Monocular Depth Estimation using Atrous Convolutions: A Survey

Darshan Rajopadhye[1], Pravin Shelke[2], Apoorva Dhalpawar[3], Prof. V. V. Waykule[4]

[1,2,3]Undergraduate Students, Dept. of Computer Engineering, AISSMS COE, Pune, Maharashtra, India
[4]Professor, Dept. of Computer Engineering, AISSMS COE, Pune, Maharashtra, India

---***---

Abstract - *Estimating the depth from any image refers to the operations and the algorithms that extracts the spatial structures of that scene. To make it more simpler, it refers to the task of obtaining the measure of distance of, ideally, every point in the image. Getting the exact information of the relative spacing of objects is an important feature required in many forthcoming technologies such as Robotics, AR and autonomous driving.*

The traditional method to obtain the depth measures from any image involves the use of stereo pairs, i.e Left Right pairs of the image. Now this requirement in itself is too much to ask from the use in this era of ease, but these models also require other things such as a high amount of ground truth depth data. This type of data is very rare to the public and is not freely available as it requires high quality sensors for collecting the ground truth data.

Monocular depth estimation, as the name suggests, uses only the single view images from the users to correctly identify the depth measures as efficiently as the traditional systems. This is achieved by not treating the problem as supervised but an unsupervised learning problem. In this paper we propose a new network structure for the existing unsupervised learning method for monocular depth estimation by using the atrous convolutions which are proven to increase the efficiency of object detection models.

Key Words: Monocular Depth Estimation, Supervised Learning, Unsupervised Learning, Ground Truth Data, Depth Maps.

1. INTRODUCTION

Estimating depth from images is one of the basic and most important tasks of computer vision and thus as a result there has been a lot of work and advancement in this field. There have been various methods that overcame the previous versions and then later were succeeded by more accurate methods.

We focus on the deep learning based unsupervised method, to be more precise, the one proposed by Godard et al [1] and check whether the atrous convolutions , which have already been proven successful in semantic segmentation [2][3] make any impact on the model in terms of the performance metrics.

1.1 Atrous Convolutions

Atrous convolutions allow to explicitly control the receptive field size of a filter without introducing additional parameters. This is achieved by skipping pixels at a certain rate (also called filter dilation), and thus inserting zero holes (french: trous) into the filter (see Fig 1). This helps with aggregating information from different spatial scales, while also allowing for larger feature maps, as less downsampling is required in order to achieve a large field of view (FOV).

Dilated convolutions introduce another parameter to convolutional layers called the dilation rate. This defines a spacing between the values in a kernel. A 3x3 kernel with a dilation rate of 2 will have the same field of view as a 5x5 kernel, while only using 9 parameters. Imagine taking a 5x5 kernel and deleting every second column and row.

This delivers a wider field of view at the same computational cost. In figure we can see the base matrix is the input image from which the highlighted features are considered and computed to be the feature in the top matrix which is the dense feature matrix. These highlighted features from the input image shift to the right based on the value of input stride until all the features are considered. So we get a 3X3 dense matrix, equivalent to the 7X7 input matrix when the input stride is 1.

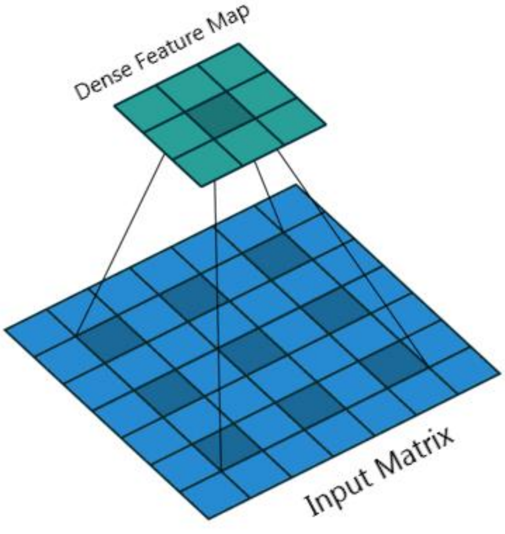

Fig -1: 2D convolution using a 3 kernel with a dilation rate of 2 and no padding

2. RELATED WORK

[1] “Unsupervised Monocular Depth Estimation with Left Right Consistency” by Clément Godard, Oisin Mac Aodha and Gabriel J. Brostow. The main findings from their work are that they use an unsupervised approach for single image depth estimation without use of ground truth values. They exploit epipolar geometry constraints and generate disparity images by training the network with an image reconstruction loss. They proposed a novel training loss that enforces consistency between the disparities produced relative to both the left and right images.

