Acoustic Classification and Optimization for Multi-Modal Rendering of Real-World Scenes

Carl Schissler[1], Christian Loftin[2], Dinesh Manocha[3]
University of North Carolina at Chapel Hill

Abstract—We present a novel algorithm to generate virtual acoustic effects in captured 3D models of real-world scenes for multimodal augmented reality. We leverage recent advances in 3D scene reconstruction in order to automatically compute acoustic material properties. Our technique consists of a two-step procedure that first applies a convolutional neural network (CNN) to estimate the acoustic material properties, including frequency-dependent absorption coefficients, that are used for interactive sound propagation. In the second step, an iterative optimization algorithm is used to adjust the materials determined by the CNN until a virtual acoustic simulation converges to measured acoustic impulse responses. We have applied our algorithm to many reconstructed real-world indoor scenes and evaluated its fidelity for augmented reality applications.

Index Terms—Sound propagation, material optimization, recognition

1 Introduction

Recent advances in computer vision have made it possible to generate accurate 3D models of indoor and outdoor scenes from a sequence of images and videos. The resulting models are frequently used for visual rendering, physics-based simulation, or robot navigation. In many applications including computer-aided design, teleconferencing, and augmented reality, it is also important to augment such scenes with synthetic or realistic sound effects. It has been shown that good sound rendering leads to an improved sense of presence in virtual and augmented environments [1], [2].

The simplest methods for generating acoustic effects in AR/VR are based on artificial reverberation filters which use simple parametric decay models. However, the specification of these parameters is time consuming and these techniques can't simulate many effects, such as diffraction or direction-dependent reverberation. Instead, the most accurate algorithms are based on sound propagation and dynamically compute an impulse response (IR), or transfer function, based on the current position of the source(s) and the listener within the environment. The sound effects heard by the listener are computed by convolving the impulse response with unprocessed or dry source audio.

In order to simulate sound propagation within a real-world scene, a 3D surface representation is needed, usually in the form of a triangle mesh. Another important requirement for sound propagation is the need for accurate acoustic material properties for the 3D scene representation. These properties include absorption and scattering coefficients. They specify how sound waves interact with surfaces in the scene and can strongly influence the overall acoustic effects in the scene, including the reverberation time. The material properties depend on a variety of factors including the angle of sound incidence, frequency, acoustic impedance, thickness, surface roughness, and whether or not there is a resonant cavity behind the surface [3], [4], [5].

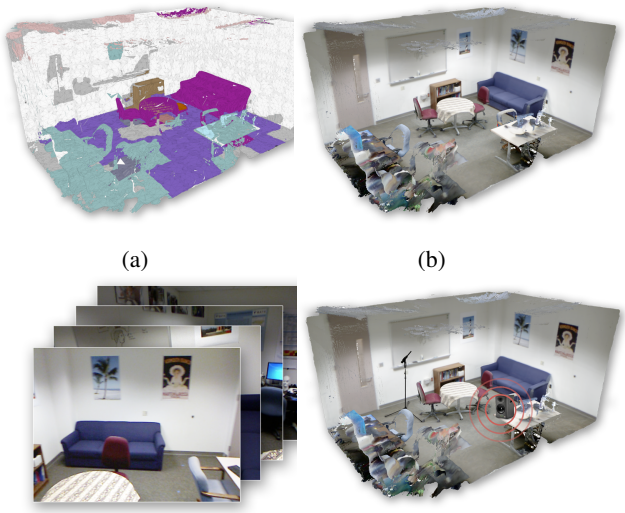


Fig. 1: Our approach automatically estimates the acoustic materials, (a), of 3D reconstructions of real-world scenes, (b), using deep learning material classifiers applied to RGB camera images, (c). We optimize the material absorption coefficients to generate sound propagation effects that match acoustic measurements of the real-world scene using a simple microphone and speaker setup (d). The synthetic sound effects are combined with visual renderings of captured models for multimodal augmented reality.

The geometric models used in current sound propagation systems are either synthetically generated (e.g., using a CAD system) or reconstructed from sensor data using computer vision techniques. However, it is difficult to determine the appropriate material properties for the 3D mesh. Current sound propagation techniques have relied on tables of measured acoustic material data that are assigned to the scene triangles or objects by a user [6].

- *Website: http://gamma.cs.unc.edu/AClassification/*
- *E-mail: [1]schissle@cs.unc.edu, [2]cloftin@cs.unc.edu, [3]dm@cs.unc.edu*

However, assigning these properties is a time-consuming manual process that requires an in-depth user knowledge of acoustic materials. Furthermore, the resulting simulations may not match known acoustic characteristics of real-world scenes due to inconsistencies between the measured data and actual scene materials.

Recent improvements in data acquisition and computer vision techniques have resulted in a large body of work on reflectance acquisition [7] which is targeted towards realistic visual rendering. However, these methods may not be directly useful for sound propagation, as they do not account for acoustic phenomena (e.g., frequency-dependent wave effects).

Main Results: We introduce a novel technique for automatically determining the acoustic material properties of 3D-reconstructed real-world scenes for multimodal augmented reality applications. Our approach builds on recent advances in computer vision and 3D scene reconstruction and augments them with a few simple acoustic impulse response measurements. We apply a convolutional neural network to the images of the real-world scene and use the result to classify the materials associated with each triangle of the 3D reconstructed model (Section 3.1). These materials are used to initialize an optimization algorithm that iteratively adjusts the frequency-dependent absorption coefficients until the resulting acoustic simulation, computed using path tracing, is similar to the measured impulse responses from the real scene (Section 3.2). The resulting 3D model and the acoustic material characteristics are used to simulate realistic sound propagation effects for augmented reality. We have evaluated our technique on several room-sized scenes and show that it is able to generate impulse responses that closely match the ground-truth measurements in those rooms (Section 5). We also present a preliminary user study that demonstrates the subjective plausibility of the sound produced by our algorithms (Section 6).

Our overall approach for capturing the acoustic characteristics of real-world scenes is designed to be simple and practical. In particular, we make little assumptions about captured acoustic data and ensure that the optimization algorithm can compute the absorption coefficients quickly. Furthermore, we use very simple acoustic sensors (e.g., a simple microphone and speaker) to capture the impulse responses of real-world scenes. To the best of our knowledge, this is the first approach for automatic computation of acoustic material properties from 3D reconstructed models for augmented reality applications.

2 Related Work

In this section we give a brief overview of prior work in sound propagation, acoustic material properties, 3D reconstruction, and visual material segmentation.

2.1 Sound and Multi-modal Rendering

It is well known that realistic sound effects can improve the sense of immersion in virtual or augment reality. Further, a greater correlation between audio and visual rendering leads to an improved sense of spaciousness in the environment [2], [8]. In order to generate realistic sound effects, we need the 3D model of the environment along with the acoustic characteristics of the materials.

Algorithms that simulate the propagation of sound in a virtual environment can be broadly divided into two major classes: wave and geometric. Wave-based techniques numerically solve the wave equation and are the most accurate sound propagation approaches. These include offline approaches like the finite-element method [9], adaptive rectangular decomposition [10], and the boundary-element method [11]. However, the computational complexity of these techniques increases dramatically with the size of the acoustic space and the simulation frequency, making them impractical for interactive applications. Other techniques like the equivalent source method [12] are designed for real-time auralization and perform considerable precomputation, but are in practice limited to static scenes and low to middle frequencies (e.g. $1-2$KHz). Wave-based precomputation approaches have also been extended to handle dynamic sources [13], but the memory and precomputation cost of these is substantial.

On the other hand, geometric sound propagation techniques are better suited for interactive applications and can generate plausible effects. These approaches simplify the computation of acoustic effects by assuming that the wavelength of sound is much smaller than the size of primitives in the scene. As a result, they are less accurate at low frequencies, where wave effects become more prominent. Image source methods [14], beam tracing [15] and frustum tracing [16] have been proposed for computing specular reflections, while Monte Carlo path tracing is commonly used for computing diffuse reflections [16], [17], [18], [19]. Geometric techniques tend to approximate diffraction as a special case of sound transport. The uniform theory of diffraction (UTD) [20] is frequently used for interactive applications, while the more accurate Biot-Tolstoy-Medwin (BTM) method is better suited for offline simulation [21]. All these methods assume that a 3D geometric model of a scene with acoustic material properties is given and the resulting sounds are generated using real-world recordings or synthesized sounds.

2.2 Acoustic Materials

The properties of an acoustic material determine how incident sound interacts with the material: how it is reflected, scattered, and transmitted through the material. While complex bidirectional reflection distribution functions (acoustic BRDFs) have been used to describe the reflectance of sound with respect to the incoming and outgoing directions [22], a lack of measured data for most material types limits their usefulness. More commonly, the reflection characteristics of an acoustic material are specified using a frequency-dependent absorption coefficient $\alpha \in [0,1]$ that determines the fraction of incident sound pressure absorbed with each reflection [4]. α is generally a function of the incident angle, θ_i, though it is common to instead average $\alpha(\theta_i)$ over θ_i to compute the random incidence absorption coefficient, α_{rand}. This value is commonly measured according to ISO 354 [23]. A sheet of the material to be tested is placed in a reverberation chamber with a known reverberation time. The change in reverberation time with the material present is used with the Sabine equation to estimate α_{rand}. A table of measured sound absorption coefficients for 66 common architectural material types can be found in [3]. The absorptive properties of a surface can also be described by the sound energy reflection coefficient, $R = \sqrt{1-\alpha}$. Our optimization approach is formulated in terms of R, rather than α.

