CHAPTER 5: AUDIO-VISUAL OBJECT TRACKING FOR MULTIPLE OBJECTS[1]

This chapter describes an audio-visual object tracking (AVOT) neural network that reduces tracking error and drift by using audio of the impact sounds from object collisions, rolling, etc. It may be used in conjunction with other neural networks to augment visually based object detection and tracking methods. Using the synthetic Sound-20K audio-visual dataset, AVOT outperforms single-modality deep learning methods, when there is audio from object collisions.

5.1 Introduction

Deep learning has enabled state-of-the-art techniques for image classification and object detection in images and video (Liu et al., 2016; Redmon et al., 2015a; Ren et al., 2015c). Object tracking classifies bounding boxes for each object in a video over time. These methods are useful for applications in autonomous driving (Geiger et al., 2012), mobile robotics (Schulz et al., 2001), person tracking (Checka et al., 2001), speaker recognition (Spors et al., 2001; Qian et al., 2019), and 3D reconstruction (Prisacariu et al., 2015). For more granularity beyond bounding boxes, object segmentation provides pixel-level annotations (Voigtlaender et al., 2019; Perazzi et al., 2016). These existing object trackers achieve real-time performance and continue to improve on accuracy and the number of classes that they can detect.

However, occlusion, similar object categories, and smaller object sizes remain a challenge for visually based trackers (Liu et al., 2016). Auditory cues can assist in these exacting areas, especially when similar and/or smaller objects are of a different material (Aytar et al., 2016). In this paper, we propose an audio-visual object tracker (AVOT) that augments visual only trackers with fused audio in a jointly trained end-to-end model. It is evaluated using synthetic Sound-20K dataset (Zhang et al., 2017c), consisting of tabletop sized objects of different geometry and materials. The data contains videos with multiple objects of various shapes (e.g. bottle, knife, etc.) and materials (e.g. steel, wood, etc.) colliding in a virtual scene.

[1] This chapter previously appeared as an article in the International Conference on Robotics and Automation (ICRA). The original citation is as follows: Justin Wilson and Ming C. Lin. Avot: Audio-visual object tracking of multiple objects for robotics. 2020a

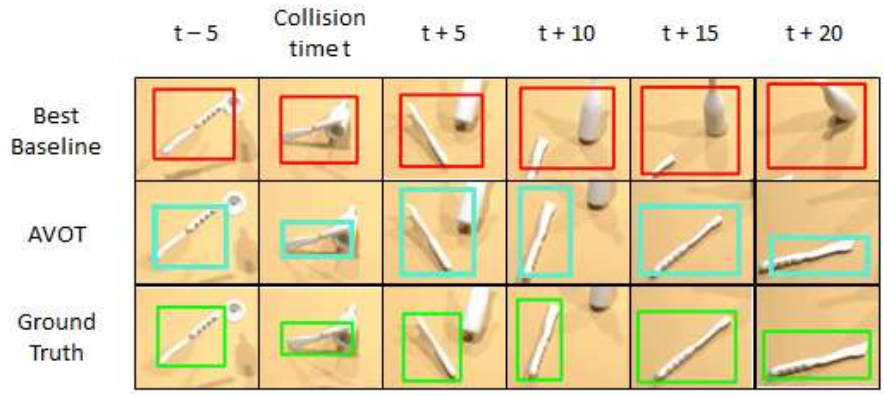

Figure 5.1: An example failure case improved by our audio-visual object tracker. (Top row) best baseline, CSRT in this case, incorrectly latches to the wrong object after collision. (Middle row) our AVOT method continues to correctly track the object post-collision. (Bottom row) ground truth annotated by the experimenter. For clarity, we show the bounding box for only one of the objects being tracked, although the methods track both objects. Please see the Supplementary Video for more demonstration.

Colliding includes objects colliding within the scene, with each other, rolling, etc. We use videos with one, two, or three colliding objects.

Other than speaker recognition, this is the first use of an audio-visual neural network for tracking tabletop sized objects and enhancing visual object trackers. The key contributions of this work include:

- An end-to-end, jointly trained audio-visual object tracker (AVOT) to enhance visual object tracking;
- Ground truth bounding box annotations for Sound-20K audio-visual dataset with 1, 2, and 3 object scenes;
- Scheduler for object detection re-initialization based on audio onset detection when using multimodal tracking.

