Deep Depth Completion of a Single RGB-D Image

Xuewen Yang

July 20 2018

Abstract

The goal of this work is to complete the depth channel of an RGB-D image. Commodity-grade depth cameras often fail to sense depth for shiny, bright, transparent, and distant surfaces. To address this problem, the author train a deep network that takes an RGB image as input and predicts dense surface normals and occlusion boundaries. Those predictions are then combined with raw depth observations provided by the RGB-D camera to solve for depths for all pixels, including those missing in the original observation. This method was chosen over others (e.g., inpainting depths directly) as the result of extensive experiments with a new depth completion benchmark dataset, where holes are filled in training data through the rendering of surface reconstructions created from multiview RGB-D scans. Experiments with different network inputs, depth representations, loss functions, optimization methods, inpainting methods, and deep depth estimation networks show that our proposed approach provides better depth completions than these alternatives.

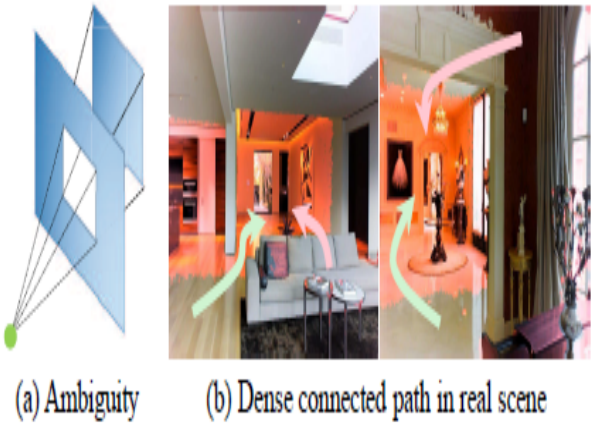

Figure 1. Using surface normals to solve for depth completion. (a) An example of where depth cannot be solved from surface normal. (b) The area missing depth is marked in red. The red arrow shows paths on which depth cannot be integrated from surface normals. However in real-world images, there are usually many paths through connected neighboring pixels (along floors, ceilings, etc.) over which depths can be integrated (green arrows).

1. Introduction

Depth sensing has become pervasive in applications as diverse as autonomous driving, augmented reality, and scene reconstruction. Despite recent advances in depth sensing technology, commodity-level RGB-D cameras like Microsoft Kinect, Intel RealSense, and Google Tango still produce depth images with missing data when surfaces are too glossy, bright, thin, close, or far from the camera. These problems appear when rooms are large, surfaces are shiny, and strong lighting is abundant C e.g., in museums, hospitals, classrooms, stores, etc. Even in homes, depth images often are missing more than 50% of the pixels (Figure 1).

2. Related Work

There has been a large amount of prior work on depth estimation, inpainting, and processing.

Depth estimation from a monocular color image is a long-standing problem in computer vision. Classic methods include shape-from-shading and shape-from-defocus [4]. Other early methods were based on hand-tuned models and/or assumptions about surface orientations [3]. Newer methods treat depth estimation as a machine learning problem, most recently using deep networks. For example, Eigen et al. first used a multiscale convolutional network to regress from color images to depths [1] [2].

Many methods have been proposed for filling holes in depth channels of RGB-D images, including ones that employ smoothness priors, fast marching methods, Navier-Stokes, anisotropic diffusion, background surface extrapolation, color-depth edge alignment, low-rank matrix completion, tensor voting, Mumford-Shah functional optimization, joint optimization with other properties of intrinsic images, and patch-based image synthesis.

3. Method

In this paper,the author investigate how to use a deep network to complete the depth channel of a single RGB-D image.The investigation focuses on the following questions: how can they get training data for depth completion?, what depth representation should they use?, and how should cues from color and depth be combined?.

3.1. Dateset

To create the dataset,the author utilize existing surface meshes reconstructed from multi-view RGB-D scans of large environments. There are several datasets of this type, including Matterport3D, ScanNet, SceneNN, and SUN3D, to name a few.They use Matterport3D. For each scene,they extract a triangle mesh M with x1-6 million triangles per room from a global surface reconstruction using screened Poisson surface reconstruction. Then, for a sampling of RGB-D images in the scene, the author render the reconstructed mesh M from the camera pose of the image viewpoint to acquire a completed depth image.

3.2. Depth Representation

The author focus on predicting surface normals and occlusion boundaries. Since normals are differential surface properties, they depend only on local neighborhoods of pixels. Moreover, they relate strongly to local lighting variations directly observable in a color image. For these reasons, previous works on dense prediction of surface normals from color images produce excellent results [1]. Similarly, occlusion boundaries produce local patterns in pixels (e.g., edges), and so they usually can be robustly detected with a deep network.

References

[1] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015. 1, 2

[2] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014. 1

[3] A. Saxena, M. Sun, and A. Y. Ng. Make3D: learning 3D scene structure from a single still image. *IEEE TPAMI*, 2009. 1

[4] S. Suwajanakorn, C. Hernandez, and S. M. Seitz. Depth from focus with your mobile phone. In *CVPR*, 2015. 1