A Texture Superpixel Approach to Semantic Material Classification for Acoustic Geometry Tagging

The current state of audio rendering algorithms allows efficient sound propagation, reflecting realistic acoustic properties of real environments. Among factors affecting realism of acoustic simulations is the mapping between an environment's geometry, and acoustic information of materials represented. We present a pipeline to infer material characteristics from their visual representations, providing an automated mapping. A trained image classifier estimates semantic material information from textured meshes mapping predicted labels to a database of measured frequency-dependent absorption coefficients; trained on a material image patches generated from superpixels, it produces inference from meshes, decomposing their unwrapped textures. The most frequent label from predicted texture patches determines the acoustic material assigned to the input mesh. We test the pipeline on a real environment, capturing a conference room and reconstructing its geometry from point cloud data. We estimate a Room Impulse Response (RIR) of the virtual environment, which we compare against a measured counterpart.

CCS Concepts: • **Computing methodologies** → **Image processing**; **Rendering**; *Neural networks*; *Simulation evaluation.*

Additional Key Words and Phrases: audio rendering, acoustic modelling, geometry reconstruction, material recognition, semantic networks.

ACM Reference Format:

. 2018. A Texture Superpixel Approach to Semantic Material Classification for Acoustic Geometry Tagging. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The realism of digital media has increased in recent years thanks to recent advances in computer games technology [28]. Sound describes entities with respect to the acoustic environment they exist in. As listener in a sound transmission, the human auditory system is aware of acoustic characteristics manifesting in auditory cues, which enable spatial hearing abilities such as sound localisation aiding interaction tasks with objects in the world. Intrinsic acoustic characteristics are dependent on the sound transmission's wavefield, dictated by structural properties of the environments associated with boundaries and materials, as they interact with sound propagating to the listener's ears. Acoustic rendering methods simulate real or virtual auditory environments by deriving from sound propagation algorithms that discretise the geometrical representation of an environment to synthesise a wavefield. They render spatialised sound adopting signal processing chains to reproduce realistic sound transmission in the simulated wavefield, considering the listener's position, orientation and their physical characteristics, described by Head-Related Transfer Function (HRTF) [15].

In Virtual Environments (VEs), acoustic rendering can reproduce spatial hearing abilities [21], supporting architectural acoustics, cultural heritage [4, 32], and computer games [24, 27] to build compelling, realistic acoustic simulations. Recent advances in wavefield synthesis have made it easier and computationally feasible to apply to VEs [2, 27]. They

Manuscript submitted to ACM

draw on geometrical acoustics, wave-based or hybrid sound propagation algorithms, simulating sound propagation by tracing rays or beams [15]; solving the wave equations at discretised junctures of the representation of the environment or by a combination of the former techniques. They enable game designers to apply realistic, spatialised audio to games. However, the accuracy of the acoustic simulation depends on material information assigned to the scene geometry. The scene geometry, tagged with frequency-dependent absorption and scattering information, determines how sound behaves in space and affects the resulting wavefield. In games development, the process of tagging materials with appropriate acoustic data often requires the work of experts, raising costs and resources needs for large scenes.

Advances in acoustic modelling propose automatic tagging of acoustic data to scene geometry using convolutional neural networks to tag acoustic materials from stereo photographs of real environments [19]. Alternatively, using a recent camera-based material tagging system tags geometry in VEs, applying scene understanding algorithms and filtering complex geometry based on its perceptual impact on the resulting acoustic model [7]. At the core of these methods lies the problem of scene segmentation. In computer games, often, a set of meshes composes a scene, where each mesh represents an object in the scene. Acoustic materials are often assigned to all triangles composing a given mesh, allowing audio engineers to group scene geometry when assigning acoustic data. Hence, the resulting acoustic model's accuracy depends on the separation of the geometry, where ideal conditions would have each triangle mapped to its specific acoustic data. A naive approach would have the entire geometry mapping to a single acoustic material. Besides, the representation of materials in real and virtual environments adds further dimensions to the material tagging problem due to complex links between the visual representation of materials of an object and its perceptual effects on the soundscape of the environments in which it exists [5].

