Review for GA-Net: Guided Aggregation Net for End-to-end Stereo Matching

Irem Arpag

February 28, 2021

The model called "GA Net" shortly is aiming at improving and accurate disparity estimation map in problematic regions especially in reflective areas such as car surfaces and tin objects in autonomous driving. The work is based on the idea that since future-based matching is ambiguous in stereo matching in computer vision because of a/m real world problems, the authors of the paper choose to use cost aggregation process and propose a two layers neural network targeting to capture local and whole-image cost dependencies respectively (local + global method together). The first of the layers is semi global aggregation (SGA) that provides exact estimations in problematic areas by aggregating the matching costs in different directions over the whole-image, and the second one is local guided aggregation (LGA) which is a traditional cost filtering strategy dealing with thin structure and object edges resulted from down and up sampling layers.

With the proposal of these two layers, end-to-end stereo matching method, it is claiming to replace the use of 3D convolutional layers, the pioneer of it is state-of-art "GC Net" method that achieves to include cost aggregation in the training pipeline and the other significant model is "PSM Net" Network that improves the accuracy further by implementing the stacked hourglass backbone. The first visualization results of the experiment apparently show that "GA Net" with only 2 GA layers and 2 3D conv. layers performs better accuracy than "GC Net" with 19 3D conv. layers. Besides decreasing memory and computation costs, another positive point of the model is that one GA layer has only 1/100 computational complexity in terms of FLOPs comparing to 3D convolution with a speed of 15 20 fps results in building a real time "GA Net" model in the literature achieving perfect efficiency on both the Scene Flow Dataset and KITTI benchmarks.

As ablation study, performance of "GA Net" is evaluated by using different number of GA layers (0-4) in different base line settings that have only traditional 3D conv. layers. For example, the new architecture takes the best 3-pixel threshold error rate with its 3 SGA layer and 1 LGA layer on KITTI 2015 set. The answer to the question how the model provides so efficiency is that since 3D conv. layer has fixed weights, it performs the same the all locations in the whole image that cause loss of details especially in the corners of the objects, on the contrary the GA-Net's SGA layer is guided by not fixed by variable weights in different geometrical locations which creates effectiveness. When comparing both qualitative and quantitative results of "GA Net" with "GC Net" and "PSM Net" after training of Scene Flow, KITTI 2012 and 2015 test sets, the set model goes beyond current best "PSM Net" in all the evaluation metrics and takes the first rank on leader board and accepted generally as the most reliable article seen after "PSM Net" [1]. The model provides the correct matching information exactly in three difficult regions: large untextured areas, reflective areas, and target boundaries.

In the literature, there are further works that take "GA Net" as a reference or base for their research, one of them is an interesting paper [2] comparing "GA Net" with traditional SGM by applying on areal and satellite data instead of street-view dataset KITTI 2015 benchmark. The results of the study confirm the success of "GA Net" and shows the potential of the deep learning for solving such problems.

The paper is a well-written, easy to understand and a comprehensive one with its outstanding performance results.

1 REFERENCES

[1] http://programmersought.com/article/72364071691

[2] XiaY, d'Angelo P, Tian J, Reinartz P. (2020 edition) Dense Matching Comparison between classical and deep learning based Algorithms for Remote Sensing Data.