CHAPTER 8: SUMMARY AND CONCLUSIONS

8.1 Summary

This dissertation was motivated by the advances in computer vision and the possibility of realizing the potential benefits of single model audio and multimodal audio-visual learning. Whether by coupling a fluid and structure to form a rigid double body for sound synthesis or fusing audio and visual inputs for object classification, tracking, and reconstruction, audio is readily available for use along with its corresponding images when datasets are generated from video.

In addition, certain conditions lend themselves more preferably to one modality or a combination of multiple modes. For example, vision-based methods are sufficient for most static objects and scenes. However, reflective and textureless surfaces may be better suited for audio methods since visual data may be ambiguous or changing over time and viewpoint. Finally, audio-visual techniques can use scheduling to use the appropriate inputs given the current state, account for drift error of dynamic objects and scenes, and handle occlusions from cluttered scenes.

8.2 Future Work

The immediate next research steps to further enhance audio-visual performance and processing is further analysis of tasks that can capture audio in their datasets and benefit from its signal (e.g. inference, tracking, reconstruction), gating or scheduling of when single or multiple modalities are used, more neural network architectures, loss functions, and fusion models, and augmenting with even more modes.

8.3 Conclusion

In research and practice, sound is a key contributor to the level of immersion and sense of presence in virtual and imaginative environments. A distraction from any of the senses can cause a ‘break in presence’. A goal in computer graphics is to continue to enhance rendering pipelines with new technology, methods,

Fluid-structure coupling used added mass operator for sound synthesis
Pouring Sequence Neural Network (PSNN) for weight estimation of liquid
Audio-Visual Object Tracker (AVOT)
Echo-Reconstruction: audio-augmented scene reconstruction on mobile devices
3D-MOV: audio-visual LSTM autoencoder for 3D reconstruction of multiple objects from video

Table 8.1: Summary of contributions

and data. While vision-based methods cover many use cases, alternate modalities such as audio can augment the level of detail and coverage of tasks in computer vision, graphics, augmented, and virtual reality. Since many sound models are physics-based and training data generated from video, established visual pipelines and datasets can be extended to generate and use sound based on the same physics and video capture used for visual data. Much of the research conducted on visual data is also relevant to sound sources. This presents opportunities for audio-based research to advance quickly based learnings from decades of vision research as well as novel directions for fusing audio, visual, and other data for multimodal learning.