Fusing audio with visual data, AVOT achieves 77.7% IoU post-collision tracking accuracy compared to 68.6% IoU using deep-learning visual tracking, SSD– (Liu et al., 2016), and 38.4% using CSRT (Lukezic et al., 2017) for virtual scenes with multiple objects based on our annotated Sound-20K dataset of 19 tabletop sized object classes of varying geometry and materials.

5.2 Background and Related Work

While object detection methods must search over the entire search space to first detect an object, tracking algorithms can be much faster by leveraging knowledge from previous frames to reduce the search

Object Detection/Tracking Datasets

Dataset	# Class	# Img/Vid
COCO (Lin et al., 2014b)	80	330K img
DAVIS (Caelles et al., 2019)	384	10.5K img
ILSVRC (Russakovsky et al., 2015)	1000	1.4M img
KITTI (Voigtlaender et al., 2019; Milan et al., 2016)	8	10.9K img
OTB (Wu et al., 2015a)	100	100 vid
PASCAL VOC (Everingham et al.)	20	21K img
Sound-20K (Zhang et al., 2017c)	55+	20K vid
VOT2018 (Kristan et al., 2016)	35	147K img
YouTube–VOS (Xu et al., 2018)	7800+	4K+ vid

Table 5.1: In contrast to other datasets, the Sound-20K dataset contains the largest audio-visual data for object interactions in a virtual scene and provides an excellent baseline for assessing the accuracy of our AVOT method against others.

space. However, this can make error recovery difficult. Tracking is also more complex because unlike object detection bounding boxes which are class specific, tracking is object specific. It assigns identifiable bounding boxes to each object and attempts to maintain each assignment over all frames. So, tracking not only detects but also maintains bounding box assignment for each object, over all frames. More granular than bounding boxes, segmentation may also be performed for pixel-level annotations. For an attribute and performance comparison, object tracking benchmark (Wu et al., 2015a) provides attribute and performance comparisons between various methods and evaluation criteria.

The majority of object detection and tracking methods are visually based, even though some datasets are generated from videos with audio. Table 5.1 lists commonly used datasets for object detection and tracking evaluation. We add ground truth bounding box annotations to the Sound-20K dataset and use it as a baseline for assessing the accuracy of our AVOT method. While the general methodology of research areas such as speaker detection and person tracking leverage both audio and visual information, their implementations are specifically aimed at tracking human speakers (e.g. face detection is part of their pipeline). Our method aims to be applicable in a broader context and does not make assumptions about the targets. It currently can track up to nineteen object-material classes. Next we discuss object detection, tracking, and audio-visual techniques in more detail, as compared to our work.

5.2.1 Object Detection

In addition to overall classification of an image, researchers are interested in also detecting and classifying the specific objects within an image. This can be achieved by using object detection methods to locate and label each object with a class-specific bounding box. As is similar in image classification, object detection techniques require large amounts of training data but in its case, more annotations for each example. Because, in the case of object detection, training data requires both class labels and bounding box coordinates for each object. For example, the PASCAL Visual Object Classes (VOC) dataset contains images, object annotations, and segmentations for twenty different classes. Other available datasets are mentioned in Table 5.1. Unfortunately, only a few datasets make available the video and accompanying audio, making audio-visual methods more time-consuming to explore. We contribute our Sound-20K ground truth annotations to aid future audio-visual research in this area.

Video object detection: object detection can be performed not only on images but on video as well. Here, additional contextual information is available such as sound and image sequence. This temporal memory has allowed video detection to achieve start-of-the-art performance and speeds by learning lightweight scene features for mobile (Liu et al., 2019) and shifting channels along the dimension of time (Lin et al., 2019). However, video also introduces new challenges such as motion blur, defocus, and various poses. Temporal coherence can also be used to overcome these defects with flow-guided feature aggregation (Zhu et al., 2017), for instance. Finally, in addition to scene features, time shifting, and temporally coherent features, temporal propagation for on demand detection has also yielded efficiency gains (Chen et al., 2018).

