CHAPTER 6: AUDIO-VISUAL OBJECT RECONSTRUCTION FROM VIDEO[1]

This chapter describes a multimodal single and multi-frame neural network for 3D reconstructions using audio-visual inputs. Our trained reconstruction LSTM autoencoder 3D-MOV accepts multiple inputs to account for a variety of surface types and views. To the best of our knowledge, our single and multi-frame model is the first audio-visual reconstruction neural network for 3D geometry and material representation.

6.1 Introduction

Deep neural networks trained on single- or multi-view images have enabled 3D reconstruction of objects and scenes using RGB and RGBD approaches. These models generate 3D geometry volumetrically (Boscaini et al., 2016; Choy et al., 2016; Xie et al., 2018a) and in the form of point clouds (Han et al., 2019; Qi et al., 2016a, 2017a). With these reconstructions, additional networks have been developed to use the 3D geometry as inputs for object detection, classification, and segmentation in 3D environments (Atzmon et al., 2018; Qi et al., 2017b). However, existing methods still encounter a few challenging scenarios for 3D shape reconstruction (Boscaini et al., 2016).

One such challenge is occlusion in cluttered environments with multiple objects in a scene. Another is spatial resolution. Volumetric methods such as voxelized reconstructions (Maturana and Scherer, 2015) are primarily limited by resolution. Point cloud representations of shape avoid issues of grid resolution, but instead need to cope with issues of point set size and approximations. Existing methods also are challenged by transparent and highly reflective or textured surfaces. Self-occlusions and occlusions from other objects can also hinder image-based networks, necessitating the possible adoption of multimodal neural networks.

To address these limitations, we propose to use audio-visual input for 3D shape and material reconstruction. A single view of an object is insufficient for 3D reconstruction as only one projection of the object can be seen, while multi-view input does not intrinsically model the spatial relationships between

[1] This chapter is currently under review.

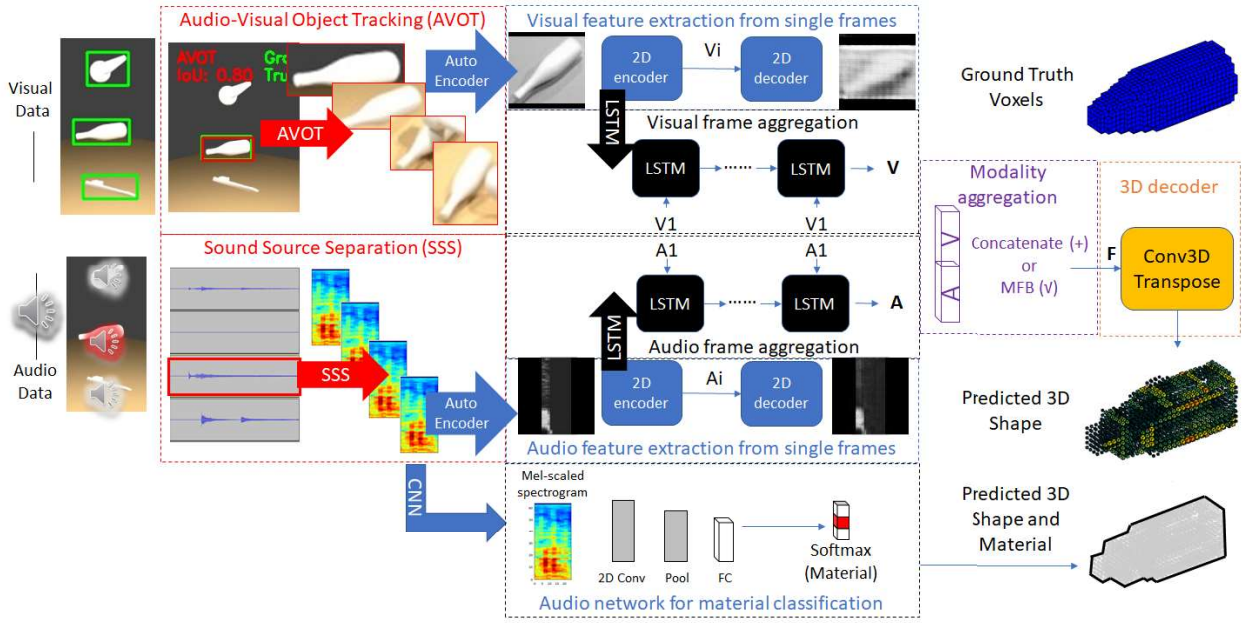

Figure 6.1: Our 3D-MOV neural network is a multimodal LSTM autoencoder optimized for 3D reconstructions of single ShapeNet objects and multiple objects from Sound20K video. During training, a LSTM autoencoder is trained to reconstruct 2D image and spectrogram inputs. 3D shape reconstructions are then generated by fine tuning the fused encodings of each modality for 3D voxel output. The network has recurrent LSTM layers for temporal consistency. Adding audio enhances learning for object tracking, material classification, and reconstruction when multiple objects collide, self-occlude, or are transparent.

views. By providing a temporal sequence of video frames, we strengthen the relationships between views, aiding reconstruction. We also include audio as an input, in particular, *impact sounds* resulting from interactions between the object to be reconstructed and the surrounding environment. Impact sounds provide information about the material and internal structure of an object, providing complementary cues to the object's visual appearance. We choose to represent our final 3D shape using voxel representation due to their state-of-the-art performance in classification tasks. To the best of our knowledge, our audio-visual network is the first to reconstruct multiple 3D objects from a single video.

Main Results: In this paper, we introduce a new method to reconstruct high-quality 3D objects from a sequence of images and sound, the main contributions of this work can be summarized as follows.

- A multimodal LSTM autoencoder neural network for both geometry and *material* reconstruction from audio and visual data is introduced;
- The resulting implementation has been tested on voxel, audio, and image datasets of objects over a range of different geometries and materials;

- Experimental results of our approach demonstrate the reconstruction of single sounding objects and multiple colliding objects in a virtual scene;
- Audio-augmented datasets with ground truth object tracking bounding boxes are available for future research.