Other models have been proposed for calculating sound scattering. The simplest use a single frequency-dependent scattering coefficient $s \in [0,1]$ that indicates the fraction of incident sound that is scattered [24]. Usually a Lambertian distribution is assumed for the scattered sound. The remainder of the sound energy $(1-s)$ is specularly reflected.

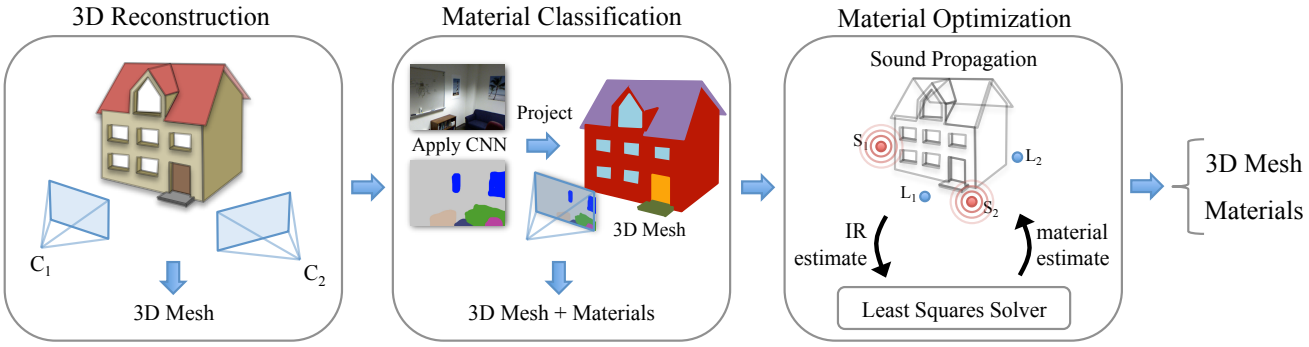


Fig. 2: Our approach begins by generating a 3D reconstruction of a real-world scene from multiple camera viewpoints. Next, a visual material segmentation is performed on the camera images, producing a material classification for each triangle in the scene. Given a few acoustic measurements of the real scene, we use the visual materials as the initialization of our material optimization algorithm. The optimization step alternates between sound propagation simulation at the measurement locations and a material estimation phase until the simulation matches the measurements. The result is a 3D mesh with acoustic materials that can be used to perform plausible acoustic simulation for augmented reality.

Many algorithms have been proposed to optimize the acoustic parameters for computer-aided design [25] or estimate the acoustic properties of real world scenes for inverse sound rendering [26], [27]. Recently, a genetic algorithm has been proposed to estimate the material properties of an existing room so that it matches measurements [28]. However, this process is time-consuming and requires many iterations. Our approach shares the same theme in terms of using acoustic measurements and is most similar in formulation to [26], but our approach is also able to handle diffuse reflections, diffraction, real-world measured IRs, and is robust to measurement noise.

2.3 3D Reconstruction

3D model reconstruction remains an active area of research in computer vision [29]. Many reconstruction approaches use two or more camera images of the same scene to estimate its structure. Passive methods use RGB images of a scene, while active reconstruction methods are based on projecting a structured light pattern into the scene [30], [31]. By analyzing the deformations of the light source with respect to the projected pattern, pixel correspondences can be found and used to compute high-quality depth maps. The relative transformation of the cameras is either known or estimated from image features. The Iterative Closest Point algorithm [32] aligns the depth image for a given frame with the structure of previously captured data, while Structure from Motion approaches match feature points (e.g., SIFT) in the images to estimate the camera poses. The depth information in the images is then fused to generate a 3D mesh in a global coordinate system. High-level plane primitive models can be used to improve the quality of the reconstruction [33]. Some of the criteria for 3D model reconstruction for sound rendering are different than for visual rendering. For example, it is important to ensure that the resulting models are watertight with no holes [34], [35]. Furthermore, we may not need to reconstruct many small features of real-world scenes. Previous work has shown that approximate mesh data can be sufficient to simulate the main acoustic characteristics of a virtual space [36], [37].

2.4 Visual Material Segmentation

Given an image of a scene, material segmentation approaches use visual characteristics to determine a material type for each pixel in the image. Liu et al. [38] combine both low-level (e.g., color) and mid-level (e.g., SIFT) image features trained in a Bayesian framework and achieve moderate material recognition accuracy. By using object and material recognition together, it has been shown that recognition results can be improved [39]. The most recent techniques are based on convolutional neural networks (CNNs) [40]. The *Materials in Context Database* (MINC) is a notable example where CNNs were trained for classification of material patches [41]. Context within the patches is used to improve the classification results, and these CNN classifiers were combined in a conditional random field (CRF) framework to perform segmentation and material recognition on every pixel in an image. Our approach builds on this work and extends these ideas to classification of acoustic materials in 3D reconstructions of real-world scenes.

3 Acoustic Materials for Reconstructed Scenes

In this section we describe our approach and how it enables sound propagation in 3D reconstructions of real-world scenes. An overview of the pipeline is shown in Figure 2. As input, our technique takes a dense 3D triangle mesh that has been reconstructed using traditional multi-camera computer vision approaches. We assume that the mesh is mostly free of holes and other reconstruction errors. Our pipeline begins by applying a CNN-based material classifier to each of the RGB camera images from the reconstruction to determine the probability that materials from a known database are present. The materials in each image are projected onto the 3D mesh and the most likely material is chosen for each material patch, where the patches are generated using a 3D superpixel segmentation algorithm. If acoustic measurements of the real scene (i.e. recorded audio samples) are available, this material information is used to initialize an optimization algorithm that iteratively refines the materials so that virtual sound propagation matches these acoustic measurements. The result is a 3D mesh and a set of materials that can be used for sound propagation and can generate virtual sounds that match those in the real environment.

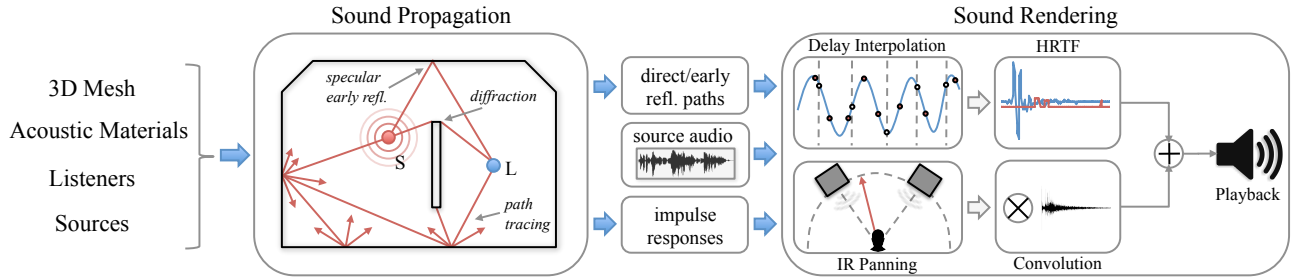


Fig. 3: The sound propagation and rendering pipeline of our system. Given the 3D mesh and optimized materials produced by our method, sound propagation is computed in the scene using separate modules for specular early reflections, diffraction, and path tracing. The result is auralized using a combination of delay interpolation for direct and early reflections paths, as well as convolution with the impulse response for late reverberation. HRTF-based spatial sound is applied to direct and early paths, while vector-based amplitude panning is used for the impulse response. The propagation path computation is also used by the optimization algorithm in Section 3.2.

3.1 Visual Material Classification for Acoustics

We present a new technique that uses the visual appearance of a real scene to estimate the acoustic material properties of the primitives. We make the assumption that there is a strong correspondence between the visual appearance of a surface and its acoustic material. For example, if a surface appears to be like brick, it is likely to have acoustic properties similar to the measured acoustic characteristics of a brick (e.g., to be highly reflective). The basis of our material classification approach is the *Materials in Context Database* (MINC) and its classifier models that have been trained for 23 common material categories [41]. From these 23 categories, we select a subset of 14 that are likely to be encountered in real scenes and discard material categories that are less relevant for acoustic simulation (e.g. hair, skin, food). We manually associate each of the categories with measured data for similar acoustic material types from [3]. For example, the MINC "brick" material category is matched with the measured absorption coefficients for the "unpainted brick" acoustic material. When there is not a one-to-one mapping between the visual material categories and the acoustic material data, we pick the most similar acoustic material in the database. This process is performed once per material category. The resulting table of material categories and their associated acoustic materials are summarized in Table 2.

The MINC CNNs were trained using 3 million material patches from 436,749 images classified by human workers on Amazon Mechanical Turk. Bell et al. [41] have shown that context, i.e. the image content surrounding a point of interest, is important in accurately classifying the materials in an image. For this reason, we choose to use images of real scenes as the input to the classification pipeline since they contain the necessary context information. For 3D scene reconstruction, a structured-light RGBD camera is used to capture the images of the scene. We use these images as the input to our material classification method. Using the approach of [33], we also generate a 3D triangle mesh for the scene with the color specified per-vertex. As part of the reconstruction, we assume that the camera projection matrix for each image is also available. These matrices are used to project the computed materials onto the mesh.

