Learning 3D Representations *for* and *by* Interaction

Humans build powerful mental representations by watching and interacting with physical entities in the world. These representations 'intuitively' capture the laws and models of the physical world, allowing for complex reasoning and interaction with the environment. Current research in computer vision, however, falls short because existing 3D representations reason solely about the geometric structure of the world. To be useful for downstream applications, 3D representations must additionally reason about physical and material properties and affordances. Furthermore, these representations must be tuned to the context of the 3D scene they are extracted from. Thus, while prior approaches can generate geometrically accurate representations [1] they often fail in unintuitive ways when made part of physically-based interaction pipelines such as simulation, rendering, 3D printing, etc.

My research focuses on learning 3D representations that are better suited for physical interaction/ reasoning through watching or interacting with the physical world — this is what I call learning 3D representations *for* and *by* interaction.

A key characteristic of these representations is **Physical Realizability** which makes them directly amenable for robotics, augmented reality, graphics and manufacturing applications. Moreover, representations learned via **interaction** and **multiview consistency** capture more plausible world models and structures. I briefly talk about these and future directions in the following paragraphs.

Physically Realizable Representations: Existing methods for 3D shape generation trade-off topological correctness for geometric accuracy [1] which causes them to produce unrealistic results for physically based tasks like rendering, simulation and 3D printing. As part of my master's thesis, I developed **Neural Mesh Flow** [project page] - a method that enables several interesting physically based applications in addition to reconstructing geometrically accurate shapes. **Neural Mesh Flow** [2] casts the shape generation problem as producing a diffeomorphic flow — via Neural ODEs — from a template shape to the target. The diffeomorphic properties *guarantee* the preservation of local mesh topology and result in physically plausible manifold meshes. Our method significantly bridges the gap between single-view 3D reconstruction and photorealistic rendering literature by generating upto **53 times** more topologically accurate meshes. I recently presented this at NeurIPS 2020 (**spotlight talk - top 4.1%)** as the first author.

Multimodal learning via interaction: A baby playing with a toy actively interacts with and gets to perceive it from multiple views and modalities, which assists in building geometrically and physically consistent mental representations. Similarly, designing strategies to **learn via interaction** can lead to physically plausible representations. One particular avenue of interest is in understanding 3D shapes via the sense of touch. I find it interesting to explore the interplay between these modalities (for instance, by curiosity-driven alignment) and learn to reconstruct an object by touching or manipulating it. I see my prior work with **Neural Mesh Flow** [2] as a key component to realizing such learning schemes as it generates physically plausible meshes that can be simulated accurately.

Learning via multiview consistency: It is possible to solve for unknown 3D geometry by exploiting the consistency between several predicted 'renderings' — via differentiable *rendering* [3]. Interestingly, this can be used to solve a critical problem in Computed Tomography (CT) reconstruction i.e. accurately resolving dynamics of fast moving objects. Clinically, this can be used for imaging subjects whose heart rate is more than 60 beats-per-minute or a fetus which is moving inside a mother's womb. My proposed framework **NeuralCT** — operates in a self-supervised manner to enable ordinary CT scanners to dynamically reconstruct a patient's active anatomy (like a 3D+time "movie") with upto **10-15 times** higher reconstruction accuracy. **NeuralCT** leverages an *object-aware* representation of patient anatomy and uses a differentiable parallel-beam Radon transform based renderer to generate **physically plausible** tissue deformations during the process of imaging.

In the future, I am interested in investigating representations that are better suited for a joint treatment of **reconstruction and interaction**. Ideally, such representations must employ both bottom-up and top-down reasoning and for tractability, be organized into a hierarchy. I find recent work on 3D scene graphs [4] particularly exciting and relevant, and I would be eager to extend such frameworks to incorporate additional attributes such as lighting, materials, inertia and affordances.

References

[1] Gkioxari, G., Malik, J. and Johnson, J., Mesh r-cnn. In *ICCV 2019*

[2] **Gupta, K**. and Chandraker, M., Neural Mesh Flow: 3D Manifold mesh generation via Diffeomorphic Flows . *In NeurIPS 2020*

[3] Zhao, S., Jakob, W. and Li, T.M., Physics-based differentiable rendering: From theory to implementation. *In ACM SIGGRAPH 2020 Courses*

[4] Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J. and Savarese, S., 3d scene graph: A structure … camera. *In ICCV 2019*