2017 IEEE International Symposium on Mixed and Augmented Reality Adjunct Proceedings

[POSTER] Depth Map Interpolation using Perceptual Loss

Ilya Makarov* Vladimir Aliev† Olga Gerasimova‡ Pavel Polyakov

National Research University Higher School of Economics, Moscow, Russia

Figure 1: Example of sample interpolation from NYUDepthv2 set. Left to right: intensity image, input depth map, output depth map, ground truth.

ABSTRACT

In this paper, we discuss a semi-dense depth map interpolation method based on convolutional neural network. We propose a compact neural network architecture with loss function defined as Euclidean distance in the feature space of VGG-16 neural network used for deep visual recognition. The suggested solution shows state-of-art performance on synthetic and real datasets. Together with LSD-SLAM, the method could be used to provide a dense depth map for interaction purposes, such as creating a first person game in AR/MR or perception module for autonomous vehicle.

Keywords: Depth Map, Semi-Dense Depth Map Interpolation, Deep Convolutional Neural Networks, Mixed Reality, FPS

Index Terms: K.10.2 [Human-centered computing]: HCI—Mixed/augmented reality; K.11.3 [Computing methodologies]: AI—Computer vision problems: Reconstruction

1 INTRODUCTION

Depth map estimation is the one of the key problems in augmented reality and computer vision. It is used for 3D reconstruction, dense SLAM [9], object detection [6] scene segmentation, etc. There are many systems providing depth map measurements [5].

However, in order to decrease conditions on tracking devices or lack of stereo setup, depth map can be estimated from monocular video sequence. This approach requires high performance computation. To overcome this difficulty modern SLAM systems estimate semi-dense [4], edge-based [19], or even sparse [3] depth, distributed according to intensity image gradient. In cases, when semi-dense or sparse depth maps can not provide full geometric information, the depth map approximation methods are focused on interpolation from low-resolution dense depth image. Such methods can be both single depth map upsampling methods [10] and also make use of intensity image [14], or even learn to interpolate using convolutional neural networks [8]. We propose a method to interpolate depth directly from semi-dense depth map using convolutional neural networks. Figure 1 illustrates the obtained results.

2 DEPTH MAP INTERPOLATION

Most of the depth map reconstruction methods focus on interpolating low-resolution depth maps. Such methods include both, traditional methods like [11, 2, 10], and deep learning based method, such as MSG-Net [8].

However, semi-dense depth interpolation methods are not widely presented in the literature. The most well-known examples are depth reconstruction methods using 2–5% of known depth values [7, 13] based on the theory of compressed sensing and allowing.

Our work is close to these methods in the sense that we can use high resolution semi-dense depth map with different spatial distributions as input. In contrast with mentioned above methods, our method needs 10–15% values to reconstruct depth map, but has $\times 33$ smaller running time (300ms on CPU compared to 10s as reported in [13]). Once trained, our method infers depth values by simple forward pass through the network, while compressed sensing methods cited above require solving optimization problem to reconstruct depth, which results in higher runtimes.

2.1 Network Architecture

The architecture of our network is shown on Figure 2. It is a residual architecture with heavy usage of dilated convolutions. Such architectures are widely used in semantic segmentation tasks[1]. The main reason for choosing this network configuration is that we need to take a full resolution semi-dense depth map as input and then produce interpolated depth with the same resolution. The intuition behind this the following: LSD-SLAM usually provide edge information with high confidence in the input semi-dense depth map that should not be downsamled too much to not loose detailed depth interpolation of the same size as the input map. Thus, we implement only one stage of $\times 2$ downsampling.

In the same time, we want our network to have big receptive field to be able to catch full spatial structure of semi-dense depth map. To achieve this property, we use dilated convolutions with gradually increasing dilation rate, so deeper layers have bigger receptive field.

The core building blocks of the encoder part are the residual and projection blocks. We use residual blocks in our architecture in order to improve interpolation quality compared to plain sequential convolutions. The residual blocks included two convolutional layers with ELU activations. In our case, ELU activations speed up convergence compared to ReLU. Projection blocks have additional 1×1 convolution in the identity path to match feature dimensions. All filters in the network are 3×3. For regularization we use dropout with rate 0.5. Last stage is upsampling to the full size. It consists of single subpixel convolution (also known as pixelshuffle) layer. Such layer helps to avoid checkerboard-pattern artefacts in the resulting depth map.

2.2 Loss function

We formulate our task as a regression problem: for input semi-dense map $\mathbf{x} \in R^{m \times n}$ find its interpolation $\hat{\mathbf{y}} \in R^{m \times n}$ minimizing the

*corresponding author, e-mail: iamakarov@hse.ru
†corresponding author, e-mail: twoleggedeye@yandex.ru
‡e-mail: olga.g3993@gmail.com

978-0-7695-6327-5/17 $31.00 © 2017 IEEE
DOI 10.1109/ISMAR-Adjunct.2017.39

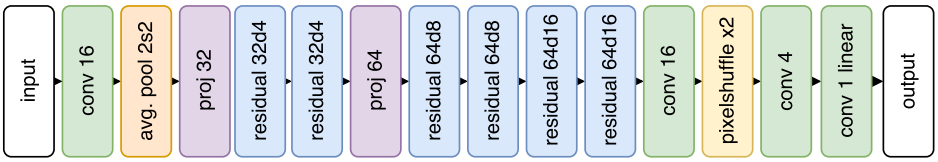

Figure 2: Neural Network architecture. All the layers except the latter convolutional layer use ELU activation function. Green are plain convultional layers, blue and violet are residual blocks, depth to space upsampling and average pooling denoted as yellow.

loss function $L(\hat{\mathbf{y}},\mathbf{y})$ constructed as a sum of VGG perceptual loss and total variation penalty:

$$L(\hat{\mathbf{y}},\mathbf{y}) = \frac{1}{WHC}\sum_{i=1}^{W}\sum_{j=1}^{H}\sum_{k=1}^{C}(\Phi(y)_{i,j,k} - \Phi(\hat{y})_{i,j,k})^2 + \alpha\|\nabla\hat{\mathbf{y}}\|_2^2,$$

Traditional loss functions such as MSE tend to give over-smoothed results and often has poor perceptual quality. We use Euclidean distance in the feature space of VGG [18] neural network, which provided better visual quality in superresolution [12], and style transfer [20]. The second term in the loss function is total variation regularizer, serving to suppress noise-like artefacts.