Our material classification approach is applied to each of the RGB camera images independently. We use a variant of the sliding-window approach detailed in [41] to apply the MINC trained GoogLeNet [42] to a grid of locations in each input image. The input to the network is a square image patch centered at the test location of size $p = d * 256/1100$ where d is the smaller of the image dimensions. The patches are extracted from the input image and scaled to 224×224 resolution. The mean of the patch is subtracted before it is passed through the CNN. At each test location, the CNN classifier predicts a probability for each material category. This grid of test locations is used to generate probability maps for all of the material categories. The probability maps are low-resolution images indicating the probability that a given material type is present at a position in the original image. The results are bilinearly filtered to the original image resolution and padded with zeros to maintain alignment before they are used to generate a final probability map for each camera image and material category.

3.1.1 Patch Segmentation

The next step is to determine the segmentation of material patches that should be used for the reconstructed 3D triangle mesh. These patches are localized groups of triangles that are assumed to be the same material. We use a 3D version of the SLIC superpixel segmentation algorithm [43] and the vertex colors computed during reconstruction from the RGB images to determine the segmentation. In our particular implementation, we are concerned with clustering triangles rather than voxels, so we cluster according to the interpolated color of each triangle's centroid. The first step in the SLIC algorithm is to convert the RGB color for each triangle centroid to the LAB color space. Then, the initial superpixel cluster centers in 3D space are determined by sampling the bounding box of the mesh at regular interval s on a cubic grid. The sampling interval s is determined using the relation $s = (V/k)^{1/3}$, where V is the volume of the mesh's bounding box and k is the desired number of cluster centers. The initial color values for the cluster centers are chosen to be the colors of the nearest triangle centroids. Thus, each cluster and triangle is described by an (X,Y,Z,L,A,B) tuple.

Next, the SLIC algorithm iteratively refines the clusters until the maximum error for all clusters is lower than a threshold or until the algorithm converges. First, each cluster considers all triangles within a $2s \times 2s \times 2s$ region around the center point in 3D space. The distance in XYZLAB space between the triangle centroid and cluster center is computed according to the standard SLIC distance metric, and the cluster label for each triangle is chosen to be the cluster with the smallest distance. Then, the XYZLAB cluster centers are recomputed as the average of all triangle centroids that belong to a cluster. The residual error between the

old and new cluster centers is determined using the L_2 norm in XYZLAB space. If the error is less than a threshold or if the error converges, the algorithm is terminated. The result is a collection of material patches that tend to closely match the visual features and boundaries in the reconstructed mesh. These patches are used as the basis of our material optimization algorithm (Section 3.2).

3.1.2 Material Projection

Next, the 2D classification results for all images are combined and applied to the reconstructed 3D triangle mesh. For each patch in the mesh, we create an accumulator p_i, initially set to zero, that stores the probability that the patch has the ith material type. Next, we project the material probabilities present in each image into the scene with the camera projection matrix. In our implementation, we perform this operation by tracing a ray for each image pixel. The patch intersected by the ray is updated by sampling from the probability map for the ith material type, and then we add the sampled probability to p_i. After this step has been carried out for every input image, we choose the final material for each patch to be the material with the largest p_i. By combining the results from many input images that are likely to have significant overlap, we achieve more robust material classification than could be achieved by using the results from a single image. Additionally, pooling the p_i for each material patch rather than for each triangle generates more robust material classifications that follow patch boundaries and are more likely to match the visual features of the mesh.

3.1.3 Mesh Simplification

The final step in preparing the reconstructed mesh for acoustic simulation is to simplify the dense triangle mesh. Dense 3D reconstructions frequently have triangles that are smaller than the smallest audible wavelength of 1.7cm, given by the speed of sound in the air and human hearing range. However, geometric sound propagation algorithms are generally more accurate when surface primitives are larger than audible sound wavelengths. Therefore, we apply acoustic mesh simplification techniques [19] to the dense 3D mesh and its material properties to increase the size of surface primitives and to reduce the number of edges for diffraction computation. The simplification algorithm involves a combination of voxel remeshing, vertex welding, and the edge collapse algorithm to reduce the model complexity. Boundaries between the patches are respected by the simplification so that no additional error is introduced. This results is a mesh that is appropriate for geometric sound propagation.

3.2 Acoustic Material Optimization

While visual material classification algorithms can achieve good results for visually salient materials (e.g., brick and grass), other material types may be ambiguous (e.g., painted walls) or not included in the training set. Furthermore, the materials for occluded areas in the scene are also unknown. At the same time, these occluded areas contribute to the acoustic effects of reflections and diffraction. In addition, when applied to acoustic material classification, the techniques developed for visual materials do not consider non-visual properties like density and the presence of hidden resonant cavities in the scene that can also affect the acoustic characteristics. The thickness and rigidity of walls also influences how sound propagates and these properties cannot be determined visually. As a result, a visual material classification algorithm used on the surfaces in a scene may not accurately classify the acoustic materials. Even if accurate material segmentation and classification information is known, the resulting sound simulated using that information may not match the real scene because the measured acoustic material data that is assigned to each material does not necessarily generalize to arbitrary scenes. Another issue is that holes in the 3D reconstructed mesh can cause the sound to 'leak' out of the scene, unnaturally decreasing the reverberation time. This problem can be mitigated by automatic hole-filling techniques [44], [45], but they do not always produce a correct result and can introduce other meshing errors.

In order to overcome these issues, we utilize captured acoustic measurements in the real-world scenes. We propose a second pipeline stage that optimizes the visually-classified material properties, computed using the algorithm in Section 3.1, so that the resulting acoustic simulation more closely matches the IRs of acoustic measurements taken from the real-world scene. One simple possibility would be to use the reverberation time, RT_{60}, and Sabine reverberation equation to globally modify the absorption coefficients to match the measured RT_{60}. However, the Sabine model is only valid for rectangular rooms and does not consider other important quantities like the ratio of direct to late sound energy. The RT_{60} also doesn't vary much throughout an environment, and so it doesn't provide much information about the spatial locality of absorption. As a result, an approach based on matching only the RT_{60} might lead to large errors with respect to other perceptually-relevant metrics.

Our formulation instead optimizes the sound energy reflection coefficient R for each material patch and simulation frequency band using an iterative least-squares approach in order to minimize the error between energy-time histograms from the simulation and energy-time histograms from measured IRs. This is similar to the approach of [26]. However our technique improves on several significant limitations. Their method makes the assumptions that all reflections are specular, that there is no significant diffraction, that all sound propagation paths are discrete and known to the optimization system, and that there is a one-to-one correspondence between the paths in the optimized and target IRs. These assumptions can only be satisfied if the optimization target IR is computed using the same simulation that is used during optimization, which is not the case for measured IRs. In addition, the approach of [26] only considers the early reflections computed via beam tracing and so it can't optimize the late reverberation present in real-world scenes that involves high-order diffuse reflections. These limitations prevent that method from optimizing acoustic materials to match real-world measurements.

Therefore, we introduce a new method that is able to handle the case of optimizing materials for sound rendering in real-world scenes.

3.2.1 Acoustic Measurments

The target of our optimization algorithm is a collection of impulse response measurements from the real scene. For each IR measurement, there is a corresponding source and listener placed within the virtual reconstructed 3D mesh. The target measured IR for a single source/listener pair is given by the time-domain signal $H_T(t)$, while the IR computed in the virtual scene for the same source/listener pair is given by the signal $H_S(t)$. To use these pressure IRs in our optimization approach, the first step is to filter them into the frequency bands used for the sound propagation simulation. This yields $H_{T,f}(t)$ and $H_{S,f}(t)$ for frequency band f. Then, the Hilbert Transform is applied to extract

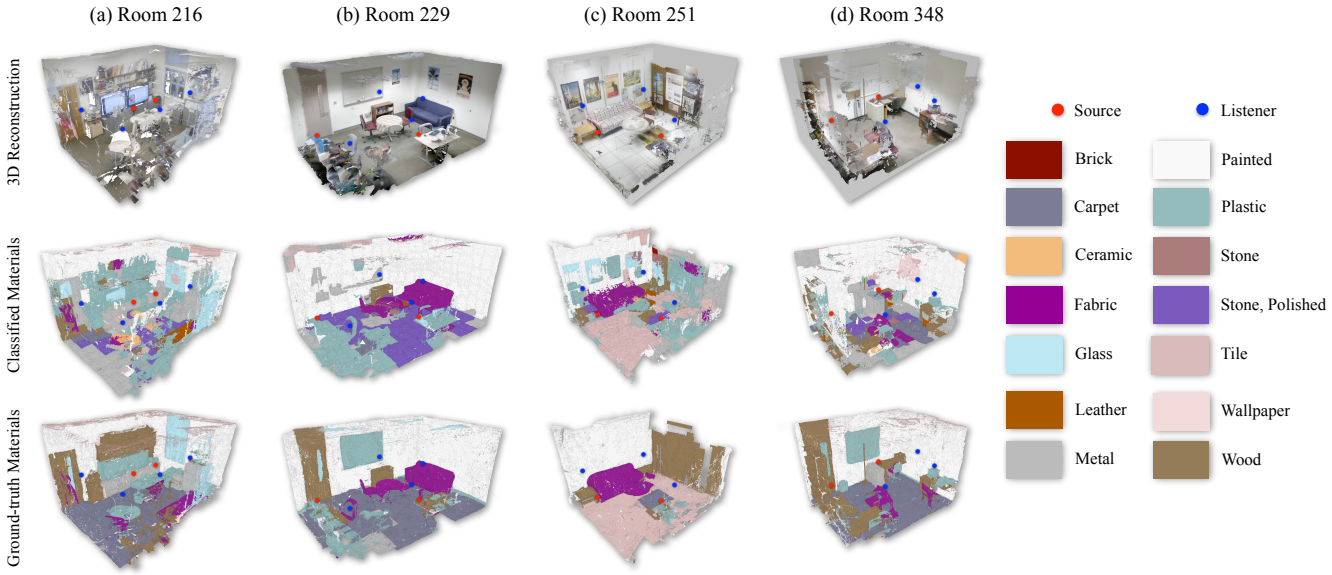


Fig. 4: The results of our visual material classification algorithm for the four benchmark scenes. Colors indicate the material category that has been assigned to each triangle of the reconstructed model. The middle row shows the results of our material classification, and the bottom row shows the manually-generated ground-truth classification that are used for validation. The source and listener positions for the acoustic measurements within the real room are shown as red and blue circles, respectively. These are used to optimize the acoustic materials present in the scenes.

