can conclude that these local geometries around the hitpoints carry much more important hints about sound, compared with global geometries.

In our dataset, the local geometry around the hitpoint is constructed and stored as follows. We firstly translate the hitpoint to the origin and then voxelize the shape into a

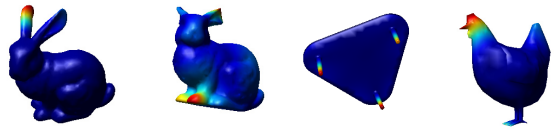


Figure 5. The relationship between local geometry and modal shape. The amplitude of modal shape is visualized by pseudo-color.

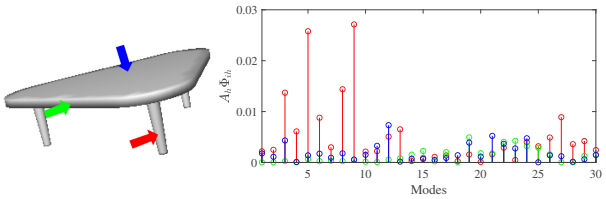


Figure 6. The excited amplitudes of modes with respect to the hitpoint.

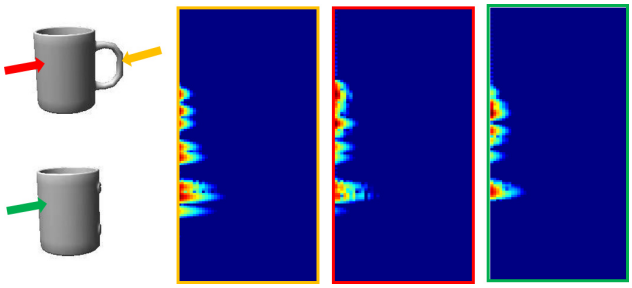


Figure 7. When striking an oak cup with (red arrow) or without (green arrow) the handhold, the generated sounds are quite similar. In comparison, striking the handhold (yellow arrow) yields a different sound.

$32\times32\times32$ grid. We drop elements which are not inside the grid centered at the origin. For each virtual object, we uniformly sample 64 hitpoints on the surface. The position of audience is uniformly sampled on the bounding sphere of the object. Each global voxel is a 3cm$\times$3cm$\times$3cm cube and each local voxel is a 0.2cm$\times$0.2cm$\times$0.2cm cube.

4. GLAVNet

With the GLAudio dataset, we are able to design deep neural networks to infer material information from both visual and auditory inputs.

4.1. Network Architecture

Since our method accepts three different inputs, we develop a multi-branch CNN as depicted in Fig. 8. It comprises four parts: a global geometry subnetwork, a local geometry subnetwork, an audio subnetwork and a fusion subnetwork. We detail them below.

Global/local subnetwork. To extract features from global/local geometry, we adopt the basic structure of VoxNet [33]. The input layer accepts a grid of fixed size: $32 \times 32 \times 32$. After passing through two 3D convolutional layers (kernel sizes: $5 \times 5 \times 5$ and $3 \times 3 \times 3$, respectively), a 3D pooling layer (kernel size: $2 \times 2 \times 2$) and a fully-connected layer, each subnetwork outputs a 384-dimensional latent vector which encodes main characteristics of the input geometry. Leaky ReLU activation is used in each convolutional layer. Since both global geometry and local geometry are voxelized with a fixed spatial resolution of $32 \times 32 \times 32$, the grid of local geometry possesses more details and hence provides more visual features that are closely related to the generated sound.

Audio subnetwork. There are various forms of representations that are suitable for modeling spectral and temporal structures of an auditory signal, among which Mel-scaled short-time Fourier transform (STFT) spectrogram has been verified to be superior for CNNs [34, 24, 48]. To generate the spectrogram of an audio clip, we firstly perform STFT to the audio clip and then compute the squared magnitude of the STFT coefficients. In our current setup, the frequencies are mapped onto Mel scale with 64 Mel bands and the time axis is spaced linearly with 40 bins, resulting in a 64×40 gray-scale image. Once the spectrogram is generated, the audio subnetwork is trained to extract useful features from it. Unlike natural images that contain complex textures, these spectrograms only have simple structures. Therefore, we use a shallow subnetwork with a 2D convolutional layer (kernel size: 5×5) and a 2D pooling layer (kernel size: 2×2) as well as a fully-connected layer, to convert each spectrogram into a 384-dimensional latent vector.

