Review for Self-Supervised Monocular Scene Flow Estimation

Irem Arpag

March 21, 2021

The paper focuses on a challenging issue of autonomous navigation, scene flow estimation, task of obtaining 3D structure and 3D motion of dynamic scenes. The approaches presented so far for scene flow estimation have their own limitations such as stereo calibration is needed for stereo camera, LIDAR (point clouds) is requiring expensive sensing devices or RGB-D camera is restricted to indoor usage that encourages the authors to propose a novel deep learning approach by monocular 3D scene flow to overcome the limitations of previous works. Although using monocular camera is an advantage regarding not requiring calibration, being flexible in usage and low cost, monocular scene flow is a highly ill-posed problem in terms of depth-scale ambiguity, determining 3D structure and motion, and occlusion problems. By taking an inverse view to solve ill-posed problem, the researchers design a single convolutional neural network, through using a standard, state-of-art, optical flow pipeline "PWCNet" by decomposing an optical flow cost volume into scene flow and depth exploiting a single joint decoder, not separate decoders for each task as in multi-task CNN methods which is complicated the processes, with the preference of self-supervised manner because of difficulty of obtaining 3D motion ground truth data and over-fitting problems of synthetic data. The significance of the approach comes from their argument of achieving good accuracy and fast-run time at the same time since none of the CNN based multi-task learning approaches accomplish them together.

Competitive accuracy, fast-run time, direct 3D scene flow estimation, resolving occlusion problem, simple and stable training set are the strong sides of the model. The first evaluation results show that the model increases the accuracy by 34 percentage when comparing previous unsupervised and self-supervised CNN models of monocular scene flow and semi-supervised fine-tuning of the said model improves the accuracy further to comparable level of semi-supervised methods. Simple and stable training is the result of exploiting a single joint decoder that brings plain training setup and better accuracy when comparing with multi-task learning approaches' multiple modules and complex training schedules.

For evaluation KITTI raw dataset, for scene flow experiment KITTI split and for fine-tunning KITTI Scene Flow Training Datasets are used and they also utilize Eigen Split for valuating monocular depth accuracy. Since data augmentation has a vital importance for achieving good accuracy with a limited training data, in the experiment the authors both use photometric augmentations and geometric augmentations to achieve monocular depth estimation and scene flow estimation together and also by adapting CAM-Convs that provides estimation depth without using camera intrinsic, they ensure good accuracy for augmented images. For self-supervised training they use Adam with 400k iterations. The practicality of the model comes from just be trained from stratch all at once and not to need any complex training strategies when comparing previous works that require stage-wise pretraining or iterative training.

As ablation study, they apply proxy loss for the purpose of reconstructing the reference image as closely as possible to prevent incorrect estimation of disparity and scene flow in the occluded areas resulted of which show that 3D points loss is the most contributing one by achieving more accurate disparity on the target image. To confirm the priority motivation of the model that is decomposing optical flow cost volumes into depth and scene flow by using a single decoder they compare single decoder with separate decoders for each task. The oblation study demonstrates that single decoder solves the imbalance and stability problem by achieving higher accuracy. Comparison of the said model shortly called Self-Mono-SF against the state-of-art multi-task CNN methods, it performs better by large margins, for example it achieves more than 40.1 percentage accuracy gain for estimating the disparity on the target image on KITTI Scene Flow Training.

Qualitative results on the KITTI Scene Flow 2015 Benchmark, Self-Mono-SF (f+) (fine-tuning) goes beyond Self-Mono-SF with its more than 400x faster run time.

The model achieves state-of-art Scene Flow accuracy among unsupervised and self-supervised monocular methods and also with fine-tuned modification it creates significant fast by proving a precise basis for future work such as an adversarial model [1] that propose a different metric learning approach for self-supervised scene glow estimation.

REFERENCES

[1] V. Zuanazzi, J. Vugt, O. Booij, P. Mettes, Adverserial Self Supervised Scene Flow Estimation, ArXiv, Nov 2020.