6.2 Related Work

Computer vision research continues to push state-of-the-art reconstruction and segmentation of objects in a scene (Dai et al., 2017c). However, there still remain research opportunities in 3D reconstruction. Wide baselines limit the accuracy of feature correspondences between views. Challenging objects for reconstruction include thin or small objects (e.g. table legs), and classes of objects that are transparent, occluded, or have much higher shape variation than other classes (e.g. lamps, benches, and tables compared to cabinets, cars, and speakers for example). In this section, we review previous work relating to 3D reconstruction, multimodal neural networks, and reconstruction network structures.

6.2.1 3D Reconstruction

Deep learning techniques have produced state-of-the-art 3D scene and object reconstructions. These models take an image or series of images and generate a reconstructed output shape. Some methods produce a transformed image of the input, intrinsically representing the 3D object structure (Odena et al., 2017; Tsai, 2018; Mao et al., 2017; Mirza and Osindero, 2014; Lun et al., 2017). 3D voxel grids provide a shape representation which is easy to visualize and works well with convolution operations (Choy et al., 2016; Girdhar et al., 2016; Riegler et al., 2017; Qi et al., 2016b; Hu et al., 2018; Wu et al., 2016). In more recent work, point clouds have also been found to be a viable shape representation for reconstructed objects (Hedman et al., 2017; Fan et al., 2017).

6.2.2 Multimodal Neural Networks

Neural networks with multiple modalities of inputs help cover a broader range of experimental setups and environments. Common examples include visual question answering (Cadene et al., 2019), vision and touch (Lee et al., 2019), and other multisensory interactions (Klemen and Chambers, 2012). Multiple

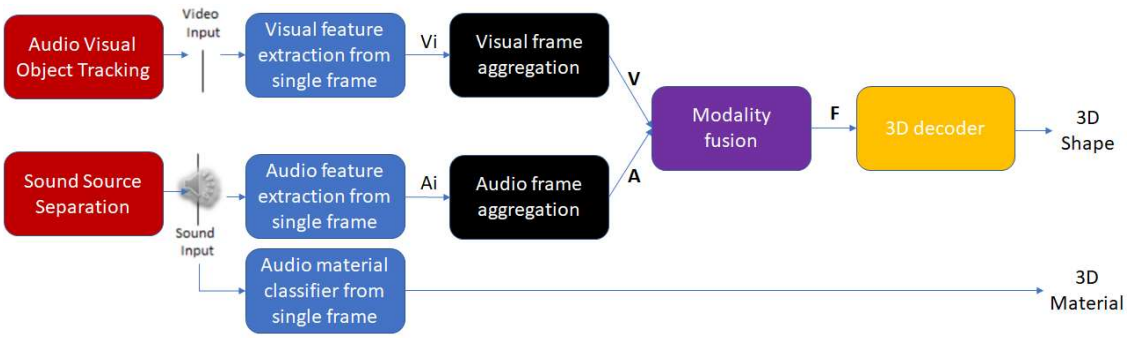

Figure 6.2: We first separate audio-visual data using object tracking (Section 6.3.1) and sound source separation (Section 6.3.2). Features from audio and visual subnetworks for each object are aggregated by LSTM autoencoders and then fused using addition, concatenation, or a bilinear model (Yu et al., 2014). Finally, 3D geometry is reconstructed by a 3D decoder and audio classified material applied to all voxels.

modes may also take the form of image-to-image translation, e.g. domain transfer (Huang et al., 2018). Using local and global cropped parts of the images (i.e. bounding boxes) have also been shown to serve as a mode of context to supervise learning (Reed et al., 2016).

Audio-visual specific multimodal neural networks have also proven effective for speech separation (Ephrat et al., 2018b) as well as sound localization (Zhao et al., 2018; Owens and Efros, 2018; Konno et al., 2020; Arandjelović and Zisserman, 2017). Audio synthesis conditioned on images is also enabled as a result of these combined audio-visual datasets (Zhou et al., 2018). Please see a survey and taxonomy on multimodal machine learning (Baltrusaitis et al., 2017) and multimodal deep learning (Ngiam et al., 2011a) for more information.

6.2.3 Reconstruction Network Structures

While single view networks perform relatively well for most object classes, objects with concave structures or classes of objects with large variations in shape tend to require more views. 3D-R2N2 (Choy et al., 2016) allows for both single and multi-view implementations given a single network. Other recurrent models include learning in video sequences (Chong and Tay, 2017; Hasan et al., 2016), Point Set Generation (Fan et al., 2017), and Pixel Recurrent Neural Network (PixelRNN) (van den Oord et al., 2016c). Methods have also been developed to ensure temporal consistency (Xie et al., 2018b) and use generative techniques (Gwak et al., 2017). T-L network (Girdhar et al., 2016) and 3D-R2N2 (Choy et al., 2016) are most similar to our 3D-MOV reconstruction neural network. Building on these related works, we fuse audio as an additional input and temporal consistency in the form of LSTM layers (Fig. 6.2).

6.3 Technical Approach

In this work, we reconstruct the 3D shape and material of sounding objects given images and impact sounds. Using audio and visual information, we present a method for reconstruction of single instance ModelNet objects augmented with audio and multiple objects colliding in a Sound20K scene from video. In this section, we cover visual representations from object tracking (Section 6.3.1) and audio obtained from sound source separation of impact sounds (Section 6.3.2) that serve as inputs into our 3D-MOV reconstruction network (Section 6.4).