5.2.2 Object Tracking

Object tracking differs from object detection in that the labels and bounding boxes are dependent. In other words, tracking attempts to establish correspondences of the same object over multiple frames, for example, one particular car in traffic over time. While object tracking has been studied for decades, numerous factors remain a challenge, such as illumination variation, occlusion, and background clutters (Wu et al., 2015a). Given the sequential nature of the task and method, tracking can be fast and efficient but also accumulate error and drift. Moreover, it is not easy for object trackers to recover from failure or an incorrect assignment to another object. Approaches such as frame skipping, Siamese trackers, and deep

learning are a few of the existing techniques being used to perform object tracking and segmentation quickly and accurately.

Frame skipping: this baseline approach *detects* every N-th frame and *tracks* frames in between. Advantages of frame skipping are realized in terms of speed and precision by efficiently tracking on a majority of the frames while still allowing for correction by performing detection every N keyframes. This can be referred to as a fixed scheduler. Dynamic schedulers have also been considered. For instance, Detect or Track (Luo et al., 2019) uses a scheduler network to determine whether to detect or track at certain frames. In our research, we propose an audio-based scheduler procedure that can alter tracking during audio onset detection in videos.

Siamese trackers: Siamese neural networks take two inputs and, with shared weights, predict if the two inputs belong to the same output class. Fully-convolutional Siamese approaches can be used for object tracking (Bertinetto et al., 2016; Li et al., 2018; Zhu et al., 2018; He et al., 2018; Yang and Chan, 2018) and unlike batch processing, online Siamese methods can perform tracking on streaming video with access only to current and previous frames (Wang et al., 2019). To improve initialization of online adaption-based deep networks such as these, offline meta-learning has been applied (Park and Berg, 2018). Asymmetric Siamese networks have also been studied and learn a linear template to search test images by cross-correlation (Valmadre et al., 2017).

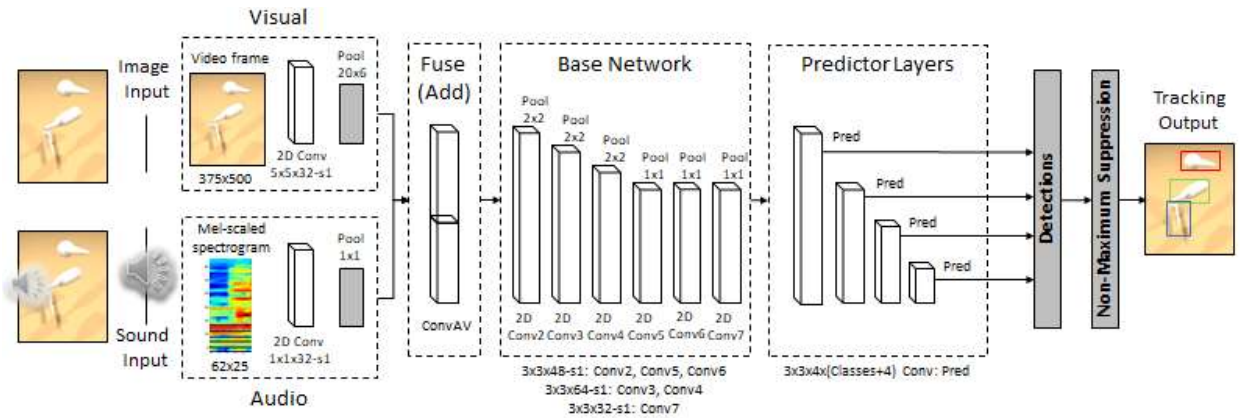

Figure 5.2: Audio-Visual Object Tracker (AVOT) neural network architecture. AVOT is a feed-forward convolutional neural network that classifies and scales a fixed number of anchor bounding boxes to track objects in a video. Here, we define an object based on its geometry and material. Convolutional layers from the visual and audio inputs are fused using an add merge layer before being input into a base network of convolutional layers similar to standard classification networks. The base is then followed by predictor layers for detection, as is done in SSD, however designed and optimized for our audio-visual dataset and task. The single best detection for each object is then selected using non-maximum suppression.