the pressure magnitude envelope from the filtered IRs [46]. The square of the IR envelope then yields the energy-time curve for each impulse response, indicating the distribution of sound energy over time. The energy-time curve for the target and simulated impulse responses and frequency band f are given by $T_f(t)$ and $S_f(t)$ respectively. The high-level goal of the optimization algorithm is to minimize the error between $S_f(t)$ and $T_f(t)$. Since the number of time samples in the IRs may be on the order of 10^5, it is necessary to perform the optimization at a lower sampling rate than the audio rendering sample rate to reduce the size of the optimization problem and increase its robustness. This is done by binning the energy present in the energy-time curves to produce energy-time histograms $S_{f,b}$ and $T_{f,b}$, where b is the bin index. Thus, our algorithm in practice minimizes the error between $S_{f,b}$ and $T_{f,b}$. The energy for bin b in each IR is given by: $S_{f,b} = \sum_{t_b \in b} S_f(t_b)$ and $T_{f,b} = \sum_{t_b \in b} T_f(t_b)$. The bin size L is a parameter that determines the time resolution of the optimization and it impacts the robustness, convergence, and performance. We used $L = 10$ms.

3.2.2 IR Registration

On each iteration of our optimization algorithm, the simulated IR must be registered with the measured target IR so that it has the same time alignment and similar amplitude. This is very important for correct operation of our algorithm. If bins $S_{f,b}$ and $T_{f,b}$ do not correspond to the same time window in the IR, then the error between them can be very large and this can lead to incorrect results as the error grows on subsequent iterations. To rectify this, we propose a method for registering the IRs that is robust to the presence of noise in the measured IR. The registration operation is performed independently for every frequency band and at each optimization iteration. The first step is to compute the cross correlation between the IRs at every time offset. The simulated IR is then shifted in time to the offset where the cross correlation is highest. Once the IRs have been time aligned, the amplitudes must be matched. A significant problem with matching them robustly is that the signal-to-noise ratio (SNR) of the measured IR may be poor due to the presence of ambient noise. This noise produces incorrect registration which can lead to poor optimization performance. As a result, we only consider the bins in the IRs that have energy over the noise floor for both IRs. Given a signal-to-noise ratio for each IR, SNR_T and SNR_S, we determine the noise floors to be $\varepsilon_T = \frac{\max(T_f(t))}{SNR_T}$ and $\varepsilon_S = \frac{\max(S_f(t))}{SNR_S}$. In the case of our measurement data, $SNR_T \approx 10^4$ and $SNR_S = \infty$. Then, an energy scale factor λ for the simulated IR that minimizes the L_2 error between all bins $T_{f,b} > \varepsilon_T$ and $S_{f,b} > \varepsilon_S$ is computed using a least-squares solver. $S_{f,b}$ is multiplied by λ to yield a simulated IR that is registered to the target measured IR. The registered IRs are used on each iteration of our algorithm to estimate the error for each IR bin. The error in decibels for bin b between the simulated and target IRs is given by $E_{f,b} = \mathrm{dB}(T_{f,b}) - \mathrm{dB}(S_{f,b})$ where $\mathrm{dB}(x) = 10\log_{10}(\frac{x}{I_0})$ and I_0 is the reference sound intensity.

3.2.3 Acoustic Simulation

A key part of our algorithm is the incorporation of sound transport information from virtual simulations within the scene's 3D reconstruction. We use a ray-based geometric sound propagation system that computes $S_f(t)$ directly as the sum of many individual ray paths, e.g. $S_f(t) = \sum \delta(t - t_j) I_{j,f}$ where $I_{j,f}$ is the sound intensity for path j and frequency band f, t_j is the propagation delay time for path j, and $\delta(x)$ is the Dirac delta function. Along with $S_f(t)$, the sound propagation system also computes a *weight matrix*, W_f, for each frequency band. W_f has rows corresponding to the impulse response bins and columns corresponding to the material patches present in the scene. For IR bin b and patch m, the entry of W_f is given by $w_{f,bm} = \frac{\sum I_{j,f} d_{j,m}}{\sum I_{j,f}}$ where $d_{j,m}$ is the number of times that path j hit material patch m during its scene traversal. Therefore, $w_{f,bm}$ represents the average number of reflections

from patch m for all paths that arrived at the listener during bin b, weighted according to the sound intensity of each path. Essentially, W_f encodes the amount of influence each material patch has on every IR bin. The weight matrix is used during the optimization procedure to estimate the best changes to make to the material patches to minimize the error between $S_{f,b}$ and $T_{f,b}$.

3.2.4 Solver System

In the unlikely case where there is just one sound propagation path per bin in the impulse response, the energy for a simulated impulse response bin is given by:

$$S_{f,b} = \frac{P}{4\pi N} \prod_d R_{jfd} \tag{1}$$

where P is the sound source's power, N is the number of primary rays emitted from the source, and R_{jfd} is the frequency-dependent energy reflection coefficient encountered along path j at reflection bounce d. Converting $S_{f,b}$ to decibels by taking the logarithm allows the energy to be expressed as a linear combination of the logarithm of reflection coefficients:

$$\mathrm{dB}(S_{f,b}) = \mathrm{dB}\left(\frac{P}{4\pi N}\right) + \sum_d \mathrm{dB}\left(R_{jfd}\right). \tag{2}$$

In the approach of [26], this relationship is used to directly solve for the reflection coefficients that produce $\mathrm{dB}(S_{f,b}) \approx \mathrm{dB}(T_{f,b})$ in a least-squares sense. Given the weight matrix W_f obtained during the simulation, the system of equations solved by [26] is roughly:

$$W_f \mathrm{dB}(R_f) = \mathrm{dB}(T_f) - \mathrm{dB}\left(\frac{P}{4\pi N}\right) \tag{3}$$

where R_f is a vector of the reflection coefficients for each material patch, and T_f is a vector of the energy for the bins of the target impulse response. After solving for $\mathrm{dB}(R_f)$, the reflection coefficients can be directly determined by inverting the decibel transform. This formulation requires a one-to-one correspondence between the propagation paths in the simulated and target impulse responses and so cannot be used in the presence of diffuse reflections or diffraction because these phenomena introduce scattering that "blurs" the boundaries between paths. Their approach also requires accounting explicitly for the effects of additional acoustic phenomena such as air absorption. In addition, since it is difficult to extract discrete paths from a measured impulse response, especially for late reverberation, the technique of [26] cannot be applied to real-world measurements.

To handle these problematic cases, we reformulate the optimization problem as an approximate iterative algorithm. In the general case where many paths are assigned to the same bin, the energy in each bin of the simulated IR is given by:

$$S_{f,b} = \frac{P}{4\pi N} \sum_j \prod_d R_{jfd}. \tag{4}$$

This produces a non-linear system of equations that is difficult to handle accurately within the framework of [26]. Therefore we make the assumption that most of the paths in the same bin will hit a similar number of patches during the scene traversal. While this assumption introduces some error, it allows the use of a similar least-squares formulation. Rather than solving directly for the reflection coefficients, we instead iteratively estimate the change in decibels to each reflection coefficient that minimizes the error between the simulated and target IRs. This enables our algorithm to automatically handle a wider range of acoustic phenomena and means that it is robust to external influence from noise. The system solved on each iteration is given by:

$$W_f \mathrm{dB}(\Delta R_f) = E_f \tag{5}$$

where $\mathrm{dB}(\Delta R_f)$ is a vector of the change in decibels for each patch's reflection coefficient, and E_f is a vector of the error between the simulated and target IRs in decibels for all IR bins. The reflection coefficients are then updated on each iteration by $R'_{f,m} = R_{f,m}\Delta R_{f,m}$. To enforce physical plausibility, the reflection coefficient should be constrained to the range $[R_{min}, R_{max}]$ encountered for typical real-world materials. We use $R_{min} = 0.3$ and $R_{max} = 0.999$. For other applications (e.g. architectural acoustics) it may be useful to apply additional constraints on material placement.