6.3.1 Object Tracking and Visual Representation

Since an entire video frame may contain too much background, we use object tracking to track and segment different objects. This tracking is performed using the Audio-Visual Object Tracker (AVOT) (Wilson and Lin, 2020). Similar to the Single Shot MultiBox Detector (SSD) (Liu et al., 2015), AVOT is a feed-forward convolutional neural network that classifies and scales a fixed number of anchor bounding boxes to track objects in a video. While 3D-MOV aggregates audio-visual features before decoding, AVOT fuses audio-visual inputs before its base network. With additional information from audio, AVOT defines an object based on both its geometry and material.

We use AVOT over other algorithms, such as YOLO (Redmon et al., 2015b) or Faster R-CNN (Ren et al., 2015b), because of the availability of audio and need for higher object-tracking accuracy given occlusions caused by multiple objects colliding. Unlike CSR-DCF (Lukezic et al., 2016), AVOT automatically detects objects in the video without initial markup of bounding boxes. For future work, a scheduler network or a combination of object trackers is worth considering as well as use of Common Objects in Context (COCO) (Lin et al., 2014a) and SUN RGB-D (Song et al., 2015; Silberman et al., 2012a; Janoch et al., 2011; Xiao et al., 2013) datasets for initialization and transfer learning.

The output from tracking is a series of segmented image frames for each object, consisting of the contents of its tracked bounding box throughout the video. These segmented frames are grayscaled and resized to a consistent input size of 88 by 88 pixels. While resizing, we maintain aspect ratio and pad to square the image. These dimensions were automatically chosen to account for the size of objects in our Sound20K dataset and to capture their semantic information. Scenes included one, two, and three colliding objects with materials such as granite, slate, oak, and marble. For our single-frame, single impact sound

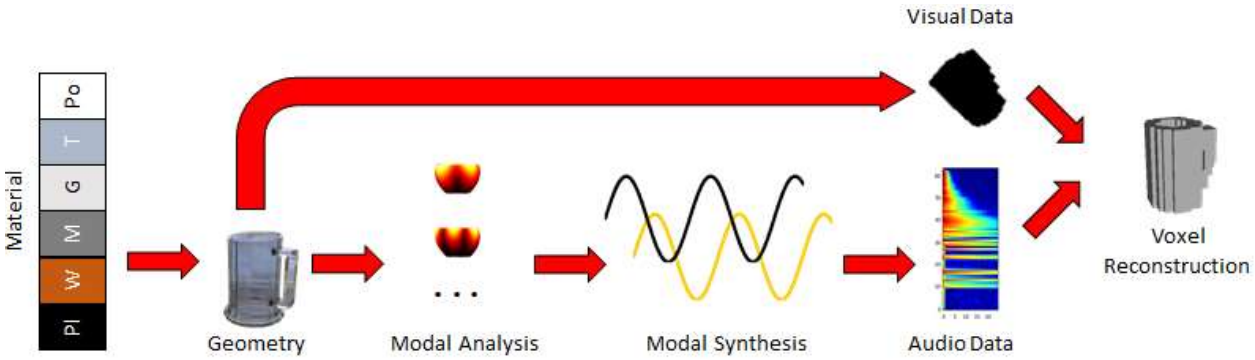

Figure 6.3: For our single impact sound analysis using ShapeNet, we build multimodal datasets using modal sound synthesis to produce spectrograms for audio input and images of voxelized objects as an estimate of shape. Please note that audio used from ISNN (Sterling et al., 2018) was generated for voxelized models as a result of the sound synthesis pipeline requiring watertight meshes. Unmixed Sound20K audio was available from the generated synthetic videos.

evaluations, we resized ShapeNet's 224 x 224 image size. For comparison, other image sizes from related work include MNIST, 28 x 28; 3D-R2N2, 127 x 127; ImageNet, 256 x 256.

6.3.2 Sound Source Separation of Impact Sounds and Audio Representation

For single frame reconstruction, we synthesize impact sounds on ShapeNet (Sterling et al., 2018), illustrated in Fig. 6.3. For multiple frames, we take as input a Sound20K video showing one or more objects moving around a scene. These objects strike one another or the environment, producing impact sounds, which can be heard in the audio track of the video. We refer to these objects, dynamically moving through the scene and generating sound due to impact and collision, as *sounding objects*. Sound20K provides mixed and unmixed audio which can be used directly or to train algorithms for sound source separation (Wang et al., 2014; Koretzky et al., 2017; Scallie et al., 2017). While prior work to localize objects using audio-visual data exists (Arandjelović and Zisserman, 2017; Zhao et al., 2018), automatically associating separated sounds with corresponding visual object tracks in the context of the reconstruction task remains an area of future work.

Initially, Sound20K and ShapeNet audio are available as time series data, sampled at 44.1 kHz to cover the full audible range. The audio is converted to mel-scaled spectrograms for neural network inputs, which effectively represent the spectral distribution of energy over time. Each spectrogram is 3 seconds for a single frame (ShapeNet) and 0.03 seconds per multi-frame (Sound20K) with an overlap of 25%. Audio spectrograms are aligned temporally with their corresponding image frames from video, forming the audio-

visual input for queries. They are generated with discrete short-time Fourier transforms (STFTs) using a Hann window function.

$$\chi(m,k) = \sum_{n=0}^{N-1} x(n+mH)w(n)exp(-2\pi ikn/N) \tag{6.1}$$

for m^{th} time frame and k^{th} Fourier coefficient with real-valued DT signal $x : \mathbb{Z} \rightarrow \mathbb{R}$, sampled window function $w(n)$ for n $\in [0 : N-1] \rightarrow \mathbb{R}$ of length $N \in \mathbb{N}$, and hop size $H \in \mathbb{N}$ (Mller, 2015).