Fusion subnetwork. After obtaining three 384-dimensional vectors from global geometry, local geometry and spectrogram, respectively, a proper multi-modal feature fusion is required before feeding these vectors to the final layer. Due to the diverse behaviors of the auditory and visual signals, linear fusion models (*e.g.*, concatenation or element-wise addition) will fail to capture the complex associations between different modalities. It is likely that some weak modalities may be suppressed by other strong ones. To avoid this issue and consider the nature of three different inputs in GLAVNet, we propose a new multi-modal fusion strategy shown in the fusion subnetwork. This subnetwork contains two Multi-modal Factorized Bilinear pooling (MFB) modules[54] and a concatenation module. Each MFB module accepts visual and auditory features, expands and fuses them in the high-dimensional space. Then, the fused feature is squeezed to produce a compact output. Two fused features are then concatenated and fed to a N-way classifier consisting of two fully-connected layers with the cross-entropy loss. Simple concatenation is adopted because the two fused features now have roughly the same distribution after MFB.

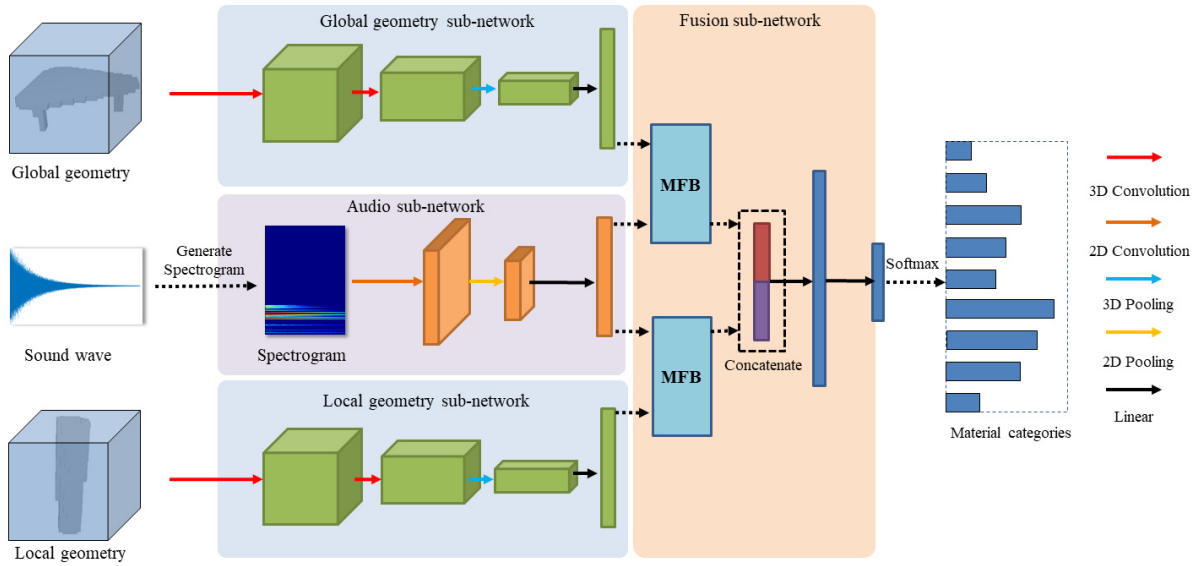


Figure 8. The network architecture of our proposed GLAVNet which consists of a global geometry subnetwork, a local geometry subnetwork, an audio subnetwork and a fusion subnetwork. After trained on our GLAudio dataset, GLAVNet can predict the fine-grained material categories based on the global geometry, the local geometry around the hitpoint and the sound.

4.2. Training Details

Our GLAVNet is implemented on top of the PyTorch framework [38]. We train it using mini-batch SGD and apply the Adam solver [27] with moment parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is fixed to 0.01. The weights of the network are all initialized using the technique described in He *et al.* [22]. Training examples are fed into GLAVNet in a mini-batch size of 64. We train the network for 100 epochs, which takes about 5 hours on one NVIDIA V100 GPU.

5. Evaluation

In this section, we conduct comprehensive experiments to evaluate our proposed method on the GLAudio dataset.