[4] “Digging Into SelfSupervised Monocular Depth Estimation” by Clément Godard, Oisin Mac Aodha, Michael Firman and Gabriel J. Brostow. A new method is used, Auto masking: The mask prevents objects moving at similar speeds to the camera (top) and whole frames where the camera is static (bottom) from contaminating the loss. A minimum reprojection loss, designed to robustly handle occlusions. A full-resolution multi-scale sampling method that reduces visual artifacts. The model is trained by minimizing the image reconstruction error.

[5] “Multi-scale Deep CNN Network for Unsupervised Monocular Depth Estimation” by Wan Yingcai, Fang Lijing and Zhao Qiankum uses unsupervised method to get left-right disparity map by left-right consistency training of left-right image of stereo camera. The DispNet, a coding decoding process and is a classical single scale network structure is implemented. Image wrap is applied with predicted disparities to reconstruct left and right images.

[6] “Estimating Depth from Monocular Images as Classification Using Deep Fully Convolutional Residual Networks” by Yuanzhouhan Cao, Zifeng Wu and Chunhua Shen uses pixel-wise classification for depth formulation instead of regression. They discretize ground truth depths into several bins and apply classification by training a fully connected deep residual network. Semantic Segmentation using fully connected random fields(CRF) is used. Batch normalization and ReLU layers are performed between convolutional layers.

[7] “Monocular Depth Prediction using Generative Adversarial Networks” by Arun CS Kumar, Suchendra M. Bhandarkar and Mukta Prasad use Monocular reconstruction i.e depth map and pose prediction from video sequences using adversarial learning. Using estimated depth and pose parameters, the generator transforms source images which are then interpolated using Spatial Transformer Networks to output the generated pair of images. Discriminator subnetwork learns to differentiate real and generated images.

[8] “Depth Estimation From a Single Image Using Guided Deep Network” by Minsoo Song and Wonjum Kim uses latent space of depth-to-depth network, which contains useful encoded features for guiding the process of depth generation.Color-to-depth network via loss and use of autoencoder which teaches decoders to learn the process of generation instead of classification. Gradients play an important role in guided networks because encoded features contain geometrical structures.

[9] “A 3D Convolutional Neural Network For Light Field Depth Estimation” by Agota Faluvegi, Quentin Bolsee, Sergiu Nedevschi, Vasile-Teodar Dadalat and Adian Munteanu uses fully convolutional 3D neural network that estimates disparity in light field images. Proposed method is parametric, lightweight and less prone to overfitting. Reflections, Refraction and noisy background can be handled with a post processing filter.

[10] “High Quality Monocular Depth Estimation via Transfer Learning” by Ibraheem Alhashim and Peter Wonka initializes encoders along with augmentation and training strategies that lead to accurate results. Encoder-Decoder architecture with skip connections. Transfer learning gives better efficiency for 3D reconstruction.

[11] “Lightweight image classifier using dilated and depthwise separable convolutions” by Wei Sun, Xiaorui Zhang and Xiaozheng He uses a joint module that reduces the computational burden with depthwise separable convolution, making it possible to apply the network model to resources or computationally constrained devices. The dilated convolution is used to increase the receptive field in the process of convolution without increasing the number of convolution parameters. Experimental results demonstrate that the proposed model makes a good compromise between the classification accuracy and the model size while maintaining the classification accuracy when the network is compressed.

[12] “Rethinking Atrous Convolution for Semantic Image Segmentation” by Liang-Chieh Chen, George Papandreou, Florian Schroff and Hartwig Adam implemented a model that employs atrous convolution with upsampled filters to extract dense feature maps and to capture long range context. The experimental results show that the proposed model significantly improves over previous DeepLab versions and achieves comparable performance with other state-of-art models on the PASCAL VOC 2012 semantic image segmentation benchmark.

3. PROPOSED SYSTEM

For the proposed system we use the basic ideas such as image reconstruction, smoothness, left and right disparity as proposed in [1], but instead of a standard ResNet50 backbone, we use the atrous convolutions in the format of a pyramid pooling.

3.1 Our architecture

We employ the idea of an Atrous Spatial Pyramid Pooling (ASPP) block [3], which contains several atrous convolutions with different atrous rates in parallel (see Fig 2). This is motivated by classical image pyramid methods [13, 14], which process images at different spatial scales. Since varying atrous rates results in filters of different spatial dimensions, an ASPP block resembles a feature map at different spatial scales. The atrous convolutions, along with a global average pooling layer, are concatenated and passed through a 1 × 1 convolution, which reduces the number of channels to 256.

Fig 3 shows the architecture of the encoder that we propose using the ASPP architecture that is shown in Fig 2. The ResNet blocks that are shown have the atrous convolutions instead of the normal convolutions and will be used in multiple as per the input image dimensions. The output from the ASPP will be further passed through the equal number of ResNet decoder blocks to get the final depth map.