6.3.2.1 Single View, Single Impact Sound

Single-view inputs are based on ShapeNet, a repository of 3D CAD models based on WordNet categories. Evaluations were performed on voxelized versions of ShapeNet's (Chang et al., 2015), ModelNet10 and ModelNet40 models (Wu et al., 2015b), and image views of these datasets from 3D-R2N2 (Choy et al., 2016). To generate audio for these objects to be used for our multi-modal 3D-MOV neural network, we use data from Impact Sound Neural Network (Sterling et al., 2018). This work synthesized impact sounds for voxelized ModelNet10 and ModelNet40 models (Wu et al., 2015b) using modal analysis and sound synthesis. Modal analysis is precomputed to obtain *modes* of vibration for each object and sound synthesized with an amplitude determined at run-time given the hit point location on the object and impulse force. The modes are represented as damped sinusoidal waves where each mode has the form

$$q_{\mathrm{i}} = a_{\mathrm{i}} e^{-\mathrm{d_i t}} sin(2\pi f_{\mathrm{i}} t + \theta_{\mathrm{i}}), \tag{6.2}$$

where f_{i} is the frequency of the mode, d_{i} is the damping coefficient, a_{i} is the excited amplitude, and θ_{i} is the initial phase.

6.3.2.2 Multi-Frame, Multi-Impact

Multi-frame inputs to our system consist of Sound20K (Zhang et al., 2017d) videos that may contain multiple sounding objects, possibly of similar sizes, shapes, and/or materials. This synthetic video dataset contains audio and video data for multiple objects colliding in a scene. Sound20K consists of 20,378 videos generated by rigid-body simulation and impact sound synthesis pipeline (Doug L. James and Pai,

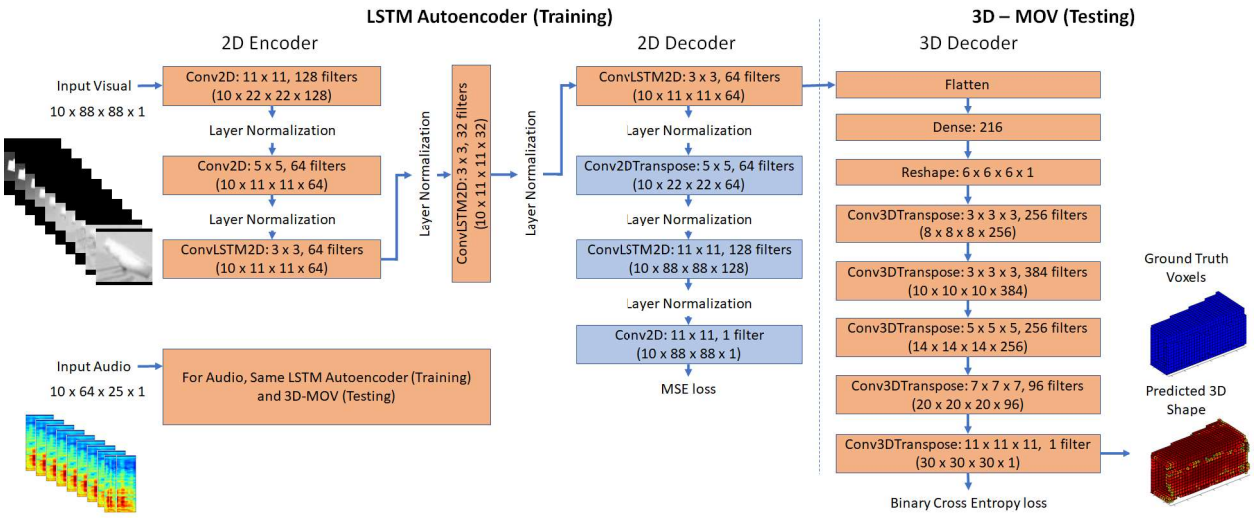

Figure 6.4: We separately train audio and visual autoencoders to learn encodings and fine-tune for our 3D reconstruction task. We replace the 2D decoder by a five deconvolutional layer 3D decoder to generate a 30^3 voxel grid. The separate audio-visual LSTM autoencoders are flattened and merged to form the dense layer. Here, the predicted 3D shape voxels are displayed based on a threshold of 0.3.

2006). Visually, Sound20K (Zhang et al., 2017d) objects can be separated from one another through tracking of bounding boxes. However, audio source separation can be more challenging, particularly for unknown objects. While Sound20K provides separate audio files for each object that can be used, the audio data can also be used to train sound source separation techniques (Wang et al., 2014; Koretzky et al., 2017; Scallie et al., 2017) to learn to unmix audio to individual objects by geometry and material. As future work, we will compare the impact on reconstruction quality and performance if we were to use combined, unmixed audio for each object. We will also compare impact of using source separated sounds versus ground truth unmixed audio.

6.4 3D-MOV Network Structure

Our 3D-MOV network is a multi-modal LSTM autoencoder optimized for 3D reconstructions of multiple objects from video. Like 3D-R2N2 (Choy et al., 2016), it is recurrent and generates a 3D voxel representation. However, to the best of our knowledge, our 3D-MOV network is the first audio-visual reconstruction network for 3D object reconstruction. After object tracking and sound source separation, we separately train autoencoders to extract visual and audio feature from each frame (Section 6.4.1). While the 2D encoder weights are reused, the 2D decoders are discarded (blue rectangles in Fig. 6.4) and re-

placed with 3D decoders for learning to reconstruct voxel outputs of the tracked objects based on given 2D images and spectrograms. Using a merge layer such as addition, concatenation, or a bilinear model (Yu et al., 2014), our method 3D-MOV fuses the results of the audio and visual subnetworks comprised of LSTM autoencoders.

6.4.1 Single Frame Feature Extraction

The autoencoder consists of two convolutional layers for spatial encoding followed by a LSTM convolutional layer for temporal encoding. As a general rule of thumb, we use small filters (3x3 and at most 5x5), except for the very first convolutional layer connected to the input, and strides of four and two for the two conv layers (Li et al., 2020). The decoder mirrors the encoder to reconstruct the image (Fig. 6.5). After each convolutional layer, we employ layer normalization, which is equivalent to batch normalization for recurrent networks (Ba et al., 2016). It normalizes the inputs across features and is defined as:

$$\mu_j = \frac{1}{m}\sum_{j=1}^{m} x_{ij}; \sigma_j^2 = \frac{1}{m}\sum_{j=1}^{m}(x_{ij} - \mu_j)^2; \hat{x}_{ij} = \frac{x_{ij} - \mu_j}{\sqrt{\sigma_j^2 + \epsilon}} \tag{6.3}$$

where x_{ij} is batch i, feature j of the input x across m features.