This approach can be easily extended to handle the case where there are multiple measured IRs. If W_f^i is the weight matrix computed for IR i, then the final weight matrix W_f used to solve the system for all IRs is formed by vertically concatenating the rows of each W_f^i. Similarly, if E_f^i is the error in decibels for IR i, then the final error vector E_f is the vertical concatenation of the various E_f^i. The final optimized materials will then incorporate information from every measured IR.

3.2.5 Optimization

The optimization begins with the initial materials for every patch in the scene as determined in Section 3.1. Then, our iterative constrained least-squares optimization algorithm is applied to modify the materials so that the simulation better matches the real scene. The main steps of our algorithm at each iteration are summarized below:

1) For each IR in the scene, build the solver system:
   a) Compute simulated energy-time curve $S_f(t)$ and weight matrix W_f.
   b) Register simulated IR $S_f(t)$ to target measured IR $T_f(t)$.
   c) Bin $S_f(t)$ and $T_f(t)$ into energy-time histograms $S_{f,b}$ and $T_{f,b}$ with bin size L.
   d) Compute, $E_{f,b}$, the error in decibels between $S_{f,b}$ and $T_{f,b}$.
2) Solve least-squares system to get change in reflection coefficients, $\Delta R_{f,m}$.
3) Apply $\Delta R_{f,m}$ to material patches $R_{f,m}$, enforcing constraints.
4) Check termination conditions.

The algorithm terminates once a maximum number of iterations has elapsed, if the average per-bin error is less than 1 decibel, or the algorithm converges to a local minimum.

4 Implementation

In this section, we describe the implementation of various components of our approach.

3D Model Reconstruction: We generated a 3D reconstruction of each real-world scene (i.e., a room) using a few hundred RGB-D images captured with a Microsoft Kinect at 640x480 resolution. The reconstruction algorithm utilizes high-level plane primitive constraints and SIFT features [33] to compute the 3D triangle mesh that is used as an input by our acoustic classification and optimization algorithm. The captured images are also used as the input to our classification algorithm. All material classifications were performed using the Caffe deep learning framework [47] and MINC patch classifier models [41] on an Nvidia GTX 960.

Sound Propagation and Rendering: An overview of our sound propagation and rendering system is shown in Figure 3. We use a combination of the image source method for specular early reflections [14], diffuse path tracing for late reverberation [17], [19], and the UTD diffraction model for approximating edge diffraction [20]. All the sounds in our system are computed in 8 octave frequency bands centered at 63.5Hz, 125Hz, 250Hz, 500Hz, 1kHz, 2kHz, 4kHz, and 8kHz. The paths computed by the underlying propagation algorithm are used in the material optimization algorithm. The result of sound propagation is an impulse response for every source and listener pair in the scene. The impulse response is spatialized using vector-based amplitude panning [48] and the unprocessed dry source audio is convolved with the IR. We also compute perceptually important sound paths such as the direct and early reflection paths separately from the impulse response so that linear delay interpolation [49] and HRTF spatial sound can be applied independently for each path. The outputs of delay interpolation and convolution rendering are mixed and then sent to the audio device for playback over headphones (see the supplementary video).

Acoustic Measurements: The ground-truth acoustic measurements for our optimization approach consist of impulse responses at various source and listener locations in the real scenes. The measurements are captured using complimentary Golay codes [50] played through a JBL LSR4328P speaker and captured by a Beyerdynamic MM1 omnidirectional measurement microphone. The measurement setup is depicted in Figure 1 (d). These measurements for each room take about $20-30$ minutes, including the time to setup the equipment. We measured 4 impulse responses for each scene that correspond to 4 different speaker/microphone pairs. Each IR was measured 20 separate times and then the results were averaged to reduce the amount of noise present. These measured IRs are used for auralization and are also used during our optimization algorithm. The positions and orientations of the microphone and speaker with respect to the 3D reconstruction were also measured and entered into the simulation. In order to correctly replicate the directional characteristics of the speaker in the sound propagation simulation, we also measured the directional transfer function of the speaker in free space at 15 azimuths and 4 elevations for a total of 60 measurements. The magnitude response of the transfer function in the 8 octave frequency bands is used to model the directional speaker in our simulation so that there is better correspondence with the measured IRs. Overall, the audio capture requirements of our approach are low and robust to measurement errors from consumer audio equipment (e.g., with nonlinear microphone or speaker frequency response). In situations with low background noise, the impulse response can even be estimated by recording the result of the user clapping their hands.

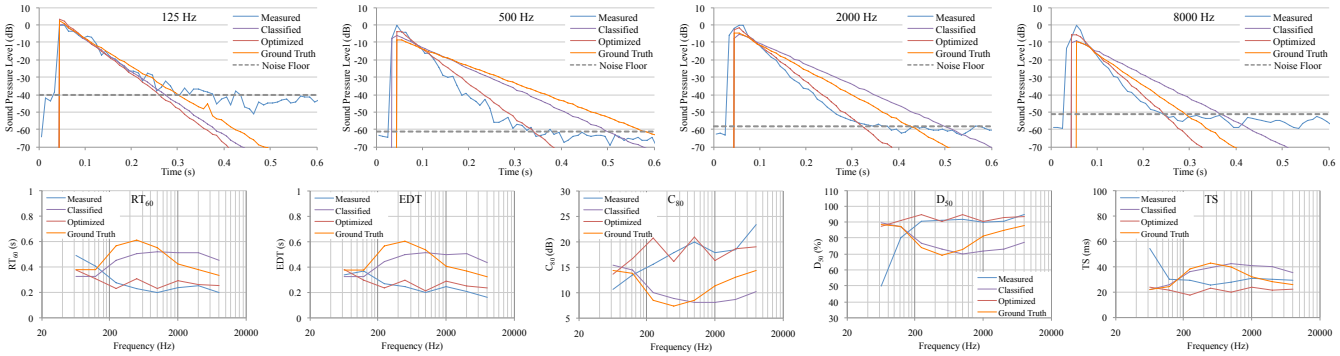
5 Results and Analysis

We have evaluated our acoustic material classification and optimization approach on several indoor real-world scenes typical of an office environment. The major characteristics and results for these scenes are summarized in Table 1. Our approach is not optimized and currently takes 6-9 hours to classify the materials in each scene. The results for the visual material classification are shown in Figure 4. We compare the output of the classification approach to a manually segmented mesh that is used as the ground truth. Overall, about 48% of the triangles in the scenes are correctly classified. For some of the materials that are incorrectly labeled (e.g. the carpet floor in Room 229 labeled as "Stone, polished"), it is possible that the CNN is unable to tell the difference between certain types of visually similar materials. A possible explanation for these results is that the input image resolution used in our implementation is less than half of the training images for the MINC dataset (1100px vertical dimension), and the RGB images contain lots of noise. There is not enough salient high-frequency information in the input images for the CNN to detect all of the material types. Another shortcoming of our automatic acoustic material classification is that the CNN cannot accurately predict materials that require higher-level knowledge of the scene. For instance, some of the test rooms contain posters on the wall that have negligible effect on the acoustics, and so the wall material (e.g. "painted") should be used. However, the CNN can be confused in these situations and produces incorrect predictions that cannot be rectified without high-level knowledge of how sound propagates through materials.

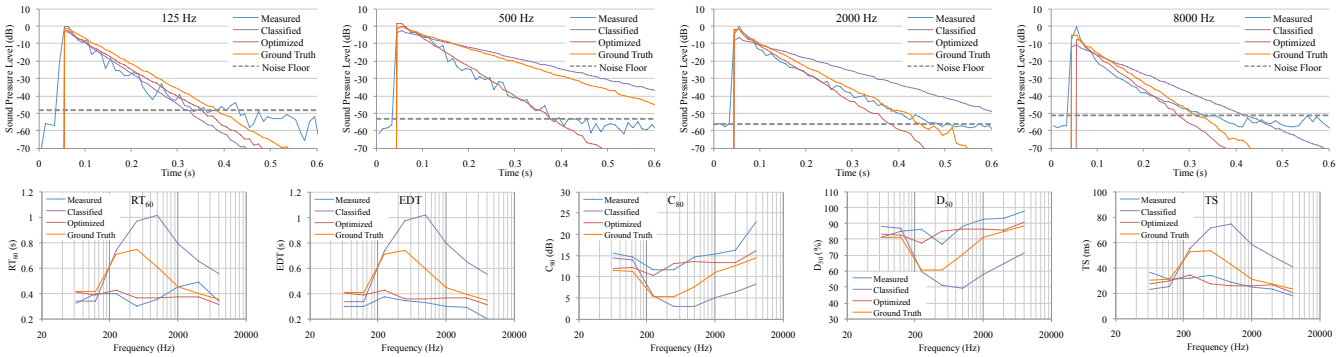
On the other hand, our acoustic material optimization algorithm for computation of absorption coefficients improves the accuracy of the simulated impulse responses with respect to the measured IRs. The results of our optimization algorithm for one IR in each scene are shown in Figure 5. Results for additional additional IRs can be found in the supplementary document. We show the energy-time histograms for the 125Hz, 500Hz, $2,000$Hz and $8,000$Hz frequency bands. For each band, we compare the measured data to the results produced directly after material classification, as well as the final results generated by our optimization algorithm. Before optimization, there are several significant mismatches with the measured data, especially in the mid-frequency bands. This can be partially explained by error during classification, though another significant factor is the table of measured absorption coefficients (Table 2), which may not always be valid for general scenes with various wall construction techniques. When the classified results are compared to the ground truth materials (before optimization), we observe similar error for both with respect to the measured data. This supports the conclusion that the material database does not match the test rooms well.

