Abstract:

Existing deep methods produce highly accurate 3D reconstructions in stereo and multiview stereo settings, i.e., when cameras are both internally and externally calibrated. Nevertheless, the challenge of simultaneous recovery of camera poses and 3D scene structure in multiview settings with deep networks is still outstanding. Inspired by projective factorization for Structure from Motion (SFM) and by deep matrix completion techniques, we propose a neural network architecture that, given a set of point tracks in multiple images of a static scene, recovers both the camera parameters and a (sparse) scene structure by minimizing an unsupervised reprojection loss. Our network architecture is designed to respect the structure of the problem: the sought output is equivariant to permutations of both cameras and scene points. Notably, our method does not require initialization of camera parameters or 3D point locations. We test our architecture in two setups: (1) single scene reconstruction and (2) learning from multiple scenes. Our experiments, conducted on a variety of datasets in both internally calibrated and uncalibrated settings, indicate that our method accurately recovers pose and structure, on par with classical state of the art methods. Additionally, we show that a pre-trained network can be used to reconstruct novel scenes using inexpensive fine-tuning with no loss of accuracy.