Deep learning: last but not least, object tracking performed using deep learning. Faster R-CNN (Ren et al., 2015a) is a real-time, state-of-the-art object tracker and four staged end-to-end neural network. First, a convolutional feature map of the image is obtained by extracting from a convolutional layer of a pre-trained CNN (e.g. ImageNet (Krizhevsky et al., 2012a), ResNet (He et al., 2016), MobileNets (Howard et al., 2017), DenseNet (Huang et al., 2017), etc.). The second stage is a Region Proposal Network, which are reference bounding boxes uniformly placed across the image. In this stage, specific regions are identified and adjusted based on the convolutional feature map from the first step. The third stage applies Region of Interest (RoI) Pooling to extract features from the convolutional map for each region. The fourth and final step then uses those features to classify the content in the bounding box (e.g. bottle, table, etc., background) and adjust the classified bounding box to a better fit, predicting $\Delta x_{center}, \Delta y_{center}, \Delta_{width}, \Delta_{height}$ from an anchor.

Single Shot MultiBox Detector (SSD) (Liu et al., 2016) is another real-time, state-of-the-art object tracker. SSD is slightly better than YOLO (Redmon et al., 2015a, 2016) in terms of speed while improving upon accuracy with additional feature layers on top of a base network [2]. Furthermore, SSD is slightly better than Faster R-CNN in terms of accuracy while eliminating object proposals with multiple feature maps of differing resolution. Although SSD uses similar default boxes, it applies them to several feature maps of different resolutions. In addition to a single unified framework for training and prediction, SSD input images are smaller at 300 x 300, compared to 512 x 512 for Faster R-CNN and 448 x 448 for YOLO. This enables faster processing over other single shot, region proposal, and pooling techniques. This permits a wider range of computer vision applications to leverage this architecture. We use SSD as both a baseline and base network for our AVOT tracker.

5.2.3 Audio-Visual Methods

Audio-visual techniques have been used for speech separation (Ephrat et al., 2018a), object and geometry classification (Zhang et al., 2017c,b; Sterling et al., 2018), and audio-visual correspondence learning (Arandjelovic and Zisserman, 2018, 2017). Most directly related to audio-visual object tracking is speaker recognition (Spors et al., 2001; Qian et al., 2019), tracking from audio-visual data using a linear prediction method (Anusha and Roy, 2015), and object detection and tracking with audio and optical sig-

[2] ImageNet VGG-16 was used as a base, but other neural networks should also produce good results.

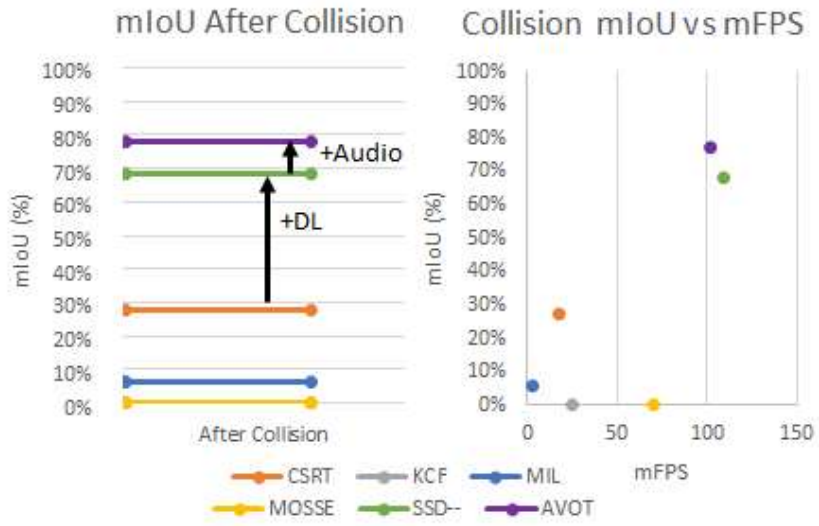

Figure 5.3: Existing object trackers performance decline when objects collide in a moving two object Sound-20K virtual scene whereas AVOT improves with audio onset. Post-collision (i.e. when there's audio), deep learning (DL) methods achieve nearly 40% higher in accuracy over other methods and AVOT further outperforms SSD– by another 10% in *mean Intersection over Union* (mIoU) with an added benefit of audio-visual input. A scheduler network gated on audio can be used to achieve the best run performance and/or the highest accuracy across all cases using multimodal trackers.

nals (Holz). For speaker recognition, a face tracking algorithm and microphone array are used to estimate speaker position. These methods fuse audio and visual data by leveraging time delays in audio and motion changes in visual. While both modalities, in theory, can distinguish these changes, one may be more adept to do so. Also, the fusion of the two can decrease uncertainty and increase reliability (Ngiam et al., 2011b). Finally, audio can also come from contact microphones or acoustical sensors to capture touch sounds and optical signals for gesture recognition (Holz). In our approach, we leverage audio from impact sounds of objects and images from video.