After optimization is applied, the impulse responses are much closer to the measured data. Overall, the optimization results are very close for the mid frequencies ($500-1000$Hz), with an average error of just $1-2$dB. At low frequencies there is a substantial amount of noise in the measured data, and the energy decay is not very smooth. This causes some error when compared to the optimized results that have a perfectly smooth energy decay curve. At high frequencies, we observe that the measured IRs tend to decay at a decreasing rate over time (e.g. not linear in decibels). This is most visible for Room 229 in the 8kHz frequency band. Our simulation does not reproduce this effect and it generates completely linear decay curves. This is an additional source of error that may require more sophisticated sound propagation algorithms to rectify. Noise in the measured IRs also makes it difficult to precisely match the energy decay for the entire impulse response decay. If the noise floor is set too high, then not all of the IR is considered during optimization. If the noise floor is set too low, such that the optimization considers the noise to be part of the IR, then it tends to produce a result with a slower incorrect decay rate. Setting the signal to noise ratio for each IR and frequency band is important for correct operation of

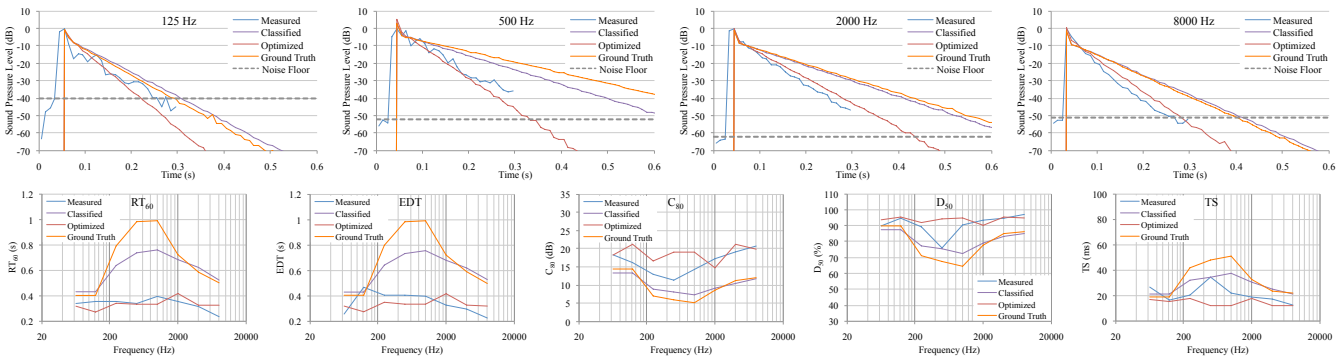
(a) Room 216, IR 1



(b) Room 229, IR 1



(c) Room 251, IR 1



(d) Room 348, IR 1

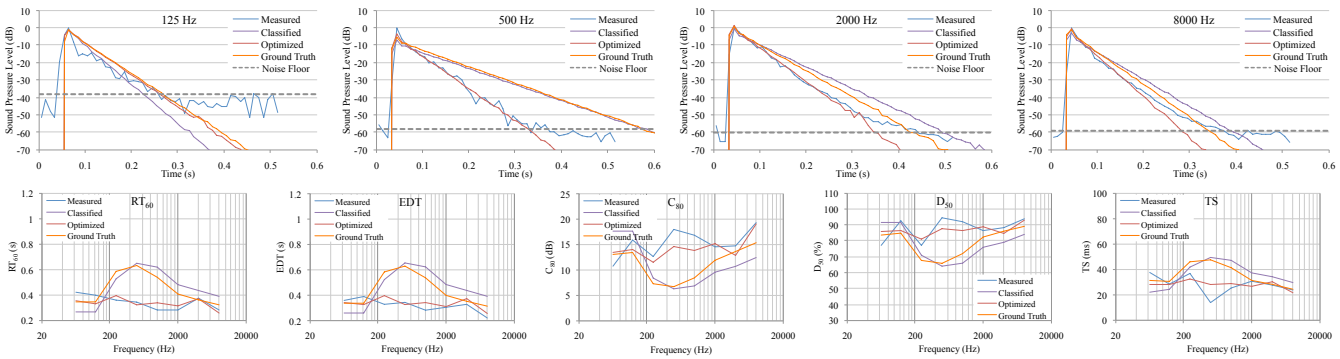


Fig. 5: The main results of our optimization approach in the benchmark scenes. We compare the energy-time curves and several standard acoustic parameters for the measured IRs (*measured*) to the results before optimization (*classified*) and the optimized results (*optimized*). We also show the results for manually-segmented materials without optimization (*ground truth*). The energy-time curves are presented for four different octave frequency bands with center frequencies 125Hz, 500Hz, 2000Hz, and 8000Hz. The noise floor corresponds to the signal to noise ratio of the measured IR for each frequency band.

Scene	Scene Complexity: Dimensions (m)	# Triangles	RT_{60} (s)	Classification: # Images	Time (hr)	% Correct	Optimization: Time (s)
Room 216	5x3.8x2.6	228,017	0.29	486	9.3	43.6	142.2
Room 229	5.3x3.6x2.6	139,545	0.44	481	9.2	52.5	154.1
Room 251	3.9x3.3x2.6	168,098	0.37	340	6.5	43.8	134.9
Room 348	4.3x3.1x2.6	131,194	0.37	481	9.2	51.4	150.5

TABLE 1: This table provides the details of the four room-sized benchmark scenarios. We give the physical dimensions of each room and the geometric complexity of the 3D reconstructed mesh models after simplification, as well as the RT_{60} values computed from the IRs. We highlight the time spent in the material classification and the optimization algorithms, as well as the percentage of scene surface area correctly classified.

the algorithm.

We also compared the results for several standard acoustic parameters: reverberation time (RT_{60}), early decay time (EDT), clarity (C_{80}), definition (D_{50}), and center time (TS). For the RT_{60} and EDT, the optimized results are close to the measured results for most frequencies, and the average error is $0.08s$. There are some mismatches at low frequencies, but these are explained by noise at the end of the IR incorrectly increasing the reported decay rates for RT_{60}. The EDT parameter is less sensitive to this noise because it only considers the first 10dB of energy decay. The C_{80}, D_{50}, and TS parameters consider the ratio of early to late sound in different ways. The results for these parameters are mixed. In some cases the optimized parameters are very close to the measured data (e.g. 1dB), but for others there are significant differences (e.g. over 5dB). A few errors are caused by noise impacting the computation of the parameters. In particular, for the computation of C_{80} and D_{50}, it is important to know the time of first arrival in the IR. If this time is not determined accurately, it can cause significant differences in the reported parameters. Other errors seem to be mostly random, so it is hard to tell what is the cause. Overall, our optimization algorithm is able to match the measured data reasonably well, though there may be room for future improvement.

Analysis: The accuracy of our approach depends on four main components: the quality of the reconstructed 3D mesh model, the resolution of the input images and machine learning approach used for material classification, the database of acoustic materials, and the sound propagation model. While we may not need to reconstruct some small features of the scene (e.g. the door knob or a book on the shelf), it is important that we get nearly watertight meshes with only small holes, otherwise the sound waves can 'leak' from those models. Furthermore, the reconstruction algorithm may not capture some hidden areas that affect the sound propagation characteristics (e.g., hidden resonant cavities or non-line-of-sight large objects). It is important to acquire high resolution input images of the scene, as that affects the accuracy of the classification algorithm. Ideally, we want to use classifier models that are trained in terms of acoustic material categories and take into account the relationship between the visual appearance and the acoustic properties, but such data does not exist. As more acoustic material databases are becoming available, we can use them to increase the fidelity of our material classification algorithm. Finally, our optimization approach is based on geometric sound propagation which may not accurately simulate all sound phenomena. As a result, the optimization may not produce impulse responses that are identical to the measured ones. Ultimately, we would like to use more accurate models for sound propagation such as wave-based methods like the Boundary Element Method, but that would increase the computational complexity of our optimization algorithm substantially. Interestingly, all these four components are active areas of research in different research communities: computer vision, learning, acoustics, and scientific computation.

Applications: The proposed technique has many possible applications where it is useful to generate physically-based sound effects for real-world scenes. On such application is teleconferencing. When a remote person is speaking, the sound from their voice can be auralized as if they are within the room. By combining our technique with head or face tracking, it may also be possible to estimate the IR between a human voice and microphone. This IR can be deconvolved with the microphone audio to approximate the dry sound from the voice. Another application is to use virtual sound sources to provide feedback for an augmented reality user interface. For example, a virtual character that is overlaid onto the real world could communicate with the user. Our technique would allow the character's voice to be auralized as if it was within the real environment. The audio could be presented through open-back headphones that produce virtual sound with little attenuation of the external sound.

6 User Evaluation

In this section, we present results from a preliminary user study that evaluates the perceptual plausibility of our acoustic classification and optimization algorithms in terms of generating plausible sounds in real-world scenes.

