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

[Paper tile] IronDepth: Iterative Refinement of Single-View Depth using Surface Normal and its Uncertainty

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

Motivation: Monocular Depth Estimation (MDE) has a wide variety of applications like autonomous driving, augmented reality, 3D reconstruction etc. The authors in this paper focus on two problems:1)Poor generalizability of existing MDE methods; 2)Poor surface normal accuracy from estimated depth maps. The key idea is that although SOTA methods show low perpixel depth errors, when their surface normals are recovered they do not capture the exact scene geometry. One such example is the walls not being flat in the reconstructed point cloud from the depth map.
Why do MDE method fail to compute normals with higher accuracy? One reason is that the training data is imbalanced, key example is the NYUv2 dataset whose histogram distribution of depths shows majority of pixels contain depth values ranging from 1 to 4 mtrs. Another reason is the view dependent inference of the depth, the depth gradient is non-uniform for a perspective camera, thus making the CNN based regression highly inaccurate since CNN tries to detect features which are translationally equivariant.
Key ideas: The authors propose Irondepth a iterative refinement method to estimate the depth using surface normals. Given a pixel's depth and normal, a neighborhood is defined in 3D space and a pixel is queried from that neighborhood, then using a technique called normal guided depth propagation, the depth of the query pixel is refined iteratively such that it can fit into that plane. This depth refinement method is formulated using a classification task to choose candidate pixels for plane fitting and depth refinement.
Pipeline and Architecture: The authors proposed a pipeline which takes a single RGB image and known camera intrinsics as input. The first step is to estimate surface normals and its aleatoric uncertainity. This was explained in their previous work where they define the importance of using aleatoric uncertainity in estimating surface normals. Parallely a D-Net or a convolutional encoder decoder generates three outputs, a inital estimated depth map, a contextual feature map and a hidden feature state. This is inturn combined with the context features, surface normal confidence (inverse of aleatoric uncertainity), and hidden state to give the updated hidden state. This is done using a Convolutional Gate recurrent unit. This module is important since this and the normal guided depth propagation combined form the classification architecture. From the updated hidden state two more outputs are derived using encoder-decoder CNNs Pr-Net and Up-Net respectively. The Pr-Net derives weights which would help in classifying the depth candidates from which the depth is to be propagated. The candidate depth pixels are calculated from the initially estimated surface normals and a set of candidate depth for a queried neighborhood pixel is generated. This is then combined with in the form of weighted sums to get the refined depth map. The Up-Net weights are then used to upsample the image back to the original resolution.
Results and Experiments: The model was tested on NYUv2 and achieved SOTA performance. The model also outperforms on cross validation with iBims-. Further the model also has a good generalization ability which was further tested on ScanNet. Middlebury2014 and KiTTI. The inference speed was tested on a 2080Ti GPU.

Applications: 1) The authors claimed that the method developed can be used to refine depth maps generated by existing depth estimation methods, this was tested with other MDE methods like TransDepth, Adabins etc which all show good improvement sin the depth and normal accuracy. This also answers an important question, that if the improvement in per pixel depth value accuracy is minimal what is the advantage of this method? The first part of this question is answered by the authors in the first application. The second crucial application is depth completion tasks where a sparse point cloud obtained from a ranging sensor could be used to perform depth completion wherein the point cloud points would act as anchor points (having high confidence) and help in performing normal guided depth propagation.

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

1)The method achieves state of the art performance in the NYU dataset and when cross validated with other data sets it also achieves good performance showing that the method is generalizable.
2)Using the normal guided propagation, the depth accuracy of any other MDE method can be improved, this is a major advantage as a sophisticated model in future could use this method to improve it’s accuracy
3)Method can be used for depth completion tasks which could use LiDaR data points to complete the depth map
4)The method achieves and captures better scene geometry when compared with the point clouds generated by other MDE methods which shows the effectiveness of the normal guided approach.

[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

1)Although the paper contributes significantly in the domain of scene reconstruction, depth completion, the main task of MDE is to estimate Depth with high accuracy which this method doesn’t show any significant improvement. The results show comparable depth metrics where although the Normal error and accuracy showed great improvements the depth error and depth accuracy remains almost the same.
2)Another weakness could be the method’s iterative nature, and its use in realtime applications although the author claims that the method runs in 66.26 ms, no further details are given on what test set this latency was tested on or what type of computations were used.

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

Although the overall concept of MDE methods in general seems a lot like black box, but the method in this paper made the use of a little bit of mathematics like using normals to refine depth pixels we get from it. The key things I learnt after reading this paper was the use of Recurrent networks, Gated Recurrent Units and ConvGRUs in the computer vision domain. Before reading this I only knew about the use of RNNs or GRUs in NLP but this was a novel application. I also learnt some new terminologies like aleatoric and epistemic uncertainties. Apart from this to improve this a sophisticated neural network model could be designed specifically to use the normal guided model to improve depth estimation. Other than this a general effort on newer MDE methods could be counted as future work.