Paper ID: 1091

M3VSNET: UNSUPERVISED MULTI-METRIC MULTI-VIEW STEREO NETWORK

Baichuan Huang[1,3], Hongwei Yi[2], Can Huang[3], Yijia He[3], Jingbin Liu[1], Xiao Liu[3]

1. Wuhan University 2. Peking University 3. Megvii Technology Limited

ICIP2021 Anchorage, Alaska.

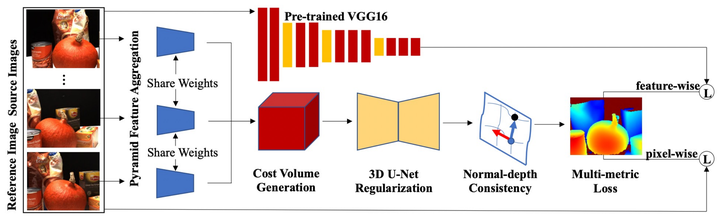
More about me

Introduction

The present Multi-view stereo (MVS) methods with supervised learning-based networks have an impressive performance comparing with traditional MVS methods. However, the ground-truth depth maps for training are hard to be obtained and are within limited kinds of scenarios. In this paper, we propose a novel unsupervised multi-metric MVS network, named M3VSNet, for dense point cloud reconstruction without any supervision. To improve the robustness and completeness of point cloud reconstruction, we propose a novel multi-metric loss function that combines pixel-wise and feature-wise loss function to learn the inherent constraints from different perspectives of matching correspondences. Besides, we also incorporate the normal-depth consistency in the 3D point cloud format to improve the accuracy and continuity of the estimated depth maps. Experimental results show that M3VSNet establishes the state of-the-arts unsupervised method and achieves better performance than previous supervised MVSNet on the DTU dataset and demonstrates the powerful generalization ability on the Tanks & Temples benchmark with effective improvement.

The code is available at https://github.com/whubaichuan/M3VSNet

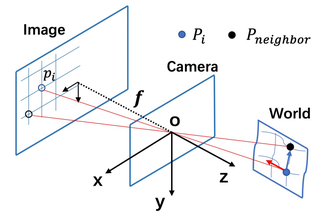
Network



The architecture of our proposed M3VSNet. It contains five components: pyramid feature aggregation, variance-based cost volume generation, 3D U-Net regularization, normal-depth consistency and multi-metric loss function.

Innovation

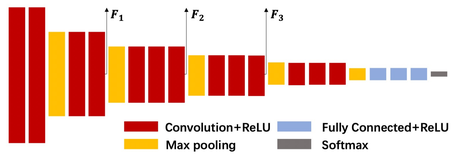
1. Normal-depth Consistency: The initial depth still contains some incorrect matching correspondences with low quality. Therefore, we incorporate the normal-depth consistency based on the orthogonality between normal and local surface tangent.



2. Multi-metric Loss: We propose a novel multi-metric loss function by considering different perspectives of matching in feature correspondence beyond pixel.

2.1 Pixel-wise Loss: For the pixel-wise loss, we only consider the photometric consistency between the reference image I_{ref} and other source images. There are mainly three parts of this loss function: photometric loss, the loss of structure similarity, the smooth of the final depth map.

2.2 Feature-wise Loss: The pixel-wise loss performs mismatch errors in some challenging scenarios. Therefore, one of the main improvements of M3VSNet is the use of feature-wise loss, which will utilize more semantic information for matching correspondences.



We extract the layer 8, 15 and 22. For every feature from the VGG16, we construct the loss based on the concept of crossing multi-views. The final feature-wise loss function is a weighted sum of different scales of features, which raises the robustness and completeness of point cloud reconstruction.

Result

1. DTU dataset: Quantitative results on the DTU's evaluation set. Three classical MVS methods, two supervised learning-based MVS methods and three unsupervised methods using the distance metric (lower is better) are listed.

Method	Mean Distance (mm)		
	Acc.	Comp.	overall.
Furu	0.612	0.939	0.775
Tola	**0.343**	1.190	0.766
Colmap	0.400	**0.664**	**0.532**
SurfaceNet	0.450	1.043	0.746
MVSNet(D=192)	**0.444**	**0.741**	**0.592**
Unsup_MVS	0.881	1.073	0.977
MVS^2	0.760	**0.515**	0.637
M^{3}VSNet(D=192)	**0.636**	0.531	**0.583**

Qualitative comparison on the DTU dataset. From left to right: ground truth, MVSNet, M3VSNet without feature-wise loss and M3VSNet

M3VSNet establishes the state-of-the-arts unsupervised learning methods for multi-view stereo reconstruction.

2. Tanks & Temples dataset: The generalization Ability. The ranking in the Leaderboard of the intermediate Tanks and Temples benchmark shows that M3VSNet is the best unsupervised MVS network until August 30, 2020.