# Single View Stereo Matching

Wenjie Niu

July 6,2018

## Abstract

*This paper show for the first time that the monocular depth estimation problem can be reformulated as two sub-problems, a view synthesis procedure followed by stereo matching, with two intriguing properties, namely i) geometrical constraints can be explicitly imposed during inference; ii) demand on labelled depth data can be greatly alleviated.*

*It's showed that the whole pipeline can still be trained in an end-to-end fashion and this new formulation plays a critical role in advancing the performance. The resulting model outperforms all the previous monocular depth estimation method as well as the stereo block matching method in the challenging KITTI data set by only using a small number of real training data.*

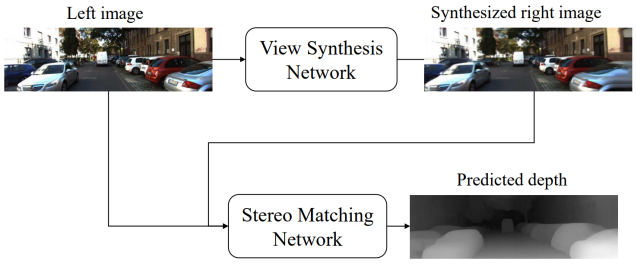


Figure 1. Pipeline of our approach on monocular depth estimation. We decompose the task into two parts: view synthesis and stereo matching. Both networks enforce the geometric reasoning capacity. With this new formulation, our approach is able to achieve state-of-the-art performance.[2]

## 1. Introduction

Authors take a novel perspective and show for the first time that monocular depth estimation problem can be formulated as a stereo matching problem in which the right view is automatically generated by a high-quality view synthesis network. The whole pipeline is shown in figure. 1. The key insights here are that i) both view synthesis and stereo matching respect the underlying geometric principles; ii) both of them can be trained without using the expensive real depth data and thus generated well; iii) the whole pipeline can be collectively trained in an end-to-end fashion that optimize the geometrically correct objectives. The method in this paper is similar idea as revealed in the Spatial Transformation Network [1].

## 2. Analysis and Approach

The whole pipeline is shown in figure. 2. By tracking this problem using two seperate steps, the paper find that both procedures obey primary geometric principles and they can trained without expensive data supply. In order to better utilize the geometric relation between two views, the paper take the idea if 1D correlation employed in DispNetC [3]. The further adopt the DispFullNet structure mentioned in [4] to achieve full resolution prediction.

Their view synthesis network is shown in the upper part of figure. 2. This network is develoed based on Deep3D [37] model.The operation of selection is the core component in this network. This model is also illustrated in figure. 2. Denote $I_l$ as the input left image, previous Depth Image-Based Rendering (DIBR) techniques choose to directly warp the left image based on estimated disparity into a corresponding right image. Suppose $D$ is the predicted disparity aligned with the left image, the procedure can be formulate as Eq: 1

$$\tilde{I}_r(i, j - D_{i,j}) = I_l(i, j), \quad (i, j) \in \omega_l \tag{1}$$

where $\omega_l$ is the image space of $I_l$ and $i$, $j$ refer to the row and column on $I_l$ respectively.

## 3. Conclusions

We explicitly encode the geometric transformation within both networks to better tackle the problems individually. Collectively training the whole pipeline results in an overall boost and we prove that both networks are able to preserve their original functionality after end-to-end training. Without using a large amount of expensive ground truth labels, we outperform all previous methods on a monocular depth estimation benchmark. Remarkably, we are the first

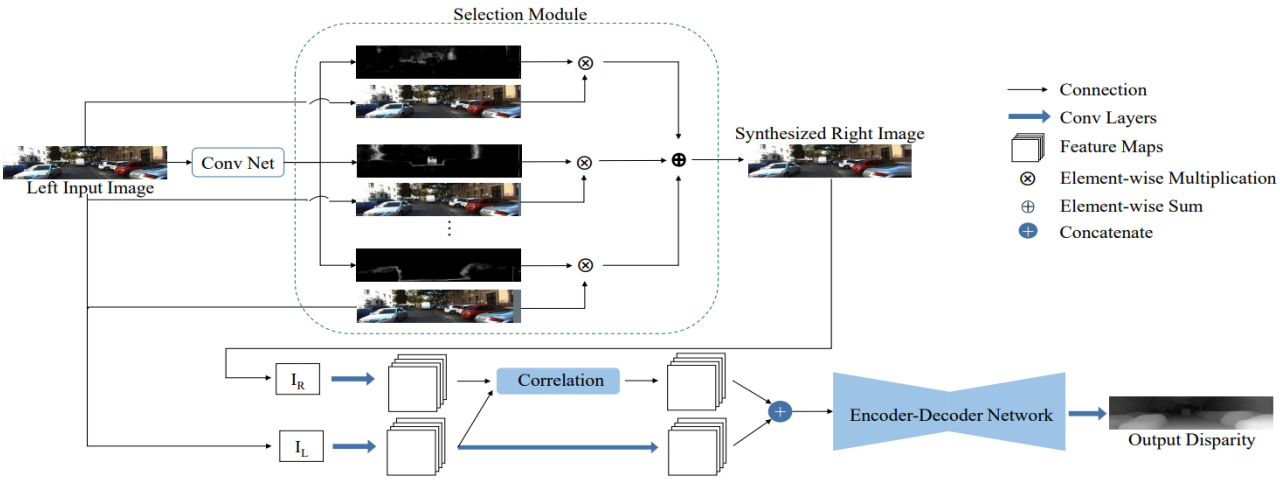


Figure 2. Details of our single view stereo matching network. Upper part is the view synthesis network. The input image is first processed by a CNN. It results in probabilistic disparity maps that help to reconstruct a synthetic right view by selectively taking pixels from nearby locations on the original left image. A stereo matching network, which is shown on the lower part of the figure, then takes both the original left image and synthetic right image to calculate an accurate disparity, which can be transformed into a corresponding depth map given the camera settings [2].

to outperform the stereo blocking matching algorithm on a stereo matching benchmark using a monocular method.

## References


[1] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In *Neural Information Processing Systems*, 2015. 1

[2] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin. Single view stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[3] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[4] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *IEEE International Conference on Computer Vision Workshop*, 2017. 1