Study Design: In the study, we compare sound auralized using measured impulse responses, referred to as *measured*, to the sound simulated using our technique both before and after absorption coefficient optimization, referred to as *classified* and *optimized*, respectively. We evaluated 2 comparison conditions: *measured* vs. *classified* and *measured* vs. *optimized*. For each case, we tested 2 source and listener pairs for each of the 4 scenes, for a total of 16 comparisons. The study was conducted as an online survey where each subject was presented with a 16 pairs of identical videos with different audio in a random order and asked two questions about each pair. The questions were (a) "which video has audio that more closely matches the visual appearance of the scene?" and (b) "how different is the audio in the videos?". The responses were recorded on a scale from 1 to 11. For the first question, an answer of 1 indicates the audio from the left video matched the visuals better, an answer of 11 indicates the audio from the right video matched the visuals better, and an answer of 6 indicates the videos were equal. On the second question,

Visual Category	Acoustic Material	Absorption 125Hz	250Hz	500Hz	1,000Hz	2,000Hz	4,000Hz
Brick	Brick, unglazed	0.02	0.02	0.03	0.04	0.05	0.07
Carpet	Carpet, heavy, on concrete	0.02	0.06	0.14	0.37	0.60	0.65
Ceramic	N/A	0.10	0.10	0.10	0.10	0.10	0.10
Fabric	Fabric Drapery, lightweight	0.03	0.04	0.11	0.17	0.24	0.35
Glass	Glass, ordinary window	0.35	0.25	0.18	0.12	0.07	0.04
Leather	Leather seating	0.44	0.54	0.60	0.62	0.58	0.50
Metal	Steel	0.05	0.10	0.10	0.10	0.07	0.02
Painted	Gypsum board, 1/2", with 4" airspace	0.29	0.10	0.05	0.04	0.07	0.09
Plastic	N/A	0.10	0.10	0.10	0.10	0.10	0.10
Stone	Concrete, rough	0.01	0.02	0.04	0.06	0.08	0.10
Stone, polished	Concrete, smooth	0.01	0.01	0.02	0.02	0.02	0.02
Tile	Tile, marble or glazed	0.01	0.01	0.01	0.01	0.02	0.02
Wallpaper	Gypsum board, 1/2", with 4" airspace	0.29	0.10	0.05	0.04	0.07	0.09
Wood	Wood, 1" panneling, with airspace	0.19	0.14	0.09	0.06	0.06	0.05

TABLE 2: The material categories and absorption coefficient data that was used in our classification approach. For each of the visual material categories, a similar acoustic material and its absorption coefficients were chosen from [3]. For the "Ceramic" and "Plastic" categories, there was no suitable measured data available so a default absorption coefficient of 0.1 was assigned for all frequencies.

an answer of 1 indicates the videos sounded extremely similar, while an answer of 11 indicates the videos sounded very different. Our research hypotheses were: (1) the *optimized* case has sound with the same level of audio-visual correlation as the *measured* case and better correlation than the *classified* case; (2) The sound generated in the *optimized* case will be more similar to that from the *measured* case than the sound from the *classified* case.

Study Procedure: The study was completed by a total of 19 subjects between the ages of 18 and 61, made up of 17 males and 2 females. The average age of the subjects was 28, and all subjects had normal hearing. At the start of the study, subjects were given detailed instructions and filled out a demographic information questionnaire. Subjects were required to use either headphones or earbuds when taking the study, and they were asked to calibrate their listening volume using a test audio clip. Subjects were then presented the 16 pairs of videos and responded to the questions for each pair. The subjects were allowed to replay the videos as many times as needed, and they were able to move forward and backward in the study to change their answers if necessary. After rating the 16 video pairs, the study was completed.

6.1 User Evaluation Results

The main results of our user evaluation are summarized in Figure 6. A two-tailed one-sample t-test across all scenes was used to test hypothesis 1. For question (a), the subjects indicated the video with the audio that more closely matched the visual appearance of the scene. For the comparison between the *measured* case and the *classified* case, the *measured* case is slightly preferred ($p = 0.038$), with an average score of 5.3 across the scenes. This indicates that the raw output of our classification approach does not match the visual appearance of the scene very closely. However, the comparison between the *measured* and *optimized* cases indicates that there is no preference for either case ($p = 0.82$), with an average score of 6.1. The low significance supports our first hypothesis that the level of audio-visual correlation for the *measured* and *optimized* cases is similar. Therefore, our approach is suitable for augmented reality applications that require generating virtual audio that matches the appearance of the real-world scene.

For question (b), the audio differences between the 3 cases were considered. When comparing the audio from the *measured* and *classified* case, the average user score was 9.3, suggesting strong differences in the audio. On the other hand, the audio from the *optimized* case was more similar to the measured audio, with an average user score of 5.9. When the second hypothesis is evaluated using a two-tailed Welch's t-test, we find that the *optimized* sound has much fewer differences as compared to the *measured* audio than the sound without optimization ($p < 0.001$). This suggests that the optimization step is important for generating sound that is close to the real-world scene.

Overall, the responses of the subjects varied significantly across individuals, producing large standard deviations. Some subjects could reliably tell the difference between the sound conditions, but other subjects seemed to be guessing, especially for the question concerning audio-visual correlation. The inclusion of more subjects and expert listeners could improve the quality of these results.

7 Conclusions, Limitations, and Future Work

We have presented a novel technique for acoustic material classification and optimization for 3D reconstructions of real-world scenes. Our approach uses a CNN classifier to predict the material categories for 3D mesh triangles, then iteratively adjusts the reflection coefficients of the materials until simulated impulse responses match corresponding measured impulse responses. We evaluated this technique on several room-sized real-world scenes and demonstrated that it can automatically generate acoustic material properties and plausible sound propagation effects. We used the results for multimodal augmented reality that combines real-world visual rendering with acoustic effects generated using sound propagation. Our initial results are promising and we also conducted a preliminary user study that suggests that our simulated results are indistinguishable from the measured data.

Our approach has some limitations. The accuracy of our approach is governed by the sensor resolution and underlying 3D model reconstruction algorithms. The input images for our material classification system have low resolution and our approach may not work well in this case. Moreover, the current approach of assigning measured material data to the MINC material categories can produce incorrect results. The number of categories is small

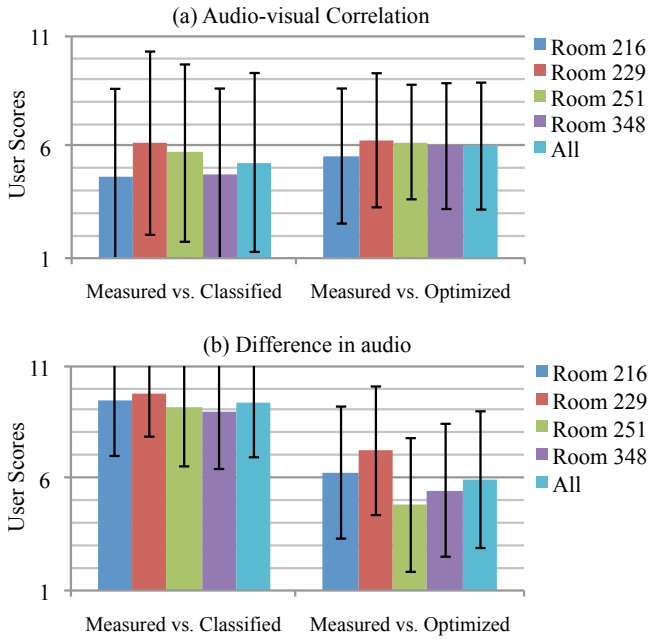


Fig. 6: We compare the user evaluation results for *measured* case versus the *classified* and *optimized* cases for 4 scenes and 2 questions. For question (a), a score below 6 indicates higher audio-visual correlation for the first method in the comparison, whereas a score above 6 indicates higher audio-visual correlation for the second method. For question (b), the higher the score, the more dissimilar the audio for the two cases under comparison. Error bars indicate the standard deviation of the responses.

and therefore the current MINC CNN model doesn't handle all real-world material variation. However, the material optimization technique proposed in Section 3.2 can be used to adjust the absorption coefficients so that the simulation is more consistent with acoustic measurements from the real scene. It is possible that the optimization may not converge to physically-accurate absorption coefficients because our approach may get stuck in local minima. Our technique also may not work well in scenes with many dynamic objects (e.g. humans) that can affect the sound. However, the impact of these objects is usually negligible if they are small in proportion to the size of the scene.

There are many avenues for future work in this domain. Most work in computer vision has been targeted towards 3D model reconstruction for visual rendering, and we need different criteria and techniques for sound rendering, as described in Section 2. Similarly, there is lack of measured data corresponding to acoustic-BRDFs for most real-world materials. It would be useful to extend recent work on visual material property acquisition [7] to acoustic materials. There is not always a one-to-one mapping between the visual material categories and the acoustic material data. This can be improved by training CNN models for new material types that disambiguate between specific acoustic material categories (e.g. painted vs. unpainted brick). Our approach also lacks high-level knowledge of the scene and materials, and so can produce incorrect material predictions where high-level knowledge is needed. Introducing additional material categories or features, such as the surface normal or size of the room, may help to classify problematic materials. Another possible avenue for improvement would be to try alternative machine learning models such as the support vector machine (SVM) or genetic algorithms. For instance, an SVM could be used to classify materials based on features extracted by the CNN [51]. We have only considered the effects of the absorption/reflection coefficient on the impulse response. It may be possible to achieve better results by simultaneously optimizing for other material attributes like the scattering coefficient s, or by considering acoustic metrics like RT_{60}, the early decay time (EDT), or clarity (C_{80}) as constraints. We would also like to further evaluate our technique on larger indoor scenes with varying reverberation effects (e.g. cathedrals, concert halls), as well as outdoor scenes.