5.1. Dataset Separation

For better evaluation, we separate our synthesized examples in GLAudio dataset into a training set and several different testing sets. We first randomly select 4 3D objects and all their corresponding sounds from GLAudio and form a testing set named *GLAudioTG*. Clearly, all shapes in GLAudioTG are never seen during training. Then, we randomly select 57 pairs of geometry and material from the remaining examples. This enables that these geometry-material pairs are precluded from the training set, leading to a new testing set named *GLAudioTGM*. We ensure that all geometries in this testing set are available in the training set. For both GLAudioTG and GLAudioTGM, we conduct clustering on the hitpoints (from 64 to 32) based on the similarity of their generated sounds, because many hitpoints on a give object yield almost identical sound. Consequently, the sizes of GLAudioTG and GLAudioTGM are 2,176 and 1,824, respectively. From the remaining examples, without considering their geometries and materials, we randomly select 3,000 ones for additional testing, *i.e.*, *GLAudioTH*. This only guarantees that the geometry-material-hitpoint triples are not involved in training. All real examples form a testing set *GLAudioTR*. The final remaining synthesized examples are used in training.

5.2. Comparison with Previous Methods

To our knowledge, there are no existing methods proposed to infer material categories (coarse-grained and fine-grained) with local geometries. We still choose the following methods as baselines to compete via training their networks on our GLAudio dataset:

- **SoundNet** [3] extracts features directly from raw audio waveform via 1D convolutions. These features can be used in the task of material categorization. For this model, we test 5 and 8 convolutional layers, respectively. Only impact sounds in GLAudio are used for training.

- **ISNN-AV** [48] is an audio-visual network that uses spectrograms of impact sounds and voxelized global shapes to estimate an object's geometry and material. Unlike GLAVNet, ISNN-AV does not consider the impact of the hitpoint.

Coarse-grained material recognition. We first compare our method against these previous methods on coarse-grained material recognition. The quantitative measurements on different testing sets are listed in Table 2. On GLAudioH, our model performs on par with the state-of-the-art method ISNN-AV, both of which achieve very high accuracy. This is because all geometry-material pairs actually exist in the training set and only hitpoints vary. As hitpoints are densely sampled, many examples in GLAudioH may have analogues in the training set, which makes the accuracy extremely high. It is more reasonable to test the performance on GLAudioTG and GLAudioTGM. These two testing sets contain either geometries or geometry-material pairs that are absent from the training set, making the task of material recognition challenging. On both testing sets, our method achieves higher accuracy than its competitors. ISNN-AV performs worse on GLAudioTGM than on GLAudioTG probably because the global geometry of an object in GLAudioTGM may mislead the model to output the material of the same object (the same geometry but different material) in the training set. With the local geometry as an evident (*i.e.*, in GLAVNet), our model achieves the superior performance with reasonable FLOPs during testing. Here, the testing set GLAudioT contains all examples in GLAudioTH, GLAudioTGM and GLAudioTG, which provides the weighted average score of each model.

Fine-grained material recognition. Since GLAudio contains fine-grained material categories, we also test these methods on fine-grained material recognition. The results are provided in Table 3. On both GLAudioTG and GLAudioTGM, GLAVNet significantly outperforms other models, with an accuracy of 54.5% and 62.2%, respectively. In comparison, both SoundNet and ISNN-AV that ignore any information of hitpoint achieve much lower accuracy. We also test on two sub-categories: metal and wood, both of which can be subdivided into fine grains. Quantitative results reveal that our method still performs significantly better than others. The confusion matrices of our model and ISNN-AV are presented in Fig. 9. From the matrices, we observe that our model can distinguish different fine-grained materials better than ISNN-AV. Local geometries around hitpoints contain rich and complementary information for inferring materials, compared to global geometries and sounds. Specifically, we see that some metals which are difficult to discriminate from sounds and global geometries (*e.g.*, aluminum), are identified by the model from local geometries around the hitpoints.

5.3. Ablation Studies

Now, we conduct ablation studies to validate the effectiveness of the proposed two modules in GLAVNet: the local geometry subnetwork and the fusion subnetwork. The bottom halves of Table 2 and Table 3 list the corresponding accuracy on different testing sets.

Effectiveness of the local geometry subnetwork. To show the benefit of local geometries in our task, we design two variations of our model: GLAVNet(GG) which feeds our model with two identical global geometries and GLAVNet(LL) which feeds our model with two identical local geometries. Without local geometries, our model achieves roughly the same performance as ISNN-AV on different testing sets. This is to be expected because both GLAVNet(GG) and ISNN-AV use only global geometries and sounds as input. GLAVNet(LL) slightly outperforms GLAVNet(GG) and ISNN-AV, indicating that local geometries are more closely related to sounds. However, global geometries are still required since they can further improve the accuracy as compared in Table 2 and Table 3.