6.4.2 Frame Aggregation

In chronological order, the training video frames make a temporal sequence. LSTM convolutional layers are used to preserve content and spatial information. To generate more training sequences, we perform data augmentation by concatenating frames with strides 1, 2, and 3. For example, we use a skipping stride of 2 to generate a sequence of every other frame. We use a 10-frame sequence size as a sliding window technique for aggregation of the encodings. The encoder weights learned here are used to then learn 3D decoder weights to output a 3D voxel reconstruction based on audio-visual inputs from audio-augmented ModelNet with impact sound synthesis and Sound20K video.

6.4.3 Modality Fusion and 3D Decoder

After encoding our inputs with LSTM convolutional layers, we flatten to a fully connected layer for each audio and visual subnetwork. These dense layers are fused together prior to multiple Conv3D trans-

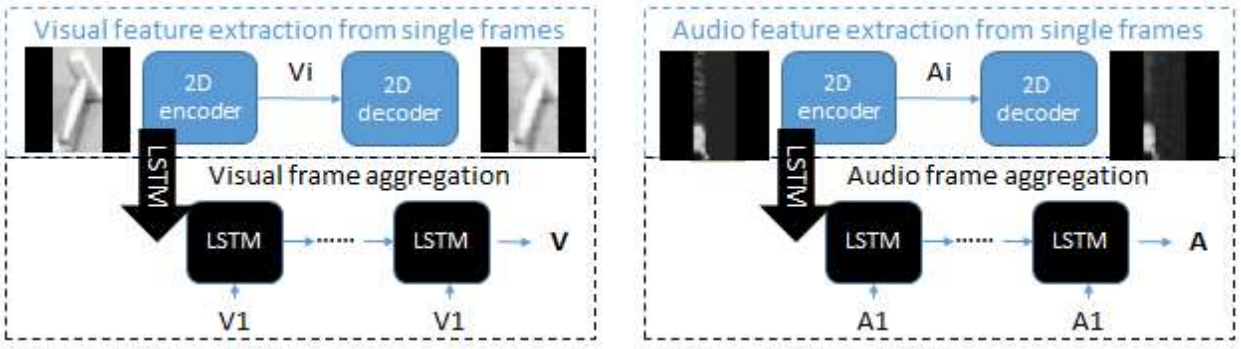

Figure 6.5: Hidden layer representations V_i and A_i are trained to spatially encode object geometry and impact sounds, where i is each video frame. These learned weights are subsequently used during test time to generate 3D shapes from audio-visual inputs. For sequence modeling, LSTM layers are reliable for temporal consistency and establishing dependencies over time. More specifically, we use convolutional LSTM layers rather than fully connected to also preserve spatial information.

pose layers for the 3D decoder. Prior work in multimodal deep learning, such as visual question and answering systems, have merged modalities for classification tasks using addition and MFB (Yu et al., 2014). A 3D decoder accepts the fusion of audio-visual LSTM encodings and maps it to a voxel grid with five deconvolutional layers, similar to T-L Network (Girdhar et al., 2016). Unlike T-L's 20^3 voxel grid, we use 30^3 voxels for greater resolution and apply a single, audio-based material classification to all voxels. Deconvolution, also known as fractionally-strided or transposed convolution, results in a 3D voxel shape by broadcasting input X through kernel K (Zhang et al., 2020).

$$\sum_{i=0}^{h}\sum_{j=0}^{w} Y[i:i+h, j:j+w] \mathrel{+}= (X[i,j] * K) \tag{6.4}$$

6.5 Results

In this section, we present our implementation, training, and evaluation metrics along with 3D-MOV reconstructed objects (Fig. 6.8). Please see a comparative analysis of loss and accuracy against baseline methods by dataset and number of views. For each of the datasets ShapeNet and Sound20K, we evaluate the network architecture described in Section 6.4 against audio, visual, and audio-visual methods using binary cross entropy loss and intersection over union (IoU) reconstruction accuracy.

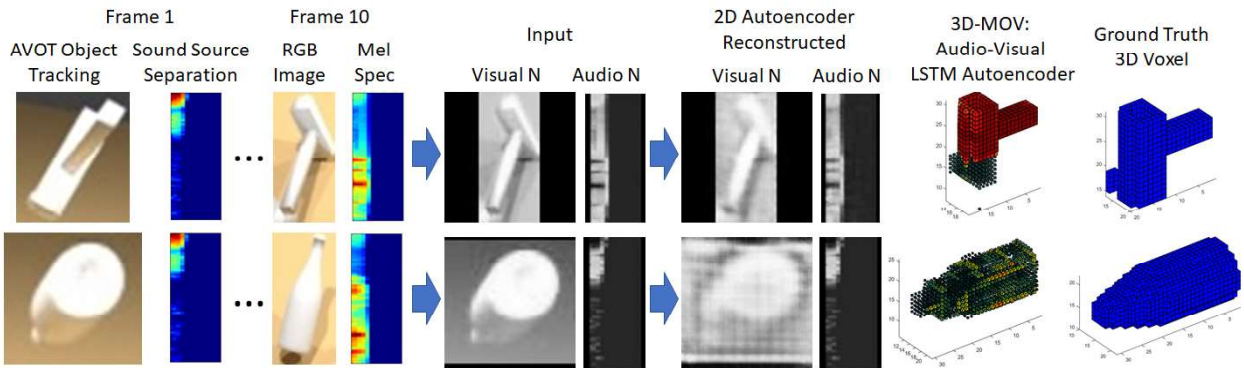

Figure 6.6: Reconstructed objects from using multiple frames and impact sounds. Please see our supplementary materials for a complete review of results for ShapeNet and Sound20K datasets using binary cross entropy loss and reconstruction accuracy comparing audio, visual, and audio-visual methods by number of views. Our method is able to obtain better reconstruction results for concave internal structures and scenes with multiple objects by fusing temporal audio-visual inputs.