Acknowledgments

The authors would like to thank the anonymous reviewers for their feedback, as well as the anonymous participants of the user study. This work is supported in part by Army Research grant W911NF-14-1-0437.

References

[1] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *Journal of the Audio Engineering Society*, vol. 52, no. 6, pp. 618–639, 2004.

[2] P. Larsson, A. Väljamäe, D. Västfjäll, A. Tajadura-Jiménez, and M. Kleiner, "Auditory-induced presence in mixed reality environments and related technology," in *The Engineering of Mixed Reality Systems*. Springer, 2010, pp. 143–163.

[3] M. D. Egan, *Architectural Acoustics*. McGraw-Hill Custom Publishing, 1988.

[4] H. Kuttruff, *Acoustics: An Introduction*. CRC Press, 2007.

[5] H. S. Seddeq, "Factors influencing acoustic performance of sound absorptive materials," *Aust. J. Basic Appl. Sci*, vol. 3, no. 4, pp. 4610–7, 2009.

[6] J. H. Rindel and C. L. Christensen, "Odeon, a design tool for noise control in indoor environments," in *Proceedings of the International Conference Noise at work. Lille*, 2007, pp. 1–9.

[7] M. Weinmann and R. Klein, "Advances in geometry and reflectance acquisition," in *SIGGRAPH Asia 2015 Courses*, ser. SA '15, 2015, pp. 1:1–1:71. [Online]. Available: http://doi.acm.org/10.1145/2818143.2818165

[8] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[9] L. L. Thompson, "A review of finite-element methods for time-harmonic acoustics," *J. Acoust. Soc. Am*, vol. 119, no. 3, pp. 1315–1330, 2006.

[10] N. Raghuvanshi, R. Narain, and M. C. Lin, "Efficient and accurate sound propagation using adaptive rectangular decomposition," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 5, pp. 789–801, 2009.

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

[12] R. Mehra, N. Raghuvanshi, L. Antani, A. Chandak, S. Curtis, and D. Manocha, "Wave-based sound propagation in large open scenes using an equivalent source formulation," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 2, p. 19, 2013.

[13] N. Raghuvanshi, J. Snyder, R. Mehra, M. Lin, and N. Govindaraju, "Precomputed wave simulation for real-time sound propagation of dynamic sources in complex scenes," in *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4. ACM, 2010, p. 68.

[14] J. Borish, "Extension to the image model to arbitrary polyhedra," *The Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1827–1836, June 1984.

[15] T. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West, "A beam tracing approach to acoustic modeling for interactive virtual environments," in *Proc. of ACM SIGGRAPH*, 1998, pp. 21–32.

[16] M. Taylor, A. Chandak, L. Antani, and D. Manocha, "Resound: interactive sound rendering for dynamic virtual environments," in *MM '09: Proceedings of the seventeen ACM international conference on Multimedia*. ACM, 2009, pp. 271–280.

[17] M. Vorländer, "Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm," *The Journal of the Acoustical Society of America*, vol. 86, no. 1, pp. 172–178, 1989.

[18] J. J. Embrechts, "Broad spectrum diffusion model for room acoustics ray-tracing algorithms," *The Journal of the Acoustical Society of America*, vol. 107, no. 4, pp. 2068–2081, 2000.

[19] C. Schissler, R. Mehra, and D. Manocha, "High-order diffraction and diffuse reflections for interactive sound propagation in large environments," *ACM Transactions on Graphics (SIGGRAPH 2014)*, vol. 33, no. 4, p. 39, 2014.

[20] N. Tsingos, T. Funkhouser, A. Ngan, and I. Carlbom, "Modeling acoustics in virtual environments using the uniform theory of diffraction," in *Proc. of ACM SIGGRAPH*, 2001, pp. 545–552.

[21] U. P. Svensson, R. I. Fred, and J. Vanderkooy, "An analytic secondary source model of edge diffraction impulse responses," *Acoustical Society of America Journal*, vol. 106, pp. 2331–2344, Nov. 1999.

[22] G. Mückl and C. Dachsbacher, "Precomputing sound scattering for structured surfaces," in *Proceedings of the 14th Eurographics Symposium on Parallel Graphics and Visualization*, 2014, pp. 73–80.

[23] ISO, "ISO 354, Acoustics—Measurement of sound absorption in a reverberation room." *International Standards Organisation*, no. 354, 2003.

[24] C. L. Christensen and J. H. Rindel, "A new scattering method that combines roughness and diffraction effects," in *Forum Acousticum, Budapest, Hungary*, 2005.

[25] M. Monks, B. M. Oh, and J. Dorsey, "Audioptimization: goal-based acoustic design," *Computer Graphics and Applications, IEEE*, vol. 20, no. 3, pp. 76–90, 2000.

[26] K. Saksela, J. Botts, and L. Savioja, "Optimization of absorption placement using geometrical acoustic models and least squares," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. EL274–EL280, 2015.

[27] G.-P. Nava, "Inverse sound rendering: In-situ estimation of surface acoustic impedance for acoustic simulation and design of real indoor environments," Ph.D. dissertation, University of Tokyo, 2006.

[28] C. L. Christensen, G. Koutsouris, and J. H. Rindel, "Estimating absorption of materials to match room model against existing room using a genetic algorithm," in *Forum Acusticum 2014, At Krakow, Poland*, 2014.

[29] K. Chen, Y.-K. Lai, and S.-M. Hu, "3d indoor scene modeling from rgb-d data: a survey," *Computational Visual Media*, vol. 1, no. 4, pp. 267–278, 2015.

[30] J. Batlle, E. Mouaddib, and J. Salvi, "Recent progress in coded structured light as a technique to solve the correspondence problem: a survey," *Pattern recognition*, vol. 31, no. 7, pp. 963–982, 1998.

[31] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *IEEE Computer Vision and Pattern Recognition.*, vol. 1. IEEE, 2003, pp. I–195.

[32] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*. IEEE, 2011, pp. 127–136.

[33] M. Dou, L. Guan, J.-M. Frahm, and H. Fuchs, "Exploring high-level plane primitives for indoor 3d reconstruction with a hand-held rgb-d camera," in *Computer Vision-ACCV 2012 Workshops*. Springer, 2013, pp. 94–108.

[34] M. Sormann, C. Zach, J. Bauer, K. Karner, and H. Bishof, "Watertight multi-view reconstruction based on volumetric graph-cuts," in *Image analysis*. Springer, 2007, pp. 393–402.

[35] S. Y. Bao, A. Furlan, L. Fei-Fei, and S. Savarese, "Understanding the 3d layout of a cluttered room from multiple images," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*. IEEE, 2014, pp. 690–697.

[36] S. Siltanen, T. Lokki, L. Savioja, and C. Lynge Christensen, "Geometry reduction in room acoustics modeling," *Acta Acustica united with Acustica*, vol. 94, no. 3, pp. 410–418, 2008.

[37] S. Pelzer and M. Vorländer, "Frequency-and time-dependent geometry for real-time auralizations," in *Proceedings of 20th International Congress on Acoustics, ICA*, 2010, pp. 1–7.

[38] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz, "Exploring features in a bayesian framework for material recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 239–246.

[39] D. Hu, L. Bo, and X. Ren, "Toward robust material recognition for everyday objects." in *BMVC*, vol. 13, 2011, p. 14.

[40] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3828–3836.

[41] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 00, pp. 3479–3487, 2015.

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[43] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[44] J. Wang and M. M. Oliveira, "A hole-filling strategy for reconstruction of smooth surfaces in range images," in *Computer Graphics and Image Processing, 2003. SIBGRAPI 2003. XVI Brazilian Symposium on*. IEEE, 2003, pp. 11–18.

[45] J. Branch, F. Prieto, and P. Boulanger, "Automatic hole-filling of triangular meshes using local radial basis function," in *3D Data Processing, Visualization, and Transmission, Third International Symposium on*. IEEE, 2006, pp. 727–734.

[46] J. S. Bendat and A. G. Piersol, "The hilbert transform," *Random Data: Analysis and Measurement Procedures, Fourth Edition*, pp. 473–503, 1986.

[47] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[48] V. Pulkki *et al.*, *Spatial sound generation and perception by amplitude panning techniques*. Helsinki University of Technology, 2001.

[49] E. M. Wenzel, J. D. Miller, and J. S. Abel, "A software-based system for interactive spatial sound synthesis," in *ICAD, 6th Intl. Conf. on Aud. Disp*, 2000, pp. 151–156.

[50] S. Foster, "Impulse response measurement using golay codes," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, vol. 11. IEEE, 1986, pp. 929–932.

[51] Y. Tang, "Deep learning using linear support vector machines," in *ICML Workshop on Challenges in Representation Learning*, 2013, pp. 1–6.