Effectiveness of the fusion subnetwork. We also test different fusion strategies in the fusion subnetwork. GLAVNet(©) leverages simple concatenation to fuse three vectors from previous subnetworks while GLAVNet($\oplus$) uses addition operation. Clearly, these two linear fusion strategies suffer from drop in accuracy since the distributions of auditory and visual features are quite different, as we explained before.

5.4. Test on Real Examples

Although our GLAVNet is trained on synthesized examples, we also test it on real examples. Specifically, on GLAudioTR which contains 43 real examples captured from 4 different objects, our GLAVNet achieves an accuracy of 53.4%, which is much higher than ISNN-AV (27.9%) and SoundNet5 (27.9%). We manually hit the real objects by a rubber hammer and record sounds by a cellphone .The geometries of real objects are captured by Microsoft Kinect. Those rough geometries are sufficient since sound is not sensitive to small variations in geometries and they are finally encoded in low-resolution voxels.These 4 objects differ in geometry from those virtual objects used for training. In this context, GLAudioTR is similar to GLAudioTG. We believe incorporating real examples into training will improve the accuracy.

6. Conclusion

In this paper, we presented a novel multimodal framework for material recognition. When adopting auditory features in such a task, we show by theoretical analysis and experimental results that local geometries around the hitpoints contain important cues that can significantly improve the accuracy of material categorization. We have constructed a new audio-visual dataset (GLAduio) that explicitly incorporate these local information related to the sound. We also have designed a new deep neural network (GLAVNet) to predict fine-grained material categories from both audio and visual inputs. We believe that our dataset is useful for

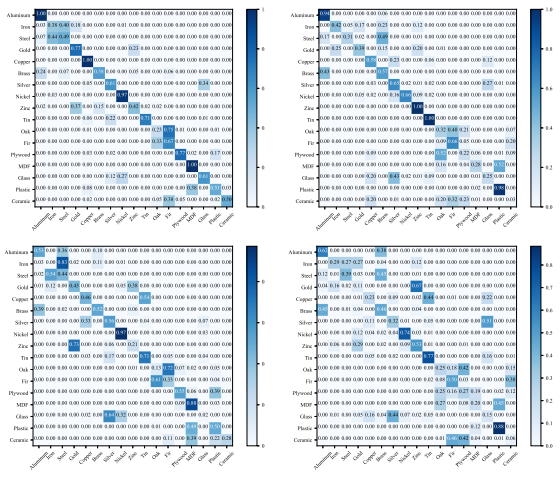
Table 2. Comparing our proposed GLAVNet with baselines for coarse-grained material categorization. GG means feeding GLAVNet with two identical global geometries and LL means feeding GLAVNet with two identical local geometries. © and $\oplus$ refer to the fusion strategies with simple concatenation and addition, respectively. The testing set GLAudioT contains all examples in GLAudioTH, GLAudioTGM and GLAudioTG. The highest accuracy on each testing set is highlighted in bold.

Method	GLAudioTH	GLAudioTGM	GLAudioTG	GLAudioT	#Params(10^6)	FLOPs(10^9)
SoundNet5	0.949	0.708	0.767	0.831	2.1	1.51
SoundNet8	0.954	0.656	0.788	0.826	6.1	0.60
ISNN-AV	**0.998**	0.822	0.831	0.901	34.8	0.61
GLAVNet	**0.998**	**0.869**	**0.850**	**0.919**	35.1	1.21
GLAVNet(GG)	0.992	0.812	0.835	0.897	—	—
GLAVNet(LL)	0.986	0.839	0.843	0.904	—	—
GLAVNet(©)	0.928	0.785	0.814	0.840	36.1	1.22
GLAVNet($\oplus$)	0.963	0.817	0.835	0.869	35.3	1.22

Table 3. Comparing our proposed GLAVNet with baselines for fine-grained material categorization. The highest accuracy on each testing set is highlighted in bold.

Method	GLAudioTH	GLAudioTGM	GLAudioTG	GLAudioT	GLAudioT(Metal)	GLAudioT(Wood)
SoundNet5	0.856	0.347	0.402	0.584	0.566	0.588
SoundNet8	0.823	0.325	0.334	0.543	0.535	0.553
ISNN-AV	**0.998**	0.403	0.398	0.658	0.658	0.688
GLAVNet	0.982	**0.622**	**0.545**	**0.753**	**0.769**	**0.765**
GLAVNet(GG)	0.990	0.392	0.434	0.663	0.651	0.719
GLAVNet(LL)	0.963	0.534	0.443	0.690	0.689	0.726
GLAVNet(©)	0.939	0.455	0.384	0.579	0.601	0.519
GLAVNet($\oplus$)	0.932	0.500	0.424	0.606	0.620	0.599

other audio-related tasks and our investigation on the relationship between local geometry and sound can spur future researches on this field. Since collecting real-world examples with high-quality geometries and sound is quite difficult, we only include a few real objects and sounds in our dataset. In the future, we would like to build a large-scale real-world dataset. Furthermore, we have only considered modal sound in impact event, other kind of sound such as acceleration noise and nonlinear thin-shell sound are still remained to include.