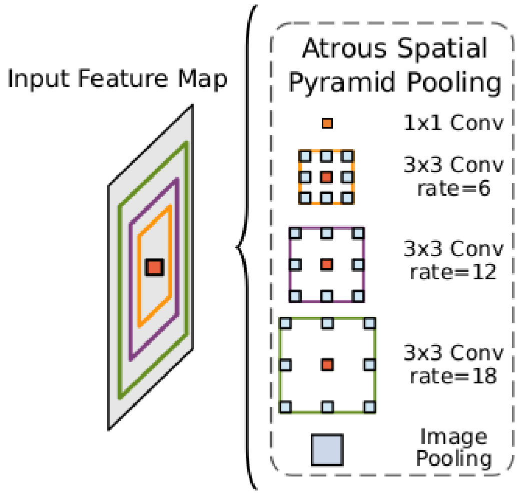

Fig -2: Atrous Spatial Pyramid Pooling (ASPP)

The ASPP module will be inserted after the final layer of the ResNet50, since this placement has proven to work well in other work [4, 15].

3.2 Mathematical Model

As we implement the basic structure proposed by Godard et al [1], we implement the same loss function that the proposed. Their loss function consists of three weighted pairs :

$$C_s = \alpha_{ap}(C^l_{ap} + C^r_{ap}) + \alpha_{ds}(C^l_{ds} + C^r_{ds}) + \alpha_{lr}(C^l_{lr} + C^r_{lr}),$$

where, the first bracketed term, C_{ap} is the appearance-based loss for both the left and right disparities. Further, C_{ds} is the disparity smoothness loss which penalizes the high derivatives in the disparity maps and lastly C_{lr} is the left right consistency loss which ensures that the left and right disparity match are same.

This loss is then computed and summed across all the four different output scales.

The output g(*i*) at index i of a 1D atrous convolution *w* is given by

$$g(i) = \sum_{k=1}^{K} x(i + rk)w(k)$$

where x is the input and K is the filter size. The filter dilation r specifies the rate at which the input is sampled. A standard convolution is a special case of an atrous convolution, where r = 1. The notion of an atrous convolution can be generalized to 2D for straightforward vision problems.

3.3 Evaluation Metrics

In order to evaluate and compare the performance of the depth estimation network proposed by us we chose the metrics that were used in the Godard et al [1] implementation as it is the baseline for our model. The five evaluation indicators are : RMSE, RMSE log, Abs Rel, Sq Rel, Accuracies. These are formulated as :

- RMSE = $\sqrt{\frac{1}{|N|} \Sigma i \in N \, || \, d \, - \, d^* ||^2}$
- RMSE log = $\sqrt{\frac{1}{|N|} \Sigma i \in N \, || \, log(d) \, - \, log(d^*) ||^2}$
- Abs Rel = $\frac{1}{|N|} \Sigma i \in N \frac{|\, d - d^* |}{d^*}$
- Sq Rel =$\frac{1}{|N|} \Sigma i \in N \frac{|\, d - d^* |^2}{d^*}$
- Accuracies: % of d

where d is the predicted depth value of pixel i, and d^* stands for the ground truth of depth. Besides, N denotes the total number of pixels with real-depth values.

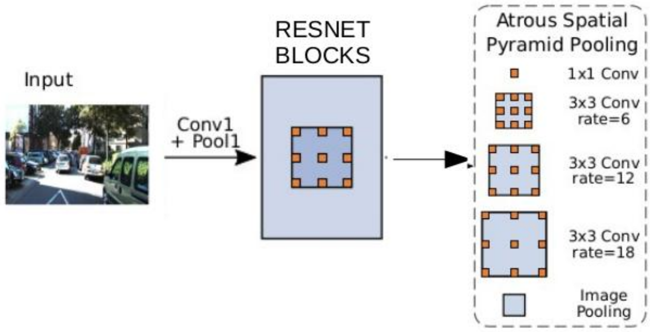

Fig -3: Final proposed architecture

3.4 Datasets

KITTI : The KITTI dataset [16] is the largest and most commonly used dataset for the sub-tasks in computer vision. It is also the commonest benchmark and the primary training dataset in the unsupervised and semi-supervised monocular depth estimation. The real images from “city”, “residential” and “road” categories are collected in the KITTI dataset, and the 56 scenes in the KITTI dataset are divided into two parts, 28 ones for training and the other 28 ones for testing, by Eigen et al. [17]. Each scene consists of stereo image pairs with a resolution of 1224×368. The corresponding depth of every RGB image is sampled in a sparse way by a rotating LIDAR sensor.