Here, we propose a novel architecture for tagging acoustic material in virtual environments, which improves upon recent work by abstracting away from camera-based systems and tests vision-based material recognition methods in real environments. Despite the significant progress made in sound propagation over the last decades, there are still many limitations in simulations for indoor and outdoor environments due to the complexity of the factors that describe a wavefield [20]. The contributions of this paper are:

- an improvement in the application of acoustic rendering to virtual environments;
- a novel architecture for recognising materials from textured meshes in complex scenes reducing the need for manual tagging of acoustic materials;
- an objective evaluation of the architecture conducted on a virtual reconstruction of a real conference room.

2 RELATED WORK

The following section provides an overview of the current state of acoustic rendering systems for virtual environments, as well as a review on methods for recognising material from visual representations.

2.1 Acoustic Rendering

Spatial sound has been shown to be influential in having a significant effect on the sense of presence and immersion for a user in a VE [26]. Factors of accurate and plausible acoustic rendering include geometry, material definitions and a room impulse which describes the attenuation of sound from a sound source to a listener and there exist approaches that tackle varying aspects of these factors. Kim et al. [16] introduced the first approach to geometry estimation using scene understanding inferring acoustic characteristics from visual representations of environments. Their pipeline identifies isotropic features in synthesised directional impulse responses, expressed as independent parameters considering

direction, time of arrival and spatial information of the sound source with respect to the listener's position. Li et al. [19] identify a novel method for acoustic simulations using convolutional neural networks to perform acoustic analysis on videos, veering away from more formal 3D scene definition. This approach synthesises RIRs for environments' representations from audio-visual scenes. Tang et al. [23] present a model for simulating sound fields using neural networks without pre-computing the wave field of an acoustic environment, predicting unseen objects with arbitrary shapes in a VE for sound propagation at interactive rates. They train a geometrical neural network on annotated meshes to infer acoustic data associated with the represented object.

A common denominator in the sound rendering methods mentioned is the problem of mapping visual representation of environments to corresponding acoustic materials, which intersects image processing and computer vision domains aimed at modelling how human vision recognise materials.

2.2 Material Recognition

Recent developments in deep learning techniques have contributed to a dramatic accuracy increase in tasks such as image classification. Specifically, convolutional neural networks have been broadly adopted to learning functions mapping between image data and various semantic descriptors, such as local object classes [22], perceptual sensitivity [11], or subjective quality [6]. For example, Lagunas et al. [18], present a method to learn similarities between materials based on their appearance and distinguish them in a feature space, informed by human perception. They describe mappings between subjective perception and physical material parameters. This is a challenging task due to the impact of low-level properties, such illumination and reflectance on the appearance of materials. The authors address this problem using deep features learned by a neural network trained on a bespoke dataset, annotated with around about 100 classes of materials, captured under different conditions, including surface shape, illuminance and reflectance, expressed by environment maps and bidirectional reflectance distribution functions. In a subjective study they encode materials in a perceptually-informed feature space, allowing for calculation of perceptual distances. Semi-supervised approaches have also been adopted in tackling such problems. For example, Gaur and Manjunath [12] propose a novel deep learning architecture to cluster materials from a given dataset, improving state-of-the-art superpixel algorithms by combining segmentation of images into perceptually meaningful pixel clusters with a novel unsupervised clustering method based on superpixel embeddings. A novel loss function uses a variable-margin that compensates the limitations of classic superpixel algorithms in segmenting texture patterns, allowing the convolutional neural network to cluster superpixel labels based on their embeddings requiring no manual supervision or annotations.

Approaches like these have demonstrated the efficacy of deep learning techniques for learning complex nonlinear mappings directly from annotated image data. This work leverages and adapts such techniques for the purpose of local material classification, allowing for the first step in mapping from image features to acoustic absorption coefficients.

3 METHOD

3.1 System Overview

We present a method for processing scene geometry generating acoustic models by predicting materials of objects composing complex scenes. A breakdown of the system shown in Figure 1 illustrates how visual representations of the environment maps to acoustic data used by sound renderers to model sound propagation in the scene.