(a) GLAudioTGM (b) GLAudioTG

Figure 9. Confusion matrices of GLAVNet (first row) and ISNN-AV (second row) on the GLAudioTGM and GLAudioTG testing sets with fine-grained categories.

Acknowledgment

We would like to thank the reviewers for their valuable feedback. This work was supported by the National Key Research and Development Program of China (No. 2018YFB1004900) and NSFC (No. 61972194 and No. 62032011).

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3

[2] Anurag Arnab, Michael Sapienza, Stuart Golodetz, Julien Valentin, Ondrej Miksik, Shahram Izadi, and Philip H. S. Torr. Joint object-material category segmentation from audio-visual cues. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015. 1, 2

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016. 2, 6

[4] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2015. 1

[5] M. J. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):828–836, 2003. 3

[6] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Opensurfaces: A richly annotated catalog of surface appearance. *ACM Trans. Graph.*, 32(4), July 2013. 2

[7] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Material recognition in the wild with the materials in context database. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3479–3487, 2015. 1, 2

[8] B. Caputo, E. Hayman, and P. Mallikarjuna. Class-specific material categorisation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1597–1604 Vol. 2, 2005. 2

[9] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. Technical Report 1512.03012, arXiv preprint, Dec. 2015. 3

[10] J. Chen, S. Shan, C. He, G. Zhao, M. Pietik ainen, X. Chen, and W. Gao. Wld: A robust local image descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1705–1720, 2010. 2

[11] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 2

[12] M. Cimpoi, S. Maji, and A. Vedaldi. Deep filter banks for texture recognition and segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3828–3836, 2015. 2

[13] Kristin J. Dana, Bram van Ginneken, Shree K. Nayar, and Jan J. Koenderink. Reflectance and texture of real-world surfaces. *ACM Trans. Graph.*, 18(1):1–34, Jan. 1999. 2

[14] Kees van den Doel and Dinesh K. Pai. The sounds of physical shapes. *Presence: Teleoperators and Virtual Environments*, 7(4):382–395, 1998. 3

[15] Yue Dong. Deep appearance modeling: A survey. *Visual Informatics*, 3(2):59 – 68, 2019. 2

[16] Zackory Erickson, Sonia Chernova, and Charles C. Kemp. Semi-supervised haptic material recognition for robots using generative adversarial networks. volume 78 of *Proceedings of Machine Learning Research*, pages 157–166. PMLR, 13–15 Nov 2017. 1, 2

[17] Zackory Erickson, Nathan Luskey, Sonia Chernova, and Charles C. Kemp. Classification of household materials via spectroscopy. *IEEE Robotics and Automation Letters*, 4(2):700–707, 2019. 1

[18] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and José García Rodríguez. A review on deep learning techniques applied to semantic segmentation. *CoRR*, abs/1704.06857, 2017. 1

[19] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):601–616, 2007. 3

[20] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 3

[21] Eric Hayman, Barbara Caputo, Mario Fritz, and Jan-Olof Eklundh. On the significance of real-world conditions for material classification. In *Computer Vision - ECCV 2004*, pages 253–266, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. 2

[22] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. 6

[23] Diane Hu, Liefeng Bo, and Xiaofeng Ren. Toward robust material recognition for everyday objects. In *2011 British Machine Vision Conference*, pages 48.1–48.11, 01 2011. 1

[24] Muhammad Huzaifah. Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. *CoRR*, abs/1706.07156, 2017. 5

[25] Doug L. James. Physically based sound for computer animation and virtual environments. In *ACM SIGGRAPH 2016 Courses*, SIGGRAPH '16, New York, NY, USA, 2016. Association for Computing Machinery. 3, 4

[26] Doug L. James, Jernej Barbič, and Dinesh K. Pai. Precomputed acoustic transfer: Output-sensitive, accurate sound generation for geometrically complex vibration sources. *ACM Trans. Graph.*, 25(3):987–995, July 2006. 4

