Paper review for "Ranjan et al: Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation"

EECS 598 Paper Review - Week 5 - Changyuan Qiu

This paper explores jointly solving 4 fundamental vision problems unsupervisedly: **single view depth prediction, camera motion estimation, optical flow and motion segmentation**, considering that they are all coupled by geometric constraints and motion of the scene. To address this joint unsupervised learning problem, the authors introduce the brand-new **Competitive Collaboration** (CC) learning framework, which is formulated as a three-player game consisting of two players competing for a resource that is regulated by the third player, or moderator, while in the meantime the two players collaborate to train the moderator to reach specific goals. In this paper's CC framework, the two competing players are a static scene reconstructor and a moving region reconstructor which compete by reasoning about static-scene and moving-region pixels in an image sequence, and the competition is moderated by a motion segmentation network that segments the static scene and moving regions and distributes training data to the competitors; and the two competitors collaborate to train the moderator such that it classifies static and moving regions correctly in alternating phases of the training cycle.

Many prior works in this field have approached estimation of these problems by coupling two or more problems together in an unsupervised learning framework, including (1) using an explainability mask to exclude evidence that cannot be explained by the static scene assumption introduced by Zhou el al. [1] which combines **depth and motion**; and (2) estimation of residual **optical flow** using a refinement network introduced by Yin et al. [2]. However, all these methods perform poorly on **depth and optical flow** benchmarks. And this paper first couple **motion segmentation** with other objectives and this allows the networks to use geometric constraints where they apply

and optical flow where they do not, and they demonstrate that this joint training results in top performance among unsupervised methods for all subproblems. They evaluate their model on all 4 tasks on the KITTI 2015 dataset and get very good results: they achieve SOTA performance on **single view depth prediction** and **camera motion estimation** among unsupervised methods; achieve SOTA performance on **optical flow** among unsupervised joint methods that solve more than one task; and introduced the first baseline for unsupervised **motion segmentation**. They also show the effectiveness of the CC framework by ablation study on depth prediction and optical flow.

Regarding limitations of this paper, one thing worth mentioning is that the author performs all evaluation on the KITTI dataset, which mainly consists of automotive images. Whether the motion segmentation trick works well for more general scenes or merely for automotive cases is under doubt.

Reference

[1] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsuper-vised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017.

[2] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1983–1992, 2018.