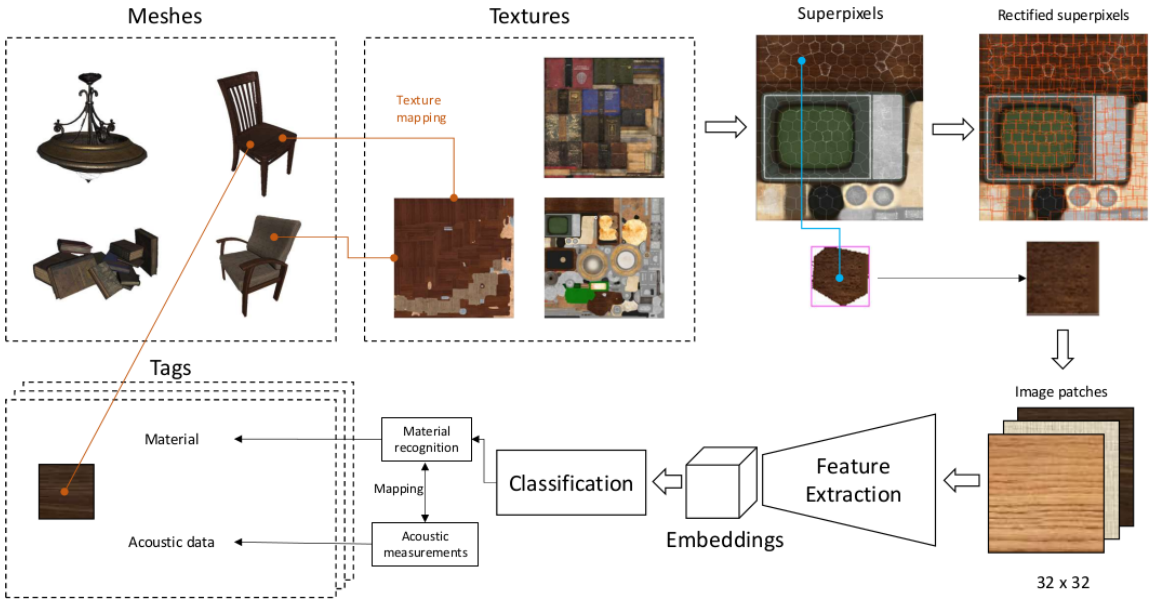


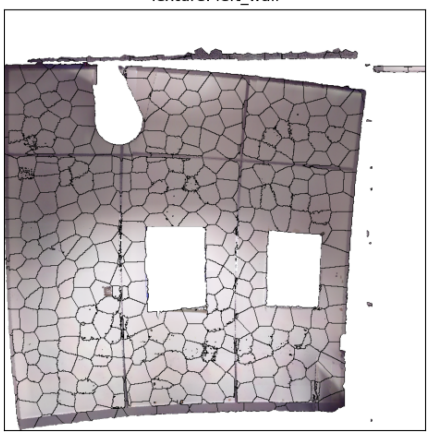
Fig. 1. Overview of the system's workflow for inference: We extract superpixels from a set of textures associated with meshes in a virtual environment to subdivide materials represented into patches which provide input for a feature extractor. A classifier assigns a label to each patch, based on its embeddings, mapping it to measurements of the corresponding material in an acoustic database. The resulting tagging process associates such information to meshes.

3.2 Material Recognition

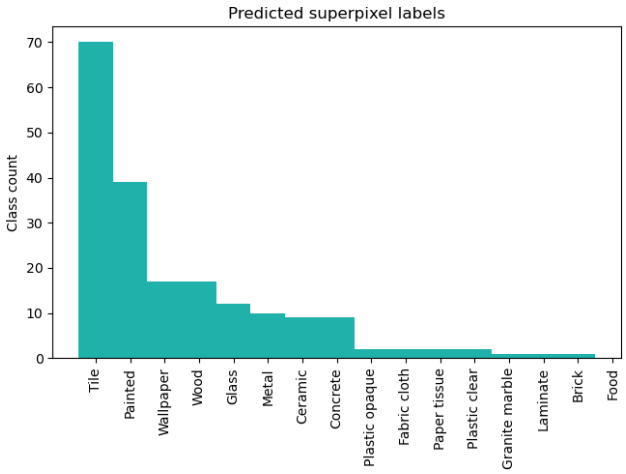
According to [29], small image patches contain enough information to distinguish materials and hence, we decompose input image textures into small image patches.

