Depth Map Interpolation and Super Resolution using Perceptual Loss

Ilya Makarov et al.

National Research University Higher School of Economics

Moscow, 2017

Why we want to estimate depth map

Autonomous navigation

- Obstacle avoidance and path planning;
- Full 3D scene reconstruction;

VR/AR/MR

- Projecting virtual objects on video;
- Adjusting the perspective and distance;

Depth Map Acquisition

Active Methods

- Time of Flight (LIDAR);
- Structured Light (Microsoft Kinect v1);

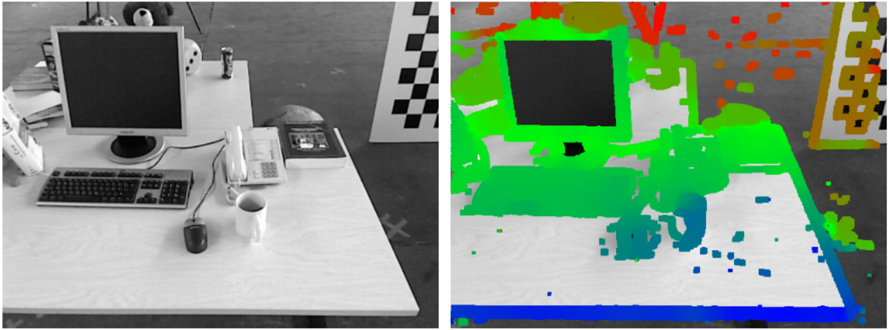
Passive Methods

- Dense Stereo;
- **Semi-dense stereo**;
- Sparse Stereo;

Sparse Depth Maps, Examples

Algorithms for depth map estimation:

- Semi-Dense Visual Odometry [2] — camera trajectory estimation from monocular video;
- Fast Tunable Stereo Reconstruction [5] — computationally efficient of stereo matching ;



Main Idea

Our goal is to suggest efficient method of depth map interpolation with different space distributions of known depth values.

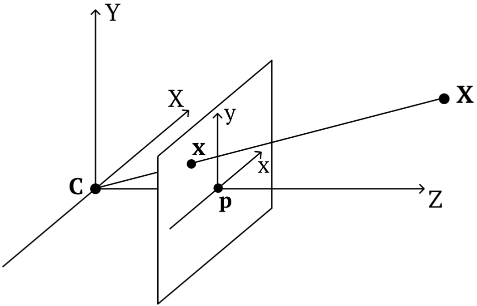
What we do:

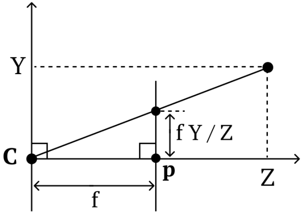
- Developed neural network with specific loss function;
- Conducted numerical experiments on synthetic and real data;
- Proved the applicability of CNNs for depth map interpolation with sparse or low resolution known depth values;

Camera Model and Depth Map

Depth Map

Let us fix a point $X \in \mathbb{R}^3$ with coordinates (x, y, z) and let $I_{i,j}$ be its image under projection on the image $\boldsymbol{I} \in \mathbb{R}^{m \times n}$. Depth map is a matrix $\boldsymbol{D} \in \mathbb{R}^{m \times n}$ such that $d_{i,j} = z$.





Problem Statement

Neural Network

Neural Network is a mapping $f(x, \boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, presented as a composition of elementary functions $f^{(i)}(x, \theta_i)$, called layers.

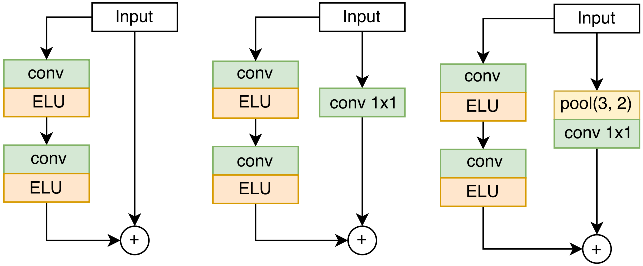
For a given sample $S = \{(\mathsf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, where $\mathsf{x}_i, \mathbf{y}_i \in \mathbb{R}^{m \times n}$ is a sparse or full depth map.