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 6

[28] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Comput. Vision*, 43(1):29–44, June 2001. 2

[29] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz. Exploring features in a bayesian framework for material recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 239–246, 2010. 1, 2

[30] H. Liu, F. Sun, B. Fang, and S. Lu. Multimodal measurements fusion for surface material categorization. *IEEE Transactions on Instrumentation and Measurement*, 67(2):246–256, 2018. 2

[31] Li Liu, Jie Chen, Paul W. Fieguth, Guoying Zhao, Rama Chellappa, and Matti Pietik ainen. From bow to cnn: Two decades of texture representation for texture classification. *International Journal of Computer Vision*, 127(1):74–109, 2019. 2

[32] Junhua Mao, Jun Zhu, and Alan L. Yuille. An active patch model for real world texture and appearance classification. In *Computer Vision – ECCV 2014*, pages 140–155, Cham, 2014. Springer International Publishing. 2

[33] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928, 2015. 5

[34] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao. Robust sound event classification using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):540–552, 2015. 5

[35] James F. O'Brien, Chen Shen, and Christine M. Gatchalian. Synthesizing sounds from rigid-body simulations. In *The ACM SIGGRAPH 2002 Symposium on Computer Animation*, pages 175–181. ACM Press, July 2002. 3

[36] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2405–2413, 2016. 1, 2

[37] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Computer Vision – ECCV 2016*, pages 801–816, Cham, 2016. Springer International Publishing. 2

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch, 2017. 6

[39] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. 3

[40] X. Qi, R. Xiao, C. Li, Y. Qiao, J. Guo, and X. Tang. Pairwise rotation invariant co-occurrence local binary pattern. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2199–2213, 2014. 1

[41] Zhimin Ren, Hengchin Yeh, and Ming C. Lin. Example-guided physically based modal sound synthesis. *ACM Trans. Graph.*, 32(1), Feb. 2013. 3

[42] Hardik Sailor, Dharmesh Agrawal, and Hemant Patil. Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification. In *Interspeech 2017*, pages 3107–3111, 08 2017. 3

[43] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, pages 1041–1044, New York, NY, USA, 2014. Association for Computing Machinery. 3

[44] Lavanya Sharan, Ruth Rosenholtz, and EH Adelson. Material perception: What can you see in a brief glance? *Journal of Vision*, 9:784–784, 08 2010. 1, 2

[45] Gaurav Sharma, Sibt ul Hussain, and Frédéric Jurie. Local higher-order statistics (lhs) for texture categorization and facial analysis. In *Computer Vision – ECCV 2012*, pages 1–12, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 2

[46] Hang Si. Tetgen, a delaunay-based quality tetrahedral mesh generator. *ACM Trans. Math. Softw.*, 41(2), Feb. 2015. 3

[47] L. Sifre and S. Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1233–1240, 2013. 2

[48] Auston Sterling, Justin Wilson, Sam Lowe, and Ming C. Lin. Isnn: Impact sound neural network for audio-visual object classification. In *Computer Vision – ECCV 2018*, pages 578–595, Cham, 2018. Springer International Publishing. 1, 2, 4, 5, 6

[49] Kimmo Valkealahti and Erkki Oja. Reduced multidimensional co-occurrence histograms in texture classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20:90–94, 02 1998. 2

[50] Manik Varma and Andrew Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. In *Computer Vision — ECCV 2002*, pages 255–271, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. 2

[51] M. Varma and A. Zisserman. Texture classification: are filter banks necessary? In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, 2003. 2

[52] M. Varma and A. Zisserman. A statistical approach to material classification using image patch exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2032–2047, 2009. 2

[53] Jia Xue, Hang Zhang, Kristin Dana, and Ko Nishino. Differential angular imaging for material recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6940–6949, 2017. 1, 2

[54] Z. Yu, J. Yu, J. Fan, and D. Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1839–1848, 2017. 5

[55] Hang Zhang, Jia Xue, and Kristin Dana. Deep ten: Texture encoding network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2

[56] zhoutong zhang, Qiujia Li, Zhengjia Huang, Jiajun Wu, Joshua B. Tenenbaum, and William T. Freeman. Shape and material from sound. In *Advances in Neural Information Processing Systems 30*, pages 1278–1288, 12/2017 2017. 3

[57] Zhoutong Zhang, Jiajun Wu, Qiujia Li, Zhengjia Huang, James Traer, Josh H. McDermott, Joshua B. Tenenbaum, and William T. Freeman. Generative modeling of audible shapes for object perception. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2