3.2.1 Training. We determine the visual material space by applying transfer learning to the OpenSurfaces dataset [3], which comprises 36 classes of surface photographs. The SLIC algorithm [1] segments input surface photographs into a set of superpixel labels, which determine regions correlated with boundaries of objects. From these resulting superpixel labels, we then generate rectified image patches encapsulating their contours through edge detection [10]. The rectified image patches are fed through a ResNet50 [13], used as a feature extractor for a classification network using a standard fully connected layer to predict class labels based on embeddings of 32x32 pixel input patches. We train the network on 13677819 input patches, composing a train set of about 9.1M images and an evaluation set of about 4.5M, adopting the standard Adam optimiser [17] to update the weights initialised from the ImageNet dataset [9]. The model usually converges in 45 epochs with a training and validation accuracy of about 0.94 and 0.83 respectively.

3.2.2 Inference. Given the set of textured meshes in a scene, we unwrap textures as images to predict materials represented. The trained ResNet50 extracts features from input image textures in complex scenes, whose embeddings enable the classifier to predict class labels associated with each input superpixels. The most frequent prediction maps to an acoustic measurement database, defining the output acoustic material.On average, the classifier takes 11.2s to determine the acoustic material for a given mesh, see Figure 2, divided in 3.8s for generating rectified patches and 7.3s to extract features and compute the mapping.



(a) An input image texture with superimposed superpixel labels.



(b) A histogram of predicted material labels based on input patches generated from superpixel labels.

Fig. 2. Example of material recognition applied on a texture from a given mesh, left image. Our system generates image patches using superpixel labels, predicting several material classes and selecting the most frequent prediction.

3.3 Acoustic Mapping

Material labels inferred from textures are associated with acoustic measurements of absorption coefficients. For every label, a one-to-many mapping groups measurements of the given material. Following the methodology in [16], we use median frequency-dependent values to determine acoustic absorption, defining acoustic materials. A single acoustic material maps to each given mesh, associating a vector of acoustic absorption coefficients to its triangles, determining the overall acoustic mapping accuracy to depend upon the mesh separation of the scene geometry. In the example texture shown in Figure 2b, "Tile" defines the acoustic material, as per predictions shown in Figure 2a.

4 EXPERIMENTAL EVALUATION METHODOLOGY

We compare acoustic models estimated with state-of-the-art wave-based audio renderers, testing whether the proposed system for the automatic tagging of scene geometry has a significant impact on the resulting acoustic model. The experimental evaluation uses a real environment as a benchmark for testing how predicted models express acoustic parameters of measurable space.

4.1 Scene Geometry Reconstruction

Wavefields are simulated from a real conference room that is $3.5346m$ deep, $2.8367m$ wide and $3.5149m$ tall. Theoretically, the dimensions determine a Schroeder frequency of $261Hz$. The room is reconstructed using the Unity game engine with $2.9M$ of triangles and $6.6M$ vertices. We reconstruct the geometry of a conference room to deploy the proposed system on a real environment and conduct the subsequent subjective experimental evaluation. We adopt a LiDAR scanner, FARO Focus^{3D} X300, to capture several point clouds scans of the room across 8 positions: 4 position points for each corner of the room at about $1m$ height and 4 additional positions at about $0.2m$ height to capture furniture and materials from different angles and enhance the spatial resolution.

Fig. 3. Real environment used for the experimental evaluation: on the left, a picture of a conference room; on the middle, a point cloud recostruction from LiDAR scans; on the right, the environment reconstructed as a set of meshes triangulated from the point cloud data. The red sphere (middle image) indicates the listener position in the virtual reconstruction used for all acoustic measurements.

4.2 RIR Measurement

The experimental evaluation compares simulated wavefields generated from the reconstructed environment to a sample measurement of the real counterpart's wavefield. We compare wavefields using Room Impulse Responses (RIRs) to describe acoustic properties dependent on geometry and materials surrounding a sound transmission between a source and a listener [30]. We capture the environments' acoustic characteristics emitting and recording an exponentially swept sine, ranging from $20Hz$ to $20kHz$, from a listening to a receiving point that is consistent between the real and reconstructed environments, see Figure 1. Applying inverse filtering, we recover the RIRs from the captured signal [14]. Given the single listener position, all RIRs are mono.

4.2.1 Real environment. In the conference room, a public address system, the dB Technologies ES 1002, emitted the exponentially swept sine converted from a laptop using an Audient ID14 DAC and ADC, which captured the signal back through an omnidirectional measurement microphone, the Earthwork M30.

