A two-view network to predict depth and ego motion from monocular sequences

January 17 2019, by Ingrid Fadelli

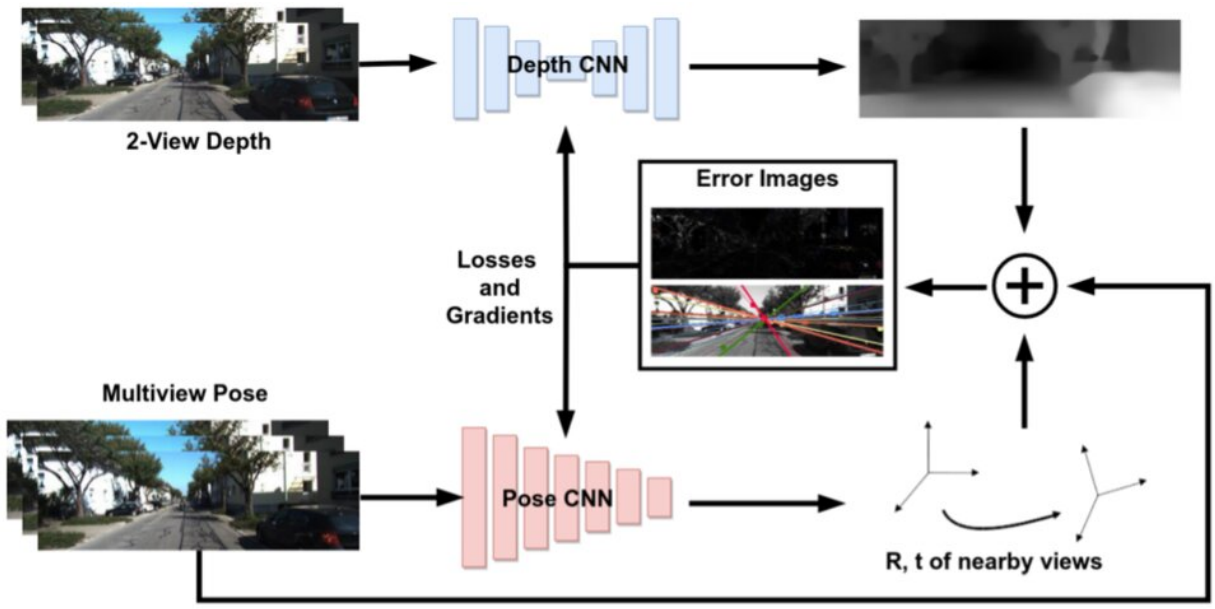

Figure 1: Overview of the training procedure. The Depth CNN predicts the inverse depth for a target view by taking in the target view and a nearby image as the input. The Pose CNN predicts the relative poses of the source views from the target, which are then warped into the target frame using the relative poses and the scene depth and the photometric errors between multiple source-target frame pairs are minimized. These are weighted by the per-pixel epipolar loss.

Credit: Prasad, Das & Bhowmick.

Researchers from the Embedded Systems and Robotics group at TCS Research & Innovation have recently developed a two-view depth network to infer depth and ego-motion from consecutive monocular sequences. Their approach, presented in a paper pre-published on arXiv, also incorporates epipolar constraints, which enhance the network's

geometric understanding.

"Our main idea was to try and predict pixel-wise depth and camera motion directly from single image sequences," Dr. Brojeshwar Bhowmick, one of the researchers who carried out the study, told TechXplore. "Traditionally, structure from motion-based reconstruction algorithms provide sparse depth outputs for salient points of interest in the image, which are tracked over multiple images using multi-view geometry. With deep learning gaining popularity in computer vision tasks, we thought of leveraging existing methods to help our cause by approaching the problem in a more fundamental manner using a combination of concepts from epipolar geometry and deep learning."

Most existing deep learning approaches to predict monocular depth and ego motion optimize the photometric consistency in image sequences by warping one view into another. By inferring depth from a single view, however, these methods might fail to capture the relation between pixels and thus to provide proper pixel correspondences.

To address the limitations of these approaches, Bhowmick and his colleagues developed a new approach that combines geometric computer vision and deep-learning paradigms. Their approach uses two neural networks, one for predicting the depth of a single reference view and one for predicting the relative poses of a set of views with respect to the reference view.