We need to find a mapping $f(\mathsf{x}, \boldsymbol{\theta}) : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$, minimizing empirical loss:

$$f(\mathsf{x}, \boldsymbol{\theta}^*) = \arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} L(f(\mathsf{x}_i, \boldsymbol{\theta}), \mathbf{y}_i)$$

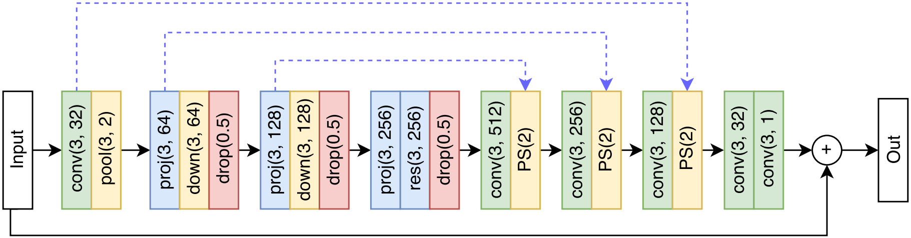
Network Architecture. Core blocks

- Encoder decreases spatial redundancy of sparse depth map;
- Decoder restores full depth map using inner representation from the auto–encoder;



Core Blocks for Encoder

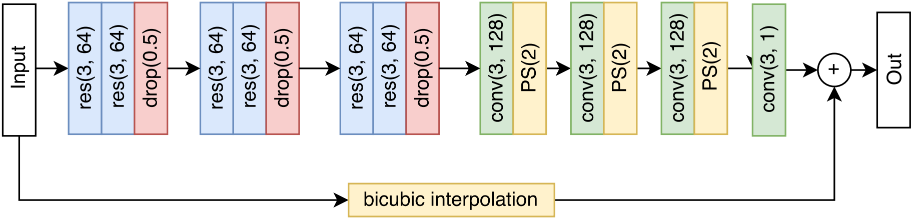
Network Architecture for Depth Map Interpolation



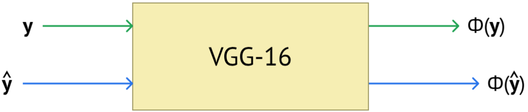
Color representation for layer functionality: yellow for decreasing spatial redundancy, green for simple convolution, red for regularizer, blue for other blocks. Dashed lines stand for concatenations of inner representations of encoder and decoder.

Network Architecture for Depth Map Superresolution

A little change for encoder architecture allows to obtain depth map superresolution:



Loss Function



$$\hat{\mathbf{y}} = f(\mathbf{x}, \boldsymbol{\theta})$$

$$L(\mathbf{y}, \hat{\mathbf{y}}) = L_{square}(\mathbf{y}, \hat{\mathbf{y}}) + \alpha L_{vgg}(\mathbf{y}, \hat{\mathbf{y}})$$

$$L_{square}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} (y_{i,j} - \hat{y}_{i,j})^2$$

$$L_{vgg}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{WHC} \sum_{i=1}^{W} \sum_{j=1}^{H} \sum_{k=1}^{C} (\Phi(\mathbf{y})_{i,j,k} - \Phi(\hat{\mathbf{y}})_{i,j,k})^2,$$

Data

- **SYNTHIA** [6] is a synthetic dataset of road scenes. Training set consists of 8000 samples, test set contains 4000;
- **NYU Depth** [8] is a real dataset from depth sensor data indoor, 2500 samples;
- **Sintel** [1] is a synthetic dataset with different scenes, 1500 samples;
- **Middlebury** [7] is a real benchmark dataset with 12 examples;

Metrics

- Mean Absolute Percentage Error

$$MAPE(x,y) = \frac{1}{WH}\sum_{j=1}^{H}\frac{|y_{i,j}-x_{i,j}|}{y_{i,j}}\cdot 100\%$$

- Root Mean Squared Error

$$RMSE(x,y) = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}(y_{i,j}-x_{i,j})^2$$

- Structural Similarity Index

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c1)(\sigma_x^2 + \sigma_y^2 + c2)}$$

