CNN-based Scene Modeling: From Depth Estimation to 3D Reconstruction

Name: Zhaofeng Niu
Laboratory's name: Interactive Media Design Lab
Supervisor's name: Hirokazu Kato

Abstract

With the rapid improvement of image sensors and computer vision technologies, scene modeling has become more and more popular. For applications like augmented reality (AR) and virtual reality (VR), quick and precise 3D reconstruction of the real world, which describes actual objects in data format that can be used for displaying and computing, is necessary. Therefore, researchers have paid lots of attention to developing an efficient yet accurate scene modeling method and have achieved many encouraging results. However, most of these methods are based on expensive depth sensors and are vulnerable to depth noises, which largely limit their application areas. In the dissertation, the author focused on the design and implementation of CNN-based scene modeling, in order to get rid of expensive depth sensors and to make a more robust reconstruction method.

Firstly, the author focuses on developing a new monocular depth estimation method for easing the task of depth acquisition. Monocular depth estimation is an essential technique for tasks like 3D reconstruction. Although many works have emerged in recent years, they can be improved by better utilizing the multi-scale information of the input images, which is proved to be one of the keys in generating high-quality depth estimations. In this chapter, we propose a new monocular depth estimation method named HMA-Depth, in which we follow the encoder-decoder scheme and combine several techniques such as skip connections and the atrous spatial pyramid pooling. To obtain more precise local information from the image while keeping a good understanding of the global context, a hierarchical multi-scale attention module is adopted and its outputs are combined to generate the final output that is with both good details and good overall accuracy. Experimental results on two commonly-used datasets prove that HMA-Depth can outperform the existing approaches.

Then the author attempts to developing a new 3D reconstruction method that is robust to depth noises. The truncated signed distance function (TSDF) fusion is one of the key operations in the 3D reconstruction process. However, existing TSDF fusion methods usually suffer from the inevitable sensor noises. In this chapter, we propose a new TSDF fusion network, named DFusion, to minimize the influences from the two most common sensor noises, i.e., depth noises and pose noises. To the best of our knowledge, this is the first depth fusion for resolving both depth noises and pose noises. DFusion consists of a fusion module, which fuses depth maps together, as well as the following denoising module, which removes both depth noises and pose noises for TSDF volumes. To utilize the 3D structural information, 3D convolutional layers are used in the encoder and decoder parts of the denoising module. Also, a specially-designed loss function is adopted to improve the fusion performance in object and surface regions. The experiments are conducted on a synthetic dataset as well as a real-scene dataset. The results prove that our method outperforms existing methods.