NYU Depth : The NYU Depth dataset [18] focuses on indoor environments, and there are 464 indoor scenes in this dataset. Different from the KITTI dataset, which collects ground truth with LIDAR, the NYU Depth dataset takes monocular video sequences of scenes and the ground truth of depth by an RGB-D camera. It is the common benchmark and the primary training dataset in the supervised monocular depth estimation. These indoor scenes are split into 249 ones training and 215 ones for testing. The resolution of the RGB images in sequences is 640×480.

4. FUTURE SCOPE

The work that we propose is in no way the most advanced work in the field. Thus there is a scope of improvement even after the changes we have made. We cannot incorporate these right now because of the limitations. Firstly the thing that we can incorporate in the next step of implementation is the development of an application. Our model outputs a depth map which has various applications right from photo editing to robotics. Thus an application can be developed by making use of this depth map and demonstrating its use. Next is the improved efficiency part. Without limited research and knowledge we came up with the proposal to add a new network to existing architecture. But with more study and research novel loss functions can be created to improve the efficiency further. And lastly as we still use the left right stereo pairs in the proposed system for the training purposes future work may also include the possibility of finding out a way to eliminate this need and make use of mono images for training purposes.

5. CONCLUSION

In this paper we have proposed a new network for the existing Godard et al [1] model that will try to improve the prediction results and accuracy. Our method will also try to reduce the number of channels between the encoder and decoder. This can decrease the number of network parameters and improve runtime, without losing predictive power.

ACKNOWLEDGEMENT

We would like to express our gratitude to our project guide, Prof. V. V. Waykule for her able support and guidance throughout the survey process of this project.

REFERENCES

[1] C. Godard, O. Aodha and G. Brostow, "Unsupervised Monocular Depth Estimation with Left-Right Consistency," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017 pp. 6602-6611.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2018.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 801–818, 2018.

[4] C. Godard, O. Aodha, M. Firman and G. Brostow, "Digging Into Self-Supervised Monocular Depth Estimation," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019 pp. 3827-3837.

[5] W. Yingcai, F. Lijing and Z. Qiankun, "Multi-scale Deep CNN Network for Unsupervised Monocular Depth Estimation," 2019 IEEE 9th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), Suzhou, China, 2019, pp. 469-473, doi: 10.1109/CYBER46603.2019.9066615.

[6] Y. Cao, Z. Wu and C. Shen, "Estimating Depth From Monocular Images as Classification Using Deep Fully Convolutional Residual Networks," in IEEE Transactions on Circuits and Systems for Video

Technology, vol. 28, no. 11, pp. 3174-3182, Nov. 2018, doi: 10.1109/TCSVT.2017.2740321.

[7] A. C. Kumar, S. M. Bhandarkar and M. Prasad, "Monocular Depth Prediction Using Generative Adversarial Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, 2018, pp. 413-4138, doi: 10.1109/CVPRW.2018.00068.

[8] M. Song and W. Kim, "Depth Estimation From a Single Image Using Guided Deep Network," in IEEE Access, vol. 7, pp. 142595-142606, 2019, doi: 10.1109/ACCESS.2019.2944937.

[9] Á. Faluvégi, Q. Bolseé, S. Nedevschi, V. Dădârlat and A. Munteanu, "A 3D Convolutional Neural Network for Light Field Depth Estimation," 2019 International Conference on 3D Immersion (IC3D), Brussels, Belgium, 2019, pp. 1-5, doi: 10.1109/IC3D48390.2019.8975996.

[10] Á. Faluvégi, Q. Bolseé, S. Nedevschi, V. Dădârlat and A. Munteanu, "A 3D Convolutional Neural Network for Light Field Depth Estimation," 2019 International Conference on 3D Immersion (IC3D), Brussels, Belgium, 2019, pp. 1-5, doi: 10.1109/IC3D48390.2019.8975996.

[11] Sun, W., Zhang, X. & He, X. Lightweight image classifier using dilated and depthwise separable convolutions. J Cloud Comp 9, 55 (2020)

[12] Rethinking atrous convolution for semantic image segmentation. arXiv 2017 LC Chen, G Papandreou, F Schroff, H Adam - arXiv preprint arXiv:1706.05587, 2019.

[13] A. Witkin, D. Terzopoulos, and M. Kass. Signal matching through scale space. International Journal of Computer Vision, 1(2):133–144, Jun 1987.

[14] L. H. Quam. Hierarchical warp stereo. In Readings in computer vision, pages 80–86. Elsevier, 1987.

[15] H. Fu, M. Gong, C. Wang, K. Batmanghelich, H. Fu, M. Gong, C. Wang, K. Batmanghelich, D. Tao, D. Ordinal, H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep Ordinal Regression Network for Monocular Depth Estimation

[16] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in the 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 3354–3361.

[17] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in Advances in neural information processing systems, 2014, pp. 2366–2374.

[18] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in European conference on computer vision. Springer, 2012, pp. 746–760.