4.2.2 Reconstructed Environment. With Steam Audio [2], we simulate wavefields of the reconstructed environment. This allows synthesis of wavefields based on acoustic geometry, considering absorption coefficients expressed across three frequency bands: low, medium and high. All simulations share the same resolution of 65536 and 16384 direct and secondary rays, respectively, with 256 bounces off solid geometry. With the same setup, we synthesise three wavefields with different acoustic materials: the *generic* with a single acoustic material for the entire scene geometry; the *tagged*, with acoustic materials assigned through manual material tagging; and *ours*, using the proposed automatic tagging.

4.3 Perceptual test

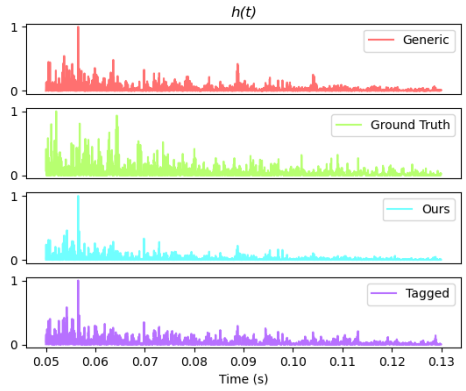
We conduct a perceptual comparison between simulated wavefields at the same positions of source and listener, see Figure 3, using an automated perceptual metric learned on subjects' responses [23]. The metric consists of a 14-layer deep neural network with filters trained on features extracted from paired input audio samples; it expresses a distance $D(x_{ref}, x_{per})$ between two input signals, where x_{ref} is a reference signal, and x_{per} is a perturbated signal. The function D considers factors including reverb and the ratio between direct and reverberated signal. We test whether the metric expresses a closer perceptual distance between the measured ground truth and the synthesised wavefields, by convolving

the RIRs to samples from the evaluation subset of a database for acoustic scene classification, which comprises 15 different acoustic environments, generating a total of 18 minutes of audio over 1620 N samples [25]. The learned metric determines the distance between the ground truth convolutions and each of the relative simulated RIRs: for each audio sample k in N, we determine perceptual distances $D(x_{ref,k}, x_{per_i,k} \; \forall \; i \in \{generic, tagged, ours\})$, where ref and per_i are the measured and simulated RIRs.

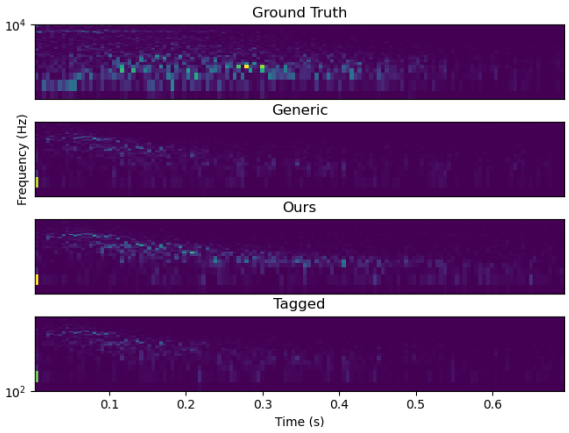
5 RESULTS

5.1 Wave Field Comparison

We compare an artistic material tag, *tagged*, which assigns acoustic geometry in the scene by applying the methodology used by game designers when tagging scenes; a *generic* tag, where all meshes map to a single acoustic material that expresses 0.1α acoustic absorption coefficient across all frequency bands; and finally, *ours*, having meshes mapping to acoustic materials inferred by proposed system. For each acoustic material, the sound renderer assigns its acoustic absorption coefficients to all triangles of the corresponding mesh. The system takes around 7 minutes to compute acoustic materials for the 38 meshes composing the reconstructed environment. Figure 4a shows the magnitude of early reflections for each acoustic simulation. Note that the ground truth shows distortion errors caused by the constant noise floor of $-55dB$ at the time of measurement. *Our* acoustic simulation marks the fastest decay in reverberation energy with a T_{60} of $0.044s$, computed parameterising the RIRs according to [8]. The ground truth is $0.43s$, and generic and tagged are respectively $0.054s$ and 0.073.



(a) Measured impulse responses from measurements shown as normalised direct energy of reflections in function of time $h(t)$.



(b) Spectrograms, computed with triangular windows, showing decay of early reflections across a logarithmic frequency range from $10^2 Hz$ to $10kHz$.

Fig. 4. A comparison between RIRs measured from a real environment and synthesised wavefields from its reconstruction.