2.3 Experiments

The network was trained on the part of SceneNet-RGBD [15] dataset with high number of high-quality depth ground truth data. We used sequence of 60000 examples for training and 5000 for validation. We used Middlebury dataset [16] to compare with [13], and NYUDepth v2 [17] real set to compare with synthetic SceneNet-RGBD (see Table 1). Semi-dense depth maps were generated by applying masks to ground truth data. For masks, we used adaptive thresholding of corresponding intensity image gradient. Along with the image, we feed network with mask of known values.

	MAPE, %	RMSE
NYUDepthv2	17.1	0.13
SceneNet-RGBD	8.2	0.11

Table 1: Almost indistinguishable results on Validation Sequences

	proposed, dB	Liu et al. [13], dB
Dolls	**33.8**	32.5
Moebius	29.1	**35.1**
Art	29.4	**34.1**
Aloe	28.4	**31.4**

Table 2: Results of comparison with [13].

In Table 2 we have compared our methods with [13], using 10% density of input values. While our method does not outperform [13], our runtime for 1000×1300 image is 300ms compared to 10 seconds as reported in [13]. For computing RMSE, data was normalized to [0, 1] range. Network was trained with Adam optimizer, using learning rate 10^{-4} for the first 40k iterations and 10^{-5} for the last 30k iterations. We used Intel Core I7-4770K with NVIDIA GTX960 GPU.

3 Discussion

We presented a method of a depth map interpolation from a semi-dense depth map using neural network. Combined with semi-dense SLAM or semi-dense stereo methods our method can be used to provide dense reconstruction for MR/AR applications. Another direction of work is the optimisation of neural network architecture to get suitable runtimes on mobile devices for real-time reconstruction. The results may be used for creating first-person shooter game for MR/AR or perception module for autonomous vehicles with low-cost sensors.

Acknowledgements

The work was supported by the Russian Science Foundation under grant 17-11-01294 and performed at National Research University Higher School of Economics, Russia.

References

[1] L.-C. Chen et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016.

[2] J. Diebel and S. Thrun. An application of markov random fields to range sensing. In *Proceedings of the 18th IC NIPS*, NIPS'05, pages 291–298, Cambridge, MA, USA, 2005. MIT Press.

[3] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *arXiv preprint arXiv:1607.02565*, 1607.02565:1–17, 2016.

[4] J. Engel, T. Schöps, and D. Cremers. *LSD-SLAM: Large-Scale Direct Monocular SLAM*, pages 834–849. Springer, Cham, 2014.

[5] S. Foix, G. Alenya, and C. Torras. Lock-in time-of-flight (tof) cameras: A survey. *IEEE Sensors Journal*, 11(9):1917–1926, Sept 2011.

[6] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision (ECCV)*. Springer, 2014.

[7] S. Hawe, M. Kleinsteuber, and K. Diepold. Dense disparity maps from sparse disparity measurements. In *2011 International Conference on Computer Vision*, pages 2126–2133, NY, USA, Nov 2011. IEEE.

[8] T.-W. Hui, C. C. Loy, and X. Tang. *Depth Map Super-Resolution by Deep Multi-Scale Guidance*, pages 353–369. Springer, Cham, 2016.

[9] C. Kerl, J. Sturm, and D. Cremers. Dense visual slam for rgb-d cameras. In *Proc. of the Int. Conf. on Int. Robot Systems (IROS)*, 2013.

[10] Y. Konno et al. Depth map upsampling by self-guided residual interpolation. In *Proceedings of the 23rd IC ICPR2016*, pages 1395–1400, New York, NY, USA, 2016. IEEE.

[11] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele. Joint bilateral upsampling. *ACM Trans. Graph.*, 26(3):1–5, July 2007.

[12] C. Ledig et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv:1609.04802*, pages 1–19, 2016.

[13] L. K. Liu, S. H. Chan, and T. Q. Nguyen. Depth reconstruction from sparse samples: Representation, algorithm, and sampling. *IEEE Transactions on Image Processing*, 24(6):1983–1996, June 2015.

[14] K. Matsuo and Y. Aoki. Depth interpolation via smooth surface segmentation using tangent planes based on the superpixels of a color image. In *2013 IEEE IC on CVW*, pages 29–36, Dec 2013.

[15] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison. Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. *arXiv preprint arXiv:1612.05079*, 2016.

[16] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *IEEE IC on CVPR*, pages 1–8, NY, USA, 2007. IEEE.

[17] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. *Indoor Segmentation and Support Inference from RGBD Images*, pages 746–760. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 1409.1556:1–14, 2014.

[19] J. J. Tarrio and S. Pedre. Realtime edge-based visual odometry for a monocular camera. In *Proceedings of the 2015 IEEE ICCV*, pages 702–710, Washington, DC, USA, 2015. IEEE Computer Society.

[20] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *Proceedings of the 33rd ICML*, pages 1349–1357. JMLR.org, 2016.