Multisensory Machine Intelligence

Ruohan Gao

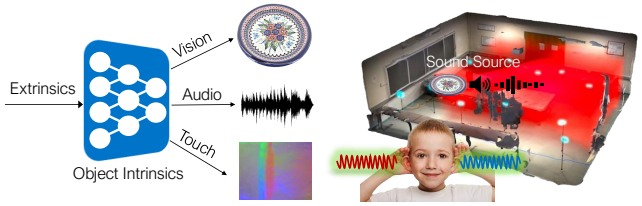
University of Maryland, College Park

The future of Artificial Intelligence demands a paradigm shift towards multisensory perception—to systems that can digest ongoing multisensory observations, that can discover structure in unlabeled raw sensory data, and that can intelligently fuse useful information from different sensory modalities for decision making. While we humans perceive the world by looking, listening, touching, smelling, and tasting, traditional form of machine intelligence mostly focuses on a single sensory modality, particularly vision. Therefore, my research, which I call **multisensory machine intelligence**, aims to enpower machines to emulate and enhance human capabilities in seeing, hearing, and feeling, ultimately enabling them to comprehensively perceive, understand, and interact with multisensory world.

Particularly, my research focuses on two important aspects of the multisensory world: **multisensory objects** (Fig. 1a) and **multisensory space** (Fig. 1b). In both aspects, we have made technical innovations on addressing three important problems—**1) Capturing:** how can we build systems to reliably and efficiently capture multisensory data from real-world objects and space? **2) Modeling:** how to effectively model multisensory objects and space with the right representation and physics-based differentiable simulation algorithms? **3) Applications:** what new cross-modal/multi-modal applications can be enabled in vision, graphics, and robotics with a unified multisensory representation?

My new faculty highlight talk will also be structured into the following two main parts:

Multisensory Objects. Our everyday activities involve perception and manipulation of a wide variety of objects. First, I study how to model the multisensory signals of real world objects, virtualizing each object by encoding its intrinsics (texture, material type, and 3D shape) with an implicit neural representation. Then we can render its visual appearance, impact sound, and tactile readings based on any extrinsic parameters. We introduce a dataset called OBJECTFOLDER that contains 1,000 implicitly represented objects each containing the complete multisensory profile of an object, and we succesfully perform Sim2Real transfer by learning from them. Vice versa, we also design differentiable inverse rendering algorithms for Real2Sim applications, where we infer a variety of physical properties of objects from their real-world observations.



(a) Multisensory Objects (b) Multisensory Space

Figure 1: **Multisensory Machine Intelligence**. My research studies two important aspects of the multisensory world: (a) multisensory objects, and (b) multisensory space.

Multisensory Space. Apart from objects, the space we live in is also multisensory. We see where the objects are and how the room is laid out, and we also hear them: sound-emitting objects indicate their locations, and sound reverberations reveal the room's main surfaces, materials, and dimensions. The two senses naturally work in concert to interpret spatial signals. Leveraging the spatial signal in videos, we devise approaches to lift a flat monaural audio signal to binaural audio by injecting the spatial cues embedded in the accompanying visual frames. We also introduce the more general visual acoustic matching task, where we transform the sound recorded in one space to another space depicted in the target visual scene. Recently, we have also proposed a differentiable room impulse response rendering framework that allows us to synthesize novel acoustic experiences through the space with any source audio. Beyond learning from passively captured videos or audio recordings, we have also explored sight and sound in embodied learning settings, exploring tasks such as embodied audio-visual navigation and echolocation in 3D environments.

My research draws inspiration from how we humans make use of all our senses in our everyday activities, thus it is highly interdisciplinary and involves various research areas, including computer vision, robotics, machine learning, augmented reality, acoustic learning, and cognitive science. My AAAI-25 new faculty highlight talk will survey my research experience and future plans on the research topics discussed above.