5.3 Technical Approach

Unlike visually based object trackers, our method defines each object by its geometry and material. With audio-visual data, the same shape (e.g. bottle) with different materials (e.g. steel vs. wood) are distinguishable and are therefore considered to be different objects. Our work also considers colliding objects. While a challenge for visually based tracking methods (Fig. 5.3), they provide auditory cues for an audio-visual object tracker. Scheduling between trackers can then be enabled based on audio availability.

Given the location of an object in the first frame of video, the object tracking task is to quickly and accurately estimate its position in all successive frames (Smeulders et al., 2014). More specifically, for each video frame in a sequence $F = f_1, f_2, ..., f_N$ where N is number of frames, obtain bounding boxes $B = b_1, b_2, ..., b_M$ where M is the number of objects.

5.3.1 AVOT Neural Network Architecture

Similar to existing object tracking architectures, AVOT is a feed-forward convolutional neural network that classifies and scales a fixed number of anchor bounding boxes to fit each object in an image. We define an object based on its geometry and material. AVOT leverages audio and visual data for a more granular definition of an object to distinguish between objects with the same appearance but different materials.

Audio input: Audio frames from Sound-20K (Zhang et al., 2017c) videos match the image frame rate of 33 frames per second. As a result, each jpeg image has a corresponding 29 ms audio wav file. The audio is converted to mel-scaled spectrograms and serve as the audio input given their performance in CNNs for other tasks (Huzaifah, 2017a). They are computed using a short-time Fourier transform with a 512 sample Hann window and 12.5% overlap. A Hanning (Hann) window was selected for its suitability for a variety of signals, good frequency resolution, and reduced spectral leakage. Each spectrogram is individually normalized and downsampled to a size of 62 frequency bins by 25 time bins (Fig. 5.4). Binning provides for appropriate fusion with image dimensions and weight matching to the logarithmic perception of frequency (Sterling et al., 2018).

Image input: image dimensions are 500 x 375 pixels. Since SSD evaluated input sizes 300 x 300 and 512 x 512 (YOLO 448 x 448), our images are augmented but input dimensions unmodified as they fall within range of previous work. For data augmentation, we use common image transformations and sampling strategy similar to SSD and YOLO. Random cropping can be especially useful for creating zoomed in and out training examples to aid the classification of small objects in PASCAL VOC and Sound-20K. Each training image randomly samples from a data augmentation sequence to make the model more robust to object size and shape (Liu et al., 2016). We use a reduced layer variation of VGG16 (Simonyan and Zisserman, 2015) as the base network leading up to our detection prediction layers. Images were

extracted from video using ffmpeg with CRF scale set to 0 (lossless) and libx264 set to vcodec (Zhang et al., 2017c). Each image is fused with its corresponding audio via an add-merge layer.

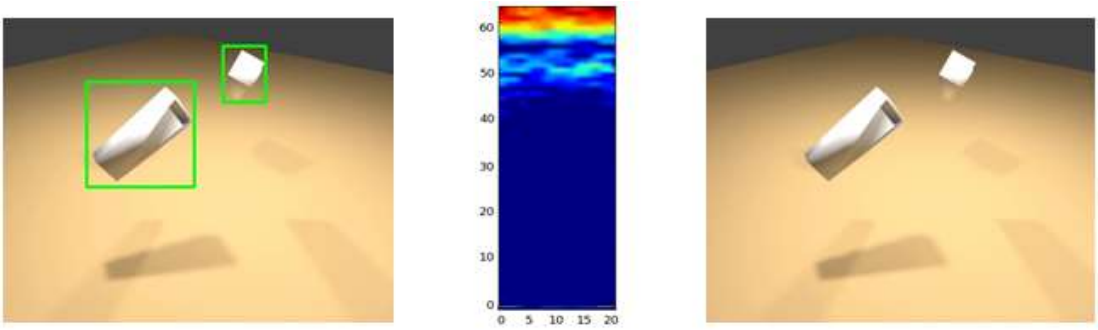

Figure 5.4: AVOT needs ground truth boxes (left), input audio from the scene video converted to a mel-scaled spectrogram (center), and input image (right) for each object during training. We predict shape offsets and confidences for all object categories where an object is defined by its geometry and material.

