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

[Paper tile] NICE-SLAM: Neural Implicit Scalable Encoding for SLAM

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

The paper discusses a real-time, robust and scalable learning-based SLAM pipeline which takes advantage of Neural implicit representations and uses RGB-D images as input to map and track the scene and camera poses respectively. The key idea of the paper is to utilize hierarchical feature grids and a small Multi-layer perceptron which is pre-trained from a convolutional occupancy network to perform occupancy predictions from an input RGB-D image. The grid based neural encoding of scenes discretizes the scene updates into smaller local scene updates instead of a complete global scene update which in turn increases the accuracy of the SLAM algorithm. This method is extensively tested on multiple datasets which cover dynamic objects, large scenes and compared with their similar architectures described in recent works.

In the related recent works, the use neural implicit representations have been done to implement SLAM. One such example given by the authors was the iMAP SLAM which takes advantage of the global feature vector to predict occupancy probabilities of the set of 3d points. But this method has some serious disadvantages. When given a large-scale scene the performance of this algorithm significantly deteriorates both in the scene reconstructions and camera tracking domains. To overcome this limitation, authors came up with a newer version of the 3d representation called convolutional occupancy networks which take advantage of the grid-based features to record geometric details and enable scene reconstruction in much finer detail. Unfortunately, this method cannot be applied in real-time which is the main requirement of a SLAM system.

The pipeline is simple an RGB-D sequence is given as an input and it outputs a camera pose and a learnable hierarchical feature grid. In the pipeline (See figure 2 in paper) the mapping is done via backpropagating through the hierarchical feature grid and the tracking is done via backpropagation to update the camera poses. The ray point sampler queries 3d points during forward pass and outputs occupancy probabilities and color prediction for the points along that ray. The same is aggregated during volume rendering to output a predicted depth and rgb image. An L1 loss is used to define geometric and photometric differences and a combination of them is used to optimize mapping and tracking. To account for dynamic objects the authors, consider the loss values as threshold and if it exceeds a certain amount the pixels are filtered and masking those pixels during mapping.

The authors test on various datasets to prove their algorithm’s performance. They test on 5 different datasets and account for metrics like ATE-RMSE, completion ratio, completeness and accuracy. The testing on replica dataset signifies that NICE SLAM produces sharper geometry with less artifacts. The testing on TUM-RGBD and Scannet datasets prove that the proposed method can handle large scale 3d scenes where other counter parts of learning based slam fail or are outclassed by NICE SLAM. It also shows that the difference between classical slam and learning based slam accuracy levels like

ATE-RMSE is reduced. It is also shown that the method is robust to dynamic objects and has the capability of filling holes and incomplete geometry. In conclusion a dense Visual slam method is proposed and tested while there are significant future improvements which can be made to improve and match the accuracy levels of classical SLAM

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

The advantages/strengths can be highlighted as follows:
1)The method is real-time and can handle large scale scenes as seen in the experimentation section of the paper. This is further proved from the architecture where the features are discretized into coarse, mid and fine level so that the updates based on backpropagation occurs in targeted voxel grids
2)The slam method also accounts for dynamic objects and effectively filters them out.
3)When compared to scene reconstruction methods which use a single large Multi-layer perceptron, the experimental validation of a grid-based representation+tiny MLPs provides much more finer level detail for mapping as well as a higher accuracy in camera pose tracking.
4)The added advantage of this architecture is that the processing time is much faster and this benefits local scene updates as seen from the experimental results which show an improvement of 2times and 3times in tracking and mapping speeds
5)Overall the use of neural implicit representations with a hierarchical feature grid shows the ability to fill small holes and at the same time extrapolate larger scenes like walls and other unobserved regions which not only improves mapping but also stabilizes tracking.

[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 limitations or weaknesses can be highlighted as follows:

1) Although the algorithm shows improvements when compared to other learning based algorithms, it is not able to match the state of the art accuracy and other SLAM benchmarks of the classical SLAM algorithms.
2) The limitation of this algorithm is that it isn't able to perform loop closures. Loop closures are of vital importance as it helps in reducing the cumulative error of the robot's estimated pose and generate a consistent global map.
3) Another major disadvantage of NICE SLAM is the generalizability of the algorithm. Since the tiny MLPs weights are frozen throughout the SLAM execution it might not be able to map and reconstruct novel scenes for which the MLP was not trained.

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

There were many important learnings which were achieved from this paper readings

1) The first was the understanding of architecture of vanilla neural implicit representations. This further led me to deep dive into the pointnet architecture which is used for segmentation and classification of problem of 3d point clouds

2) I also learnt about the convolutional occupancy networks and the use and design of hierarchical feature grids to localize and store features to make the reconstruction problem translationally equivariant.
3) The main learning of this paper was to understand the different metrics required to test and validate the slam algorithms along with the use cases and knowledge of different Visual SLAM datasets out there.

The further improvement on the work as described in the weaknesses could be to implement loop closure and also to match the accuracy levels as closely as possible to the conventional slam algorithms. Further deep dive into the papers which followed the NICE SLAM shows that there are improvements still ongoing which cover some disadvantages of NICE SLAM but has their own weaknesses. (Referring to a recent paper called E-SLAM)