6.5.1 Implementation

Our framework was implemented using Tensorflow (Abadi et al., 2015) and Keras (Chollet et al., 2015). Training was run on Ubuntu 16.04.6 LTS with a single Titan X GPU. Voxel representations were rendered based on Matlab visualization code from 3D-GAN (Wu et al., 2016). From Sound20K videos, images were grayscale with dimensions 84 x 84 x 1 and audio spectrograms were 64 x 25 x 1, zero padded to equivalent dimensions. Visual data was augmented with resizing, cropping, and skipping strides.

6.5.2 Training

Since joint optimization can be difficult to perform, we train our reconstruction autoencoder and fused audio-visual networks separately and then jointly optimize to fine-tune the final network. Mean square error is used for the 2D reconstruction loss to train the encoder to reconstruct input images and audio spectrograms. Binary cross entropy loss is calculated between ground truth and reconstructed 3D voxel grids. During testing, we reconstruct from encoded vector representation of audio-visual inputs to a 3D voxel reconstruction output.

Previous work has used symmetry induced volume refinement to constrain and finalize GAN volumetric outputs (Niu et al., 2018). Other methods have used multiple views to continuously refine the output (Choy et al., 2016). Furthermore, most adversarial generating methods create examples by perturbing

existing data, limiting the solution space. Our approach constrains the space of possible 3D reconstructions for objects in the scene by temporal consistency, aggregation, and fusion of audio and visual inputs.

6.5.3 Evaluation metrics

Methods were evaluated against voxel Intersection-over-Union (IoU), also known as the Jaccard index (Jaccard, 1901), between the 3D reconstruction and ground truth voxels as well as cross-entropy loss. This can be represented as area of overlap divided by the area of union. More formally:

$$IoU = \frac{\sum_{i,j,k}[I(p_{(i,j,k)} > t)I(y_{(i,j,k)}]}{\sum_{i,j,k}[I(I(p_{(i,j,k)} > t) + I(y_{(i,j,k)}))]} \tag{6.5}$$

where $y_{i,j,k} \in 0, 1$ is the ground truth occupancy, $p_{i,j,k}$ the Bernoulli distribution output at each voxel, $I(\cdot)$ an indicator function, and t for threshold. Higher IoU means better reconstruction quality.

Please see Table 6.1 for results by dataset against baseline methods: Figure 6.7 for example ModelNet10 reconstructions and Figure 6.8 for exemplary Sound20K reconstructions, and Figure 6.9 for training loss for audio, visual, and audio-visual.

6.5.4 Experimental Results

Average runtime for 3D-MOV audio-visual training of the Sound20K dataset was about 1.5 minutes per epoch for a sequence size of 10 and strides of 1, 2, and 3. Using 20 epochs, average training time was 30 minutes. For sequence size of 5, 10k training examples took about 4 minutes per. For sequence size of 1, 50k training examples completed in roughly 10 minutes per epoch. Methods were evaluated against voxel Intersection-over-Union (IoU), also known as the Jaccard index (Jaccard, 1901), between the 3D reconstruction and ground truth voxels as well as cross-entropy loss.

6.5.5 Datasets

We use a server with Ubuntu 16.04.6 LTS and a single Titan X GPU. Training and hold-out test splits were 80% and 20% respectively. With Sound20K sequence size of 10 and strides 1-3, we have 9,800 train-

Table 6.1: 3D-MOV was evaluated against baselines for loss (mean square error and binary cross entropy) and reconstruction accuracy (intersection over union). A view consists of both an image and audio frame. Decreases in 3D-MOV accuracy as ShapeNet views increase requires further investigation but may suggest impact sounds of different hit points are needed rather than using the same sound across views. *We use the T-L Network (Girdhar et al., 2016) fused with audio as an overall baseline comparator with 0.67 loss and 18.0% IoU for an instance of the MN10 chair class. ** Reported in (Choy et al., 2016)

Dataset		ShapeNet (Chang et al., 2015)		Sound20K (Zhang et al., 2017d)
Method	Input	1 view	5 views	10 views
3D-MOV-A (Ours)	A	21.2%	N/A	37.15%
3D-R2N2 (Choy et al., 2016)	V	56.0%**	63.1%**	N/A
3D-MOV-V (Ours)	V	22.7%	22.5%	65.7%
T-L Network (Girdhar et al., 2016)	AV	18.0%*	N/A	N/A
3D-MOV-AV (Ours)	AV	32.6%	31.0%	69.8%

ing and 1,960 test examples of RGB image and audio mel-scaled spectrograms. For ModelNet10, we used voxelized objects since the sound synthesis pipeline requires watertight meshes. Downloadable versions of the datasets used can be found for audio-visual Sound20K synthetic videos (Zhang et al., 2017d), ModelNet10 (Chang et al., 2015), and voxelized ModelNet10 with impact sounds (Sterling et al., 2018). Future work is to explore the impact of increases in dataset size on performance. Pre-processing of audio involved converting sound files to mel-scaled spectrograms. Each spectrogram is 3 seconds for a single frame (ShapeNet) and 0.03 seconds per multi-frame (Sound20K). For visual data, segmented frames are grayscaled and resized to a consistent input size of 88 by 88 pixels. While resizing, aspect ratio was maintained and padded to square the image. These dimensions were chosen to account for the size of objects in our Sound20K and ModelNet10 datasets to capture their semantic information.