Architecture: Fig. 5.2 illustrates the layers of our multimodal object tracker neural network. The early visual layer is based on (LeCun and Bengio, 1998) and audio layer based on impact (Sterling et al., 2018) and environmental sound (Huzaifah, 2017a) classification. Convolutional layers from the visual and audio inputs are fused using an add merge layer. A multiply-merge layer was also considered and resulted in a similar training loss, however, at 1.5x the number of training epochs. Fused features are then input into a base network. Given our relatively small annotated audio-visual dataset, our base network is a reduced version of the standard image classification architecture (Krizhevsky et al., 2012a). The base is then followed by predictor, or also referred to as feature or classifier layers. Upper and lower feature maps are used for detection, as is done in SSD, to promote consistency and capture fine details respectively. The single best detection for each object is then selected using non-maximum suppression.

5.3.2 AVOT Dataset

Ground truth annotations were manually labeled by the experimenter for 18 objects. Each object is unique by geometry and material. The dataset is comprised of 17 three second videos of 103 image and audio frames each. This resulted in a total of 1,752 audio and visual segments. Videos contained one, two, and three colliding objects per scene. Our training and test datasets are split 80% and 20% respectively. The test dataset randomly samples frames from each video that are held out from training and used only for evaluation. For example, a video with 100 frames will have 20 frames randomly selected for test and

the remaining 80 frames used for training. Fig. 5.5 shows loss by epoch for our AVOT tracker compared to a variation of visually based SSD.

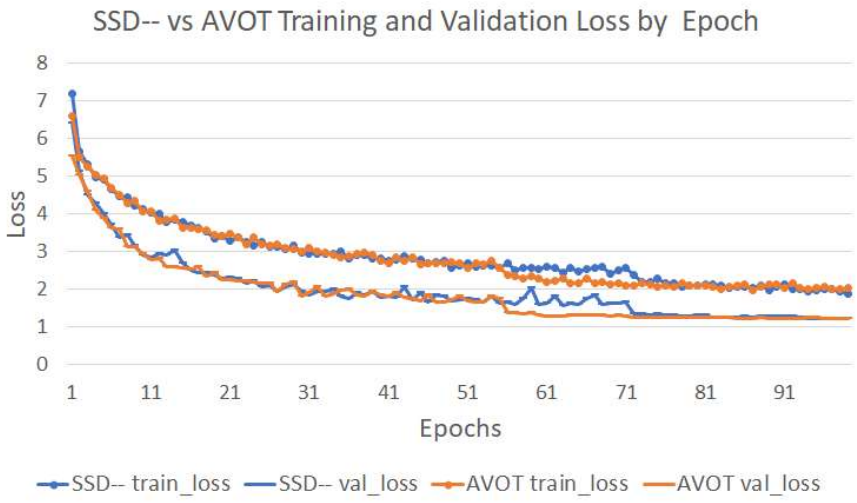

Figure 5.5: The training (circle) and validation (line) loss for SSD– (blue) and AVOT (orange). Multimodal AVOT loss seems to decrease more consistently than visual only SSD with reduced layers, denoted as SSD–.

5.3.3 Implementation Details

All models were implemented with Tensorflow (Abadi et al., 2016) and Keras (Chollet, 2015). AVOT was run with early stopping at a maximum of 100 epochs, 100 steps per epoch, and batch size of 16 (Fig. 5.5). Training was performed using an Adam optimizer (Kingma and Ba, 2015) and loss as defined by the weighted sum of localization loss (Smooth L1) and confidence loss (Softmax). We use a reduced variation SSD for predictor layers. AVOT anchor box scaling factors were set to 0.08, 0.16, 0.32, 0.64, and 0.96 and aspect ratios 0.5, 1.0, and 2.0 (Liu et al., 2016). Here, we do not use SSD aspect ratios 1/3 or 3 given a smaller number of target classes. There are five scaling factors for four predictor layers because the last scaling factor is used for the second aspect ratio box of the last predictor layer. Although fewer layers, detections are still based on small 3 x 3 kernels at each feature map offset (Liu et al., 2016).

