Please fill in the following information after you read the paper.
Name: Darshit Miteshkumar Desai; UID: 118551722; Dir id: darshit

[Paper tile] Unsupervised Learning of Depth and Ego-Motion from Video (SfMLearner)

[Summary] Describe the key ideas, experiments, and their significance.

The key idea of the paper is to make use of un-supervised learning paradigm to extract pose and monocular depth from a video sequence. It is also dubbed as SfMLearner since in the classical version of SfM we also try to solve the chicken and egg problem of finding camera pose and the sparse 3d point cloud based on the neighboring views as anchors using a correlation to set those neighboring views. In this scenario, the authors take inspiration of doing a learning-based implementation from the vision of human beings. They propose that humans learn to understand the 3D world using their eyes throughout a life time of experience in 2D projections and given a single frame of image they are able to segment depth, shape and normals of surfaces. They plan to do the same using a learning-based method. This is one of the seminal ideas which gave birth to a number of branching architectures for unsupervised learning.

The authors formulate a problem in two steps, they first assume that given a 3D representation of a scene if we have the camera poses and the 3d points we can easily back project the 3D scene onto a 2D image using that knowledge to synthesize many different views. At the same time if the reverse were to happen where we had the novel views and wanted to synthesize the camera poses and the depth then that would become an ill posed problem. They propose that this can only be overcome by predicting two things from a learning-based algorithm, camera poses as well as a 3D representation, in the context of this paper a 2.5D depth image.

Their network architecture is simple, they use a dispnet (or Unet) shaped network layer arrangement to predict depth from a single view image. In a parallel network they input a sequence of images in a similar architecture where the intermediate output of the network is the 6-DOF pose and the fully deconvolved same resolution output is a explain-ability mask. The authors also model certain limitations. They take an assumption of having a static scene without moving objects and also assume that there is no occlusion between neighboring source and target views. This is done by adding a explainability prediction network that outputs a per pixel soft mask that signifies network's belief in where direct view synthesis will succeed. Since there is no way to directly supervise this, the loss of this network is added as a regularization term. The authors also take into consideration gradient locality, this was addressed by adding a small bottleneck for the depth network that helps in constraining globally smooth gradients.

In the experiments, the author's mention that the algorithm was trained on cityscapes dataset and the KITTI dataset was used for fine tuning and benchmarking. Both of which are datasets in which a camera was mounted on a car and the poses are available too. The single view depth estimation results are comparable with other methods that use some form of depth or camera poses as supervision. On KITTI dataset it was found that the unsupervised method performs comparably well with other supervised methods. The method fails only in case where another paper considers left-right consistency loss for training. This is identified by authors as a future work direction. The method is also tested for generalization on Make3D dataset and when compared to methods which were supervised on that dataset during training, the results show that the author’s method is able to capture scene layout reasonably well. The pose estimation output is tested with ORB SLAM method which uses monocular data for pose estimation, for short frame sequences the method outperforms reasonably well when the same is compared to orbslam. But for long sequences where orb slam uses relocalization and loop closure the performance of SfMlearner deteriorates severely.

[Strengths] Consider the aspects of key ideas, experimental or theoretical validation.

The strengths of this method are mentioned below:

1) This is first in the class method to perform unsupervised learning for both depth and pose estimation on videos
2) The method is trained on unlabeled videos and yet performs comparably well for both depth and pose estimation when compared to approaches which use ground truth depth and pose for training

The method doesn’t have many strengths, since it was one of the first and seminal papers to do end to end unsupervised learning. One of the strengths could be that this work inspired a lot of follow up work and the paper writing by authors highlighted a lot of future work for researchers to work on.

[Weaknesses] Consider the aspects of key ideas, experimental or theoretical validation, writing quality, and data contribution (if relevant). Explain clearly why these are weak aspects of the paper

The disadvantages and weaknesses are as mentioned below:

1) During experimentation on KITTI dataset the model mistakes cars/bushes as distant objects since it was trained on cityscapes only.
2) The method also struggles in a typical scenario where large open space scenes are there or the objects in the scene are very close, in both the cases the method is not able to render depth properly
3) The method is also unable to predict thin structures as well since the explainability mask gives a low confidence on such objects and makes them unexplainable
4) The current framework doesn’t explicitly estimate moving obstacles and scene dynamics as well as occlusion
5) Also in this algorithm the authors used depth map which is a simplified representation of the 3d scene, a more volumetric reconstruction approach is required

[Reflection] Share your thoughts about the paper. What did you learn? How can you further improve the work?

After reading this paper, I got the intuition that most of the outside world cannot especially be mapped for ground truth which can be used for 3d reconstruction. That is why such unsupervised learning methods are critical for capturing such depth data for purposes like autonomous driving and navigation. There was a lot of future work which was highlighted by the authors. One direction was to explore the application of this method to be used simply as a segmentation and detection architecture. Another area could be to improve the method’s ability to work for internet scale videos where the camera intrinsic matrix isn’t always available. Another idea could be to use motion segmentation as a potential solution to account for dynamics of scenes like occlusion and dynamic motion of objects in the scene.