6.6 Conclusions

To the best of our knowledge, this is the first method to use audio and visual inputs from ShapeNet objects and Sound20K video of multiple objects in a scene to generate 3D object reconstructions with material. While multi-view approaches can improve reconstruction accuracy, transparent objects, interior concave structures, self-occlusions, and multiple objects remain a challenge. As objects collide, audio provides a complementary sensory input that can enhance the reconstruction model to improve results. In this paper, we demonstrate that augmenting image encodings with corresponding impact sounds refine reconstructions of multimodal LSTM autoencoder neural network outputs.

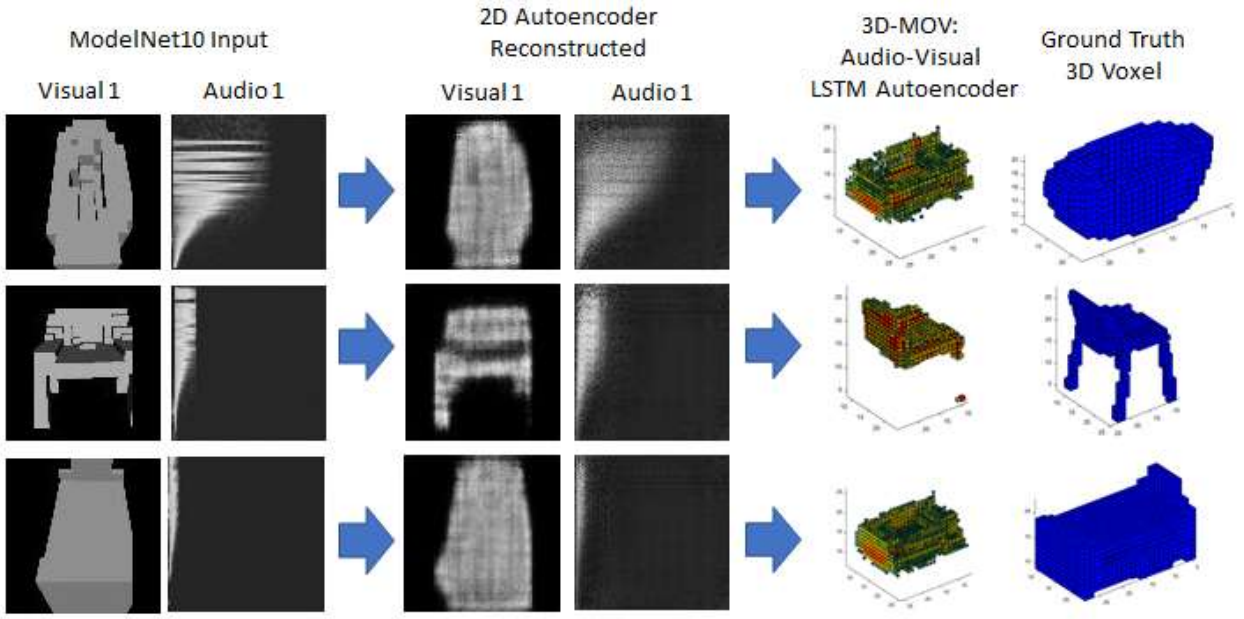

Figure 6.7: 3D-MOV-AV reconstructed image and audio inputs for single view voxelized ModelNet10 classes (top, bathtub; middle, chair; bottom, bed). These results are using a single image and single impact sound, fusing the two modalities with an addition merge layer, training for 60 epochs on a single GPU, and using a voxel threshold of 0.4. 3D-MOV-AV performs the best on ModelNet10 single views, showing audio augmenting visual data.

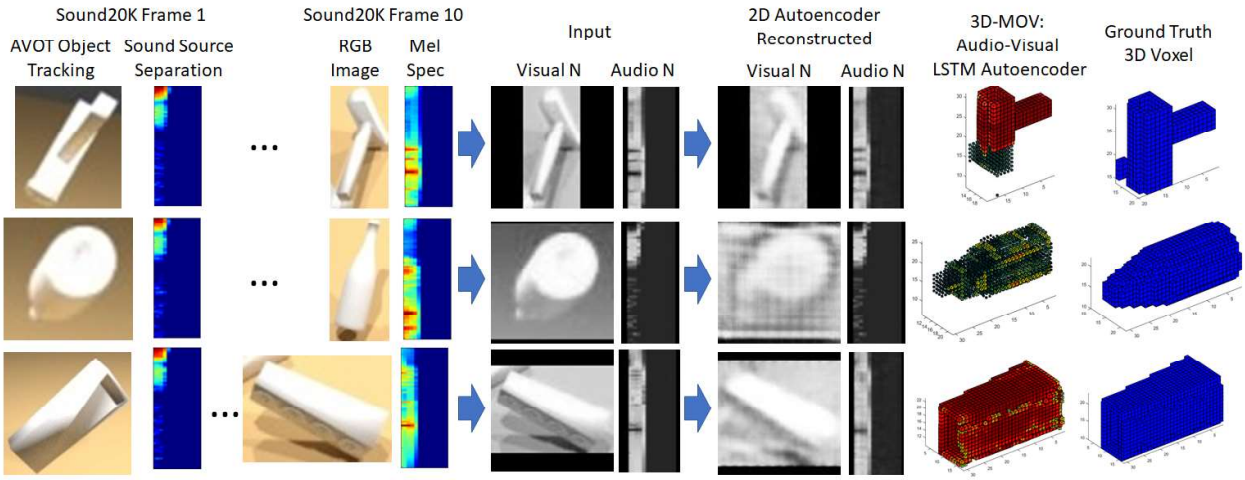

Figure 6.8: Reconstructed objects from using multiple frames and impact sounds. Please see Table 6.1 for results from ShapeNet and Sound20K datasets using binary cross entropy loss and reconstruction accuracy comparing audio, visual, and audio-visual methods by number of views. Audio-visual 3D-MOV-AV performs the best for Sound20K sequence size of 10 views.