5.2 Perceptual Distance Test

The perceptual metric between the measured ground truth and the synthesised wavefields yields three variables describing 1620 pairwise distances between the measured and synthesised RIRs, see Table 1. Given the test conditions, the correlation test uses Pearson r scores to measure the relationship between RIRs with different acoustic geometry.

Table 1. A table showing Pearson r scores for correlation between the simulated wave-fields: *tagged*, *generic* and *ours*.

		Generic	Tagged	Ours
Generic	r	1	0.85	0.854
	Sig. (2-tailed)		0	0
Tagged	r	0.85	1	0.96
	Sig. (2-tailed)	0		0
Ours	r	0.854	**0.96**	1
	Sig. (2-tailed)	0	**0**	

6 DISCUSSION

Experimental results show correlations between wave-fields generated from geometry that has been automatically tagged by our material classification system; geometry tagged with a single generic material, and geometry tagged with human supervision. A preliminary perceptual test shows that tagging the acoustic geometry using a material classifier mapped to acoustic absorption data can produce acoustic simulations that correlate to their manually tagged counterparts. As shown in the example of superpixel prediction in Figure 2b, the classifier can recognise many patches from the input image texture, associating them with labels that visually relate to the material represented, excluding outliers by selecting the most frequent prediction. In the evaluation, the mesh reconstruction process can introduce noise to the resulting texture due to the geometrical approximations caused by the point cloud's triangulation [31]. These approximations cause artefacts, resulting in incorrectly predicted material labels. While the superpixel segmentation process excludes these outliers, it prevents the feature extractor from capturing larger structures in the input texture.

The generic and tagged simulations, having a single acoustic material and approximated absorption data, are likely to subtract less energy from rays, maintaining more energy over late reflections. Hence, the limited number of rays and more specific acoustic absorption assigned to scene geometry determine shorter reverberation levels in our simulation, see Figure 4a. According to the spectral analysis shown in Figure 4b, this also causes our simulation to preserve more harmonic details above $10^3 Hz$. Overall the experimental evaluation shows that acoustic materials' assignment through our system yields results that correlate to artistic material tags. Consequently, the proposed system can tag acoustic materials in unseen complex scenes, reducing the costs of applications of wave-based methods in VEs. However, limitations in the test design, such as the single measurement point and the limited set of materials in the real environment, raise the need for an extended evaluation including different scene scales, i.e. larger rooms or outdoor environments having more sustained late reflections in their RIRs. Besides, task-based subjective experiments in VEs with subjects would further define the perceptual impact of acoustic material tagging on acoustic simulations.

7 CONCLUSIONS AND FUTURE DIRECTIONS

We addressed the problem of mapping acoustic absorption data to scene geometry in virtual environments for acoustic simulations. Methods and frameworks for material recognition have become efficient enough in recognising materials in the wild with varying factors of illumination, shape and surface characteristics. Despite their limited resolution in computer games applications, sound propagation systems benefit from acoustic material tagging. With the proposed system, we aim to integrate material recognition to wave-based methods to determine materials' acoustic properties. The next steps of this work aims to extend the system to broader scenes with larger sets of materials as well as improving the material recognition by performing multi-scale analysis of superpixels and adopting clustering paradigms to overcome the limitations of finite material space definitions.

REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. 2012. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 11 (2012), 2274–2282. https://doi.org/10.1109/TPAMI.2012.120

[2] Steam Audio. 2020. Git Repository Steam Audio. *URL: https://valvesoftware.github.io/steamaudio/downloads. html,[accessed 2019, February 27]* (2020).

[3] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. 2013. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on graphics (TOG)* 32, 4 (2013), 1–17.

[4] Umberto Berardi, Gino Iannace, and Carmine Ianniello. 2016. Acoustic intervention in a cultural heritage: The chapel of the Royal Palace in Caserta, Italy. *Buildings* 6, 1 (2016), 1.

[5] Nicolas Bonneel, Clara Suied, Isabelle Viaud-Delmon, and George Drettakis. 2010. Bimodal perception of audio-visual material properties for virtual environments. *ACM Transactions on Applied Perception (TAP)* 7, 1 (2010), 1–16.

[6] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing* 27, 1 (2017), 206–219.