Initialization: our AVOT neural network uses he_normal initialization (He et al., 2015b). For evaluation, we also initialize baseline methods with he_normal rather than fine tune on pre-trained networks. Recent research suggests equivalent performance between random initialization for training instead of

pre-trained weights (He et al., 2019). Furthermore, given a smaller dataset of Sound-20K with ground truth annotations, we have reduced the layers of baseline implementations to avoid overfitting.

Non-maximum suppression (NMS) (Neubeck and Gool, 2006): object trackers may produce more than one overlapping bounding box that are greater than the confidence and IoU thresholds for the same object. NMS is a post-process that selects the bounding box with the greatest confidence and suppresses remaining bounding boxes that overlap this maximum by some threshold. Here, NMS confidence and IoU threshold are set to 0.5 (Liu et al., 2016).

Scheduler network: Impact sounds from objects colliding emulate a type of scheduler network that can improve detections post collision. For added efficiency, only visual inputs can be processed leading up to audio onset. After, both audio and visual inputs can be used. In the case of our synthetic dataset, there is no audio prior to collision which makes audio onset easier to detect than videos with noise.

5.4 Experiments and Results

Evaluation was performed using ground truth annotations on the Sound-20K audio-visual dataset. This dataset is comprised of synthetic videos of multiple objects colliding in a scene. Training took roughly 30 minutes running on Ubuntu 16.04.6 LTS with a single Titan X GPU. We use Intersection over Union (IoU) for accuracy between ground truth and predicted object bounding boxes. As a general rule of thumb, a true positive prediction occurs when $IoU \geq 0.5$, according to the PASCAL VOC challenge. We measure the speed in mean frames per second (mFPS) with a batch size of 16 using a Titan X and cuDNN v7.4.2.

OpenCV implementations: online Multiple Instance Learning (MIL) (Babenko et al., 2009), Kernelized Correlation Filters (KCF) (Henriques et al., 2015), Discriminative Correlation Filter with Channel and Spatial Reliability (CSRT) (Lukezic et al., 2017), and an adaptive correlation filter known as Minimum Output Sum of Squared Error (MOSSE) (Bolme et al., 2010) are a few trackers available in OpenCV (Bradski, 2000). We selected to evaluate these as baselines due to their advantages in terms of accuracy and/or speed. For these methods, appearance is learned from first frame bounding boxes that are initialized with ground truth coordinates.

mIoU / mFPS Object Tracking Accuracy by Method

Method	2 Objects	3 Objects
AVOT (Ours)	**58.3%** / 101.6	**66.1%** / 101.0
CSRT (Lukezic et al., 2017)	46.9% / 17.1	30.1% / 4.7
KCF (Henriques et al., 2015)	13.5% / 24.9	1.7% / 38.6
MIL (Babenko et al., 2009)	43.0% / 2.5	21.6% / 1.6
MOSSE (Bolme et al., 2010)	7.6% / 70.4	1.0% / **74.5**
SSD– (Liu et al., 2016)	55.5% / **108.7**	65.9% / **103.8**

Table 5.2: Multiple network models were evaluated on accuracy and time using mean Intersection over Union (mIoU) and mean frames per second (mFPS). **Ours is AVOT**. Failure cases for baseline methods without audio tend to classify to the correct geometry but wrong material. By exploiting both visual and audio data, AVOT achieves the highest level of tracking accuracy, with nearly comparable best runtime performance, over existing visual tracking methods.

5.4.1 Our Results vs. Baselines

Given our limited number of training examples in our audio-visual dataset, we used a reduced layer implementation of SSD (labeled in Table 5.2 as SSD–) for a baseline and base network for AVOT. Our AVOT neural network outperforms SSD– and other baseline methods after collision (Fig. 5.3). As shown in Fig. 5.3, AVOT was able to achieve the highest level of accuracy of 80% in *mean Intersection over Union* (mIoU)– about 10% more accurate than SSD. While these results are AVOT only, we further propose a scheduler network (**Algorithm 1**) to switch between AVOT and other methods based on audio onset to maximize accuracy and performance over all frames in multimodal object tracking.