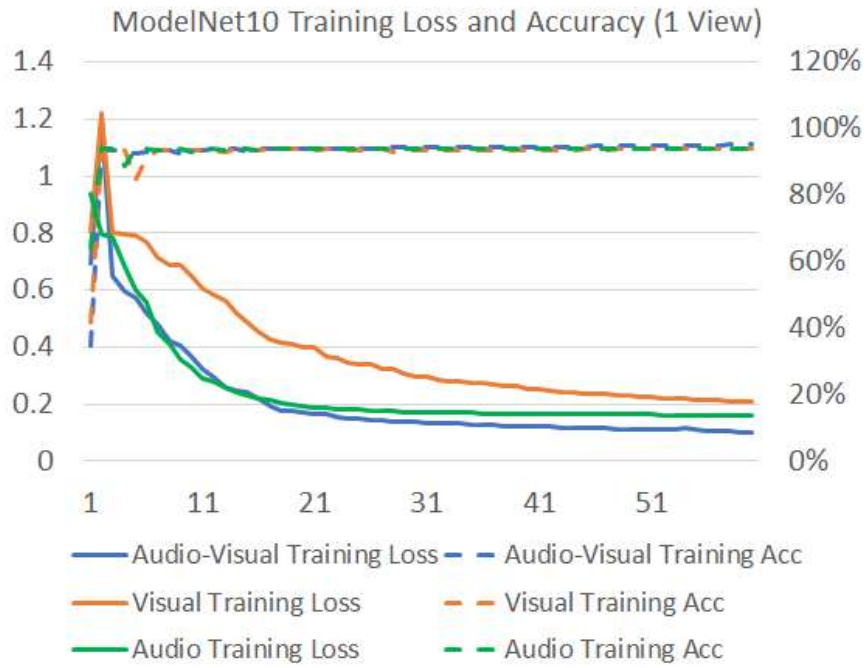

Figure 6.9: Training loss and accuracy for ModelNet10 dataset of 60 epochs. 3D-MOV-AV concludes training with the best performance in terms of binary cross entropy loss and accuracy based on ModelNet10 single views, showing audio augmenting visual data.

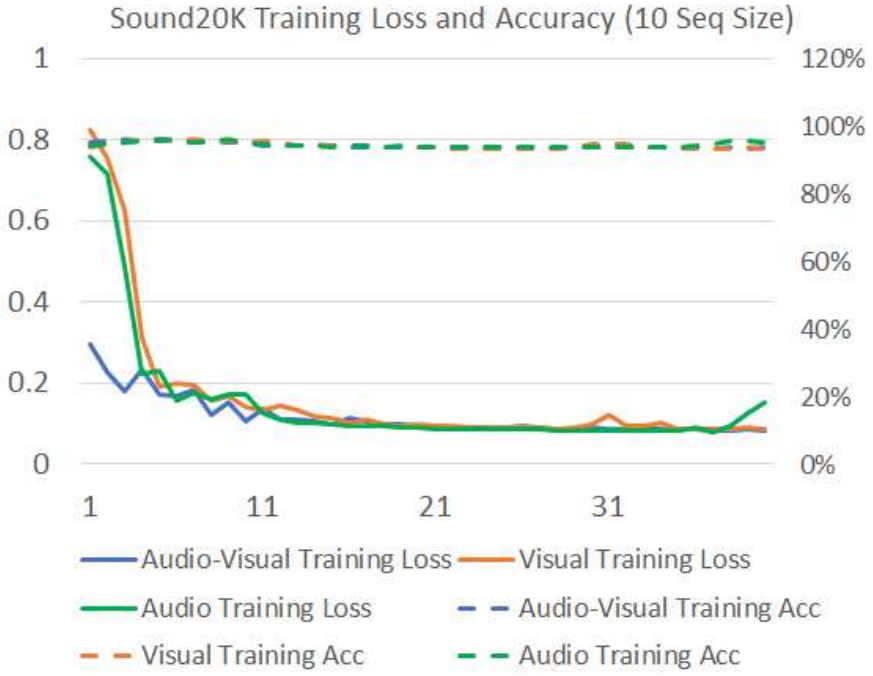

Figure 6.10: Training loss and accuracy for Sound20K dataset of 40 epochs using a sequence size of 10.

Limitations: our approach is currently implemented and evaluated with fixed-grid shapes. Further experimentation with other resolutions, residual architectures (He et al., 2015a), adaptive grids, and multi-scale reasoning (Denton et al., 2015) are worth exploring. Material classification is predicted based on audio alone, given the textureless image renderings of the datasets used. Also, only a single material is inferred for the entire geometry rather than per voxel classification. Finally, the trade-off between additional views and additional auditory inputs could be further explored.

Future Work: evaluation of other real-time object trackers, such as YOLO and Faster R-CNN, can be performed and trained on other existing datasets, such as COCO and SUN RGB-D. Further investigations can also examine how the error introduced by object tracking propagates to reconstruction error. Same applies to errors from sound source separation and being able to accurately associate unmixed sounds with their corresponding visual object tracks. Next, while audio helps classify the material of the reconstructed geometry, we assume a single material classification based on audio alone and apply that to all voxels. Research on classifying material per voxel using both audio and visual data could expand part segmentation research into reconstructing objects with different materials. Rather than being fully deterministic, fusing audio and visual information for generative models to reconstruct geometry and material may also be of interest to the research community. Then, there may be more than one possible 3D reconstruction for a given image or sound. Beyond reconstruction, audio may also enhance image and sound generation, as well as memory and attention models. For instance, image generation using an audio conditioned GAN and sound generation based on image conditioning could be explored, similar to WaveNet (van den Oord et al., 2016a) local and global conditioning techniques. Finally, testing on real data in the wild and larger datasets of annotated audio and visual data allow for future research.