[7] Mattia Colombo, Alan Dolhasz, and Carlo Harvey. 2020. A Computer Vision Inspired Automatic Acoustic Material Tagging System for Virtual Environments. In *2020 IEEE Conference on Games (CoG)*. IEEE, 736–739.

[8] Amaro A de Lima, Thiago de M Prego, Sergio L Netto, Bowon Lee, Amir Said, Ronald W Schafer, Ton Kalker, and Majid Fozunbal. 2009. Feature analysis for quality assessment of reverberated speech. In *2009 IEEE International Workshop on Multimedia Signal Processing*. IEEE, 1–5.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[10] Lijun Ding and Ardeshir Goshtasby. 2001. On the Canny edge detector. *Pattern Recognition* 34, 3 (2001), 721–725.

[11] Alan Dolhasz, Carlo Harvey, and Ian Williams. 2020. Learning to Observe: Approximating Human Perceptual Thresholds for Detection of Suprathreshold Image Transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[12] Utkarsh Gaur and BS Manjunath. 2019. Superpixel Embedding Network. *IEEE Transactions on Image Processing* 29 (2019), 3199–3212.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Martin Holters, Tobias Corbach, and Udo Zölzer. 2009. Impulse response measurement techniques and their applicability in the real world. In *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*. 1–5.

[15] Vedad Hulusic, Carlo Harvey, Kurt Debattista, Nicolas Tsingos, Steve Walker, David Howard, and Alan Chalmers. 2012. Acoustic rendering and auditory–visual cross-modal perception and interaction. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 102–131.

[16] Hansung Kim, Luca Remaggi, Sam Fowler, Philip Jackson, and Adrian Hilton. 2020. Acoustic Room Modelling using 360 Stereo Cameras. *IEEE Transactions on Multimedia* (2020).

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Manuel Lagunas, Sandra Malpica, Ana Serrano, Elena Garces, Diego Gutierrez, and Belen Masia. 2019. A similarity measure for material appearance. *arXiv preprint arXiv:1905.01562* (2019).

[19] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. 2018. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–12.

[20] Shiguang Liu and Dinesh Manocha. 2020. Sound Synthesis, Propagation, and Rendering: A Survey. *arXiv preprint arXiv:2011.05538* (2020).

[21] T. Lokki and M. Grohn. 2005. Navigation with auditory cues in a virtual environment. *IEEE MultiMedia* 12, 2 (2005), 80–86. https://doi.org/10.1109/MMUL.2005.33

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[23] Pranay Manocha, Adam Finkelstein, Zeyu Jin, Nicholas J Bryan, Richard Zhang, and Gautham J Mysore. 2020. A differentiable perceptual audio metric learned from just noticeable differences. *arXiv preprint arXiv:2001.04460* (2020).

[24] Ravish Mehra, Atul Rungta, Abhinav Golas, Ming Lin, and Dinesh Manocha. 2015. Wave: Interactive wave-based sound propagation for virtual environments. *IEEE transactions on visualization and computer graphics* 21, 4 (2015), 434–442.

[25] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. TUT database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 1128–1132.

[26] S. Poeschl, K. Wall, and N. Doering. 2013. Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence. In *2013 IEEE Virtual Reality (VR)*. 129–130. https://doi.org/10.1109/VR.2013.6549396

[27] Nikunj Raghuvanshi and John Snyder. 2014. Parametric wave field coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 1–11.

[28] Jose Luis Rubio-Tamayo, Manuel Gertrudix Barrio, and Francisco García García. 2017. Immersive environments and virtual reality: Systematic review and advances in communication, interaction and simulation. *Multimodal Technologies and Interaction* 1, 4 (2017), 21.

[29] Gabriel Schwartz and Ko Nishino. 2019. Recognizing material properties from images. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[30] Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau. 2002. Comparison of different impulse response measurement techniques. *Journal of the Audio Engineering Society* 50, 4 (2002), 249–262.

[31] Eric Turner, Peter Cheng, and Avideh Zakhor. 2014. Fast, automated, scalable generation of textured 3d models of indoor environments. *IEEE Journal of Selected Topics in Signal Processing* 9, 3 (2014), 409–421.

[32] Michael Vorländer, Dirk Schröder, Sönke Pelzer, and Frank Wefers. 2015. Virtual reality for architectural acoustics. *Journal of Building Performance Simulation* 8, 1 (2015), 15–25.