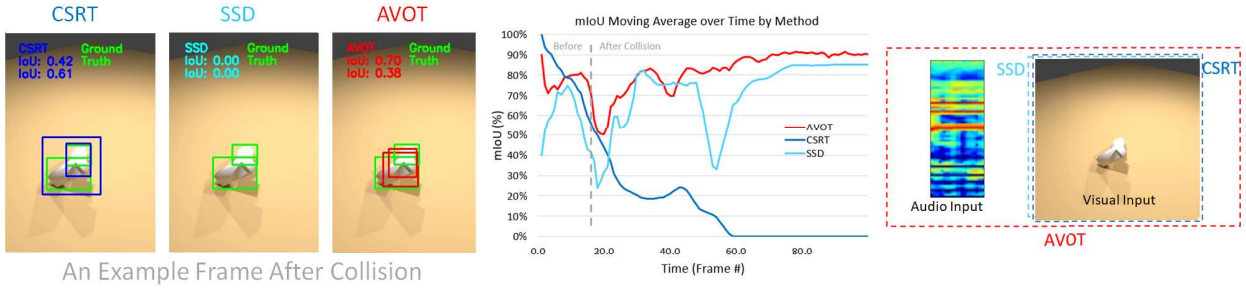

Figure 5.6: We compare CSRT and SSD to our AVOT method for multi-object tracking. Two colliding objects with the same geometry but different materials are tracked free-falling in a virtual scene from Sound-20K (Zhang et al., 2017c). CSRT is unable to maintain tracking post-collision and while SSD recovers, it temporarily loses tracking at the time of occlusion. Audio-visual AVOT maintains tracking across all frames. Please see the Supplementary Video for more demonstrations.

5.4.2 Maximization Activation

We analyzed activation maximizations to visualize the spectrogram audio and visual input which would produce the highest activation for a given volume class. They demonstrate features being learned by both modalities for the object tracking task. Please see the Supplementary Video for demonstration.

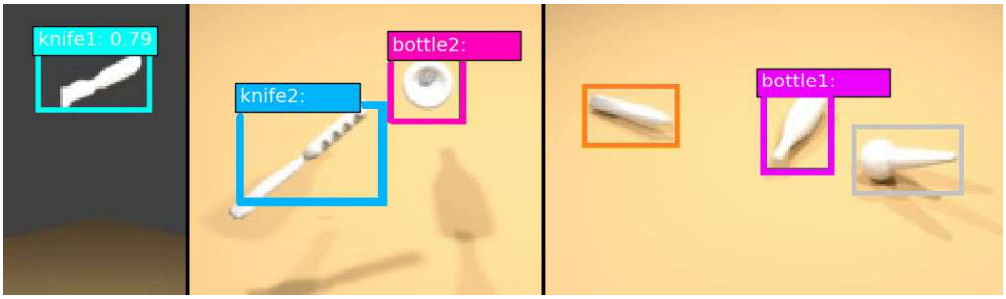

Figure 5.7: Examples of AVOT applied to virtual scene of Sound-20K with predicted bounding box. These are exemplary screenshots of AVOT performing object tracking before and after collisions for one, two, and three object virtual scenes. Notice alphanumeric labels (e.g. bottle1 and bottle1) to differentiate the same geometry with different materials.

5.5 Conclusion

We present AVOT, an end-to-end trained neural network for object tracking using audio and visual data from videos. To distinguish between similar objects with different materials, we define an object based on its geometry and material. This more granular categorization benefits from a multimodal learning approach using audio and visual data, where audio is typically available from the sources of video but are currently underutilized. By fusing audio with visual data, our audio-visual object tracker (AVOT) outperforms single-modality methods when audio is present from impact, collision, and rolling sounds while maintaining real-time performance. We evaluated against Sound-20K and make our audio-visual data along with ground truth bounding box annotations available for future research in this area.

Future work: we will expand the size of our training set by annotating more objects in the Sound-20K dataset, increase the number of object classes that we are predicting, evaluate alternative fusion methods, and perform sensitivity analysis on scaling factors and aspect ratios. We would also like to augment our audio data and experiment with a variation of our object tracker with audio only.