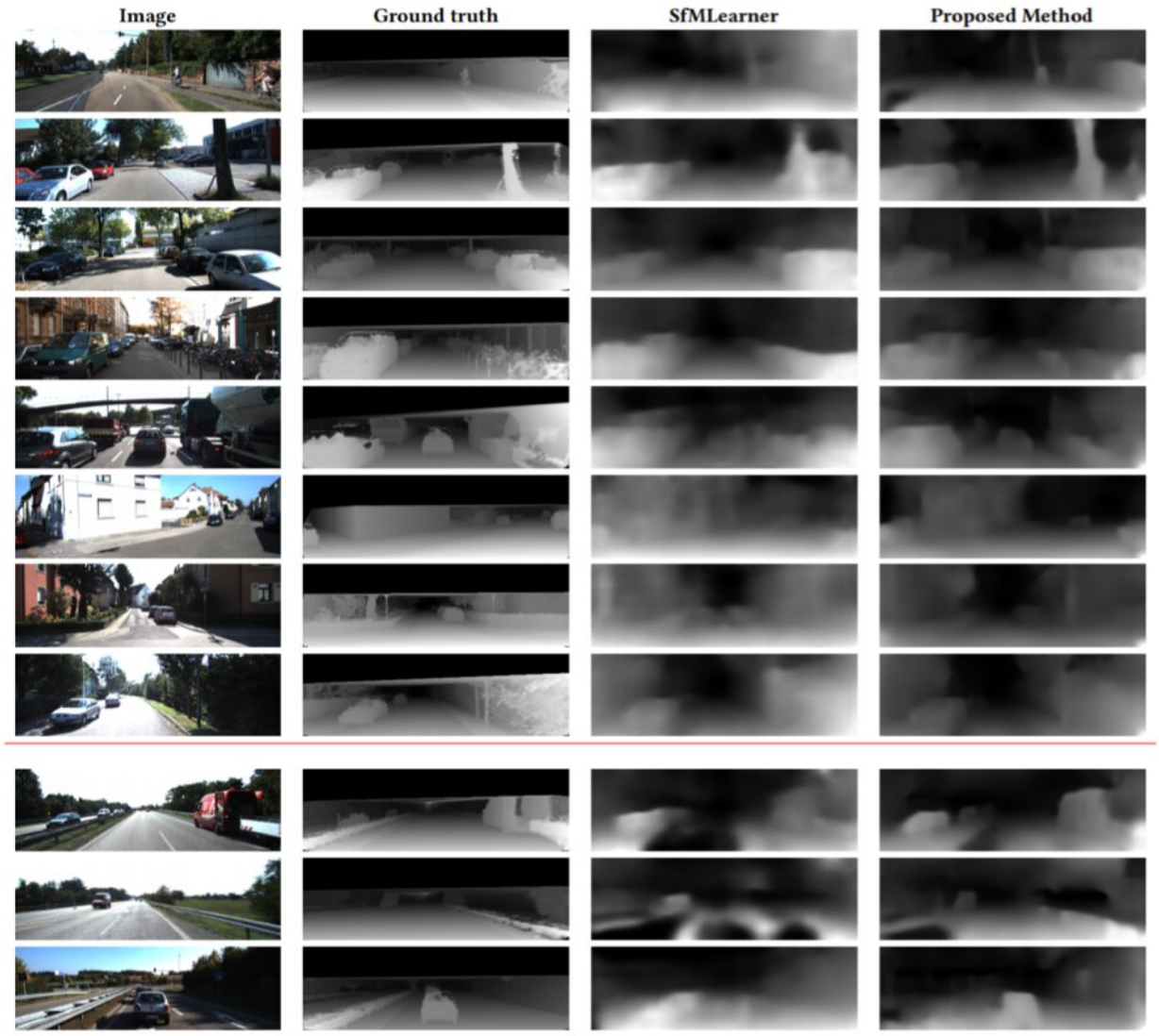

Figure 2: Results of depth estimation compared with SfMLearner. The ground truth is interpolated from sparse measurements for visualization purposes. Some of their main failure cases of SfMLearner are highlighted in the last 3 figures, such as large open spaces, texture-less regions, and when objects are present right in front of the camera. As it can be seen in the last 3 figures, our method performs better, providing more meaningful depth estimates even in such scenarios. (Pictures best viewed in color.)

Credit: Prasad, Das & Bhowmick.

"The target image scene can be reconstructed from any of the given poses by warping them based on the depth and relative poses," Bhowmick explained. "Given this reconstructed image and the reference

image, we calculate the error in the pixel intensities, which acts as our main loss. We add the novelty of using the per-pixel epipolar loss, a concept from multi-view geometry, in the overall loss, which ensures better correspondences and has the added advantage of discounting moving objects in the scene that can otherwise deteriorate the learning."

Rather than predicting depth by analyzing a single image, this new approach works by analyzing a pair of images from a video and learning inter-pixel relationships to predict depth. It somewhat resembles traditional SLAM/SfM algorithms, which can observe pixel motions over time.

"The most meaningful findings of our study are that using two views for predicting the depth works better than a single image, and that even weak enforcement of pixel level correspondences via epipolar constraints works nicely," Bhowmick said. "Once such methods mature and improve in generalizability, we could apply them for perception on drones, where one would want to extract maximum sensory information by consuming as little power as possible, which can be achieved by using a single camera."

In preliminary evaluations, the researchers found that their method could predict depth with higher accuracy than existing approaches, producing sharper depth estimates and enhanced pose estimates. However, currently, their approach can only perform pixel-level inferences. Future work could address this limitation by integrating semantics of the scene into the model, which might lead to better correlations between objects in the scene and both depth and ego-motion estimates.

"We are further probing into the generalizability of this method and other similar methods on various scenes, both indoor and outdoor," Bhowmick said. "Currently, most works perform well on outdoor data, such as driving data, but perform very poorly on indoor sequences with

arbitrary motions."

More information: Epipolar geometry based learning of multi-view depth and ego-motion from monocular sequences. arXiv:1812.11922 [cs.RO]. arxiv.org/abs/1812.11922