Email: jrenzhile@gmail.com
http://cs.brown.edu/people/ren

Zhile Ren | Research Statement

I develop new representations and algorithms for three-dimensional (3D) scene understanding from cluttered indoor RGB-D images and outdoor video sequences. I introduce novel representations for 3D object detection systems that localize objects with cuboids and describe room layouts by Manhattan structures. Using view-invariant 3D features, I capture 3D style-variations and design systems to detect small objects by modeling support surfaces. Finally, I develop cascaded prediction frameworks to model 3D contextual relationships and enable rapid understanding of scene properties including depth, motion, and segmentation.

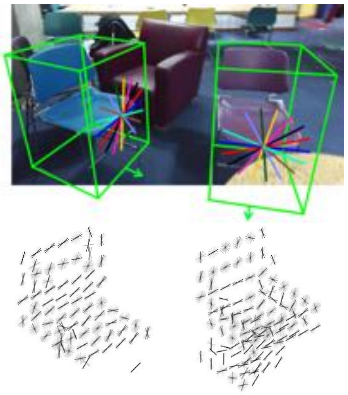

Figure 1: COG descriptor encodes orientation-invariant gradient feature for objects with different views.

Cloud of Oriented Gradient

A big challenge in representing 3D objects is how to design features that capture local appearances and are consistent when viewpoint changes. I introduce a Cloud of Oriented Gradient (COG) descriptor [1] that links the 2D appearance and 3D pose of object categories, and accurately models how perspective projection affects perceived image boundaries as captured by RGBD cameras in any orientation. I also propose a "Manhattan voxel" representation which better captures the 3D room layout of common indoor environments that follows Manhattan structure. Effective classification rules are learned via a structured prediction framework that accounts for the intersection-over-union overlap of hypothesized 3D cuboids with human annotations, as well as orientation estimation errors. The model is learned solely from annotated RGB-D images without the benefit of CAD models, but nevertheless its performance substantially exceeds the state-of-the-art on the SUN RGB-D dataset [4].

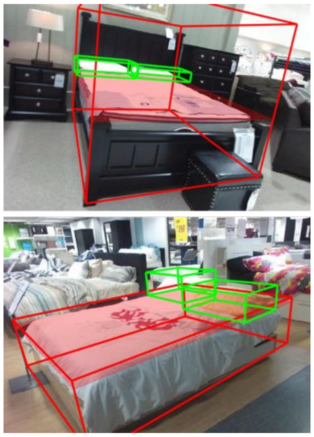

Figure 2: Modeling latent support surfaces can capture style variation of 3D objects and help detect small objects.

Modeling Latent Support Surface

Existing 3D representations for RGB-D images have limited power to represent objects with different visual styles. The detection of small objects is also challenging because the search space is very large in 3D scenes. However, much of the shape variation within 3D object categories can be explained by the location of a latent support surface, and smaller objects are often supported by larger objects. I design algorithms to use latent support surfaces to better represent the 3D appearance of large objects, and provide contextual cues to improve the detection of small objects [2]. The proposed system further improves 3D detection performances over all 19 object categories.

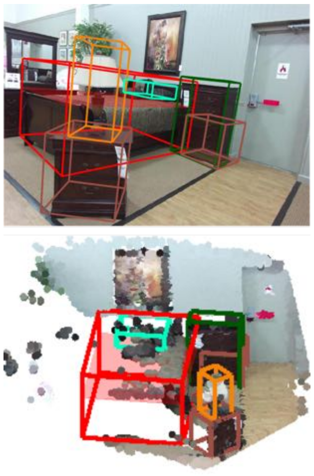

Figure 3: A sample 3D detection result.

Cascaded Prediction

Many works use Markov random fields (MRFs) to model contextual relationships of different components in their systems, and usually lead to optimizing an inefficient high-dimensional objective function. A cascaded prediction framework [5] can be interpreted as a directed graphical model with hidden variables and marginalizing the first-stage variables recovers a standard, fully-connected undirected graph. The cascaded representation is far more efficient and is used in a series of my research works to solve different vision tasks.

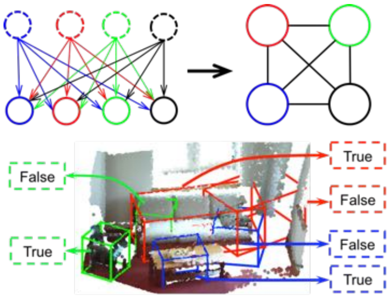

Figure 4: A graphical model of cascaded prediction and a 2-stage object detection system.

Modeling 3D Contextual Relationship Modeling contextual relationship in 3D is more effective than in 2D, because properties like 3D spatial overlap or scene layout are independent of viewing angles. I propose to adapt cascaded classification to model contextual relationships in 3D scenes. "First-stage" detections become input features to "second-stage" classifiers that estimate confidences in the correctness of detection hypotheses. The updated object detector lead to holistic scene interpretations of higher quality [1,2].

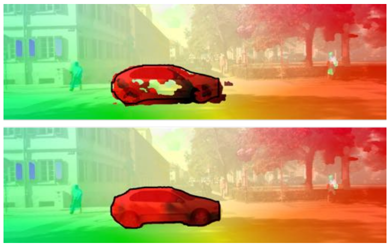

Figure 5: A noisy optical flow prediction is improved by modeling semantic segmentation, stereo and motion using a cascaded framework.

Efficiently Optimizing High-Dimensional Objectives The scene flow is the dense 3D geometry and motion of a dynamic scene. Given images captured by calibrated cameras at two frames, 3D motion field can be recovered by projecting 2D motion estimates onto a depth map inferred via stereo matching. With instance-level semantic segmentation that separates objects and backgrounds of scenes, I develop a cascaded classification model that iteratively update segmentation masks, stereo depth maps and optical flow fields [3]. At the time of publication, our algorithm was the state-of-the-art in two-frame scene flow estimation at KITTI benchmark.

Future Research

3D scene understanding is a growing research area but the success of 2D visual understanding using deep learning hasn't fully transferred to 3D. As an ongoing effort, I'm collaborating with industrial partners to utilize deep learning to train 3D detection systems in large-scale synthetically-generated datasets and hope to answer the question "How to effectively model object appearances in 3D using deep learning?" Moreover, although utilizing multi-frame inputs to infer object motions is a appealing research direction, few success happened in the past. Therefore I'm also designing algorithms to answer the question "What's the best way to model temporal information in motion estimation tasks?"

References

[1] Zhile Ren, Erik Sudderth, Three-Dimensional Object Detection and Layout Prediction using Clouds of Oriented Gradients, CVPR 2016

[2] Zhile Ren, Erik Sudderth, 3D Object Detection with Latent Support Surfaces, CVPR 2018

[3] Zhile Ren, Deqing Sun, Jan Kautz, Erik Sudderth, Cascaded Scene Flow Prediction using Semantic Segmentation, 3DV 2017

[4] Shuran Song, Lichtenberg Samuel, Jianxiong Xiao, SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite, CVPR 2015

[5] Geremy Heitz, Stephen Gould, Ashutosh Saxena, Daphne Koller, Cascaded classification models: Combining models for holistic scene understanding, NIPS 2009