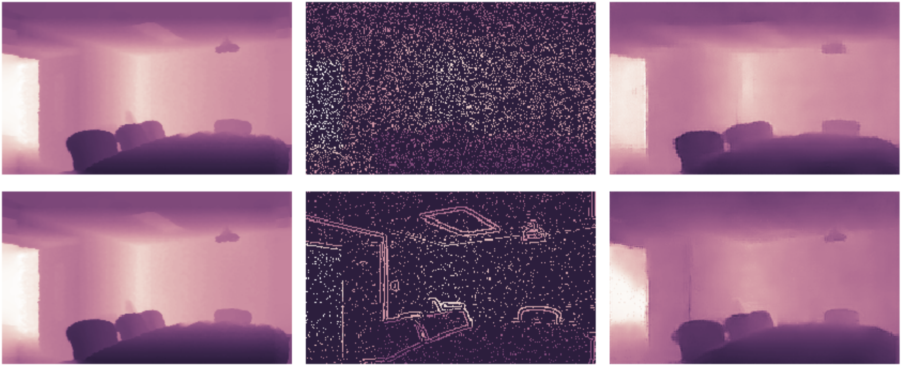
- Peak Signal to Noise Ratio

$$PSNR(x,y) = 20\log_{10}(\frac{\sqrt{MAX}}{RMSE(x,y)})$$

Results

	MAPE, %			RMSE, m		
Sparse Depth Distribution	SYNTHIA 01	NYU Depth	Sintel	SYNTHIA 01	NYU Depth	Sintel
Uniform	4.9	5.8	28.2	52.1	0.54	4.66
Grid	24.6	23.7	48.1	71.7	0.69	5.20
Gradient	8.7	11.5	36.3	54.9	0.62	4.38
Gradient + Uniform	**3.8**	**4.8**	**24.5**	**46.1**	**0.35**	**4.00**

Table: Result on depth map interpolation. Small values of MAPE and RMSE mean better quality.



Examples of our results with different input distributions. Top for uniform, bottom for combined. From left to right: ground truth, input sparse depth map, interpolated full depth map.

Results for depth map interpolation

	Our method, dB	Liu et al. [4], dB
Dolls	**35.8**	33.9
Moebius	35.7	**37.7**
Art	34.7	**37.3**
Aloe	32.9	**33.4**
Rocks	35.4	**37.5**

Table: Comparison of our method with state-of-art best method. PSNR metric. 10–15 % known depth values. Bigger value of PSNR means better quality.

One depth map processing time:

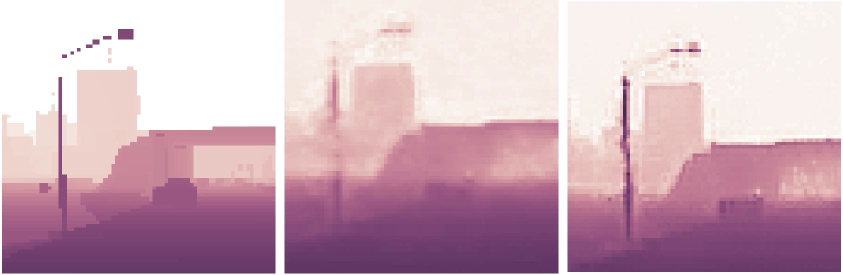
- Our method: 0.76s (CPU), 0.088s (GPU);
- Liu et al. [4]: >10s (CPU);

Results on loss function comparison

	MSE			MSE & VGG		
	MAPE,%	RMSE, m	SSIM	MAPE, %	RMSE, m	SSIM
SYNTHIA	4.1	46.9	0.94	**3.8**	**46.1**	**0.96**
NYU Depth	**4.5**	0.35	0.87	4.8	0.35	**0.89**
Sintel	**24.4**	4.1	0.55	24.5	**4.0**	**0.58**

Table: Loss function comparison. Less value of RMSE and greater value of SSIM means better quality.

Results on Loss function comparison



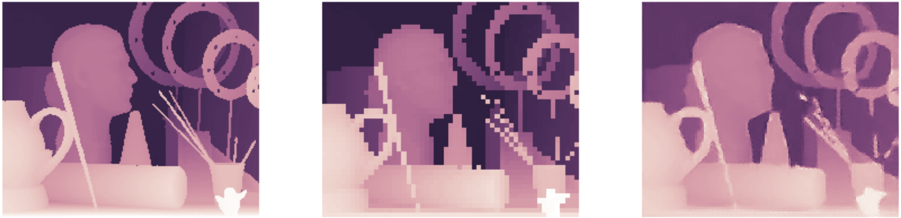
Scaled fragments of depth map. Right to left: ground truth, only MSE loss, combined loss function.

Results on Depth Map Superresolution

	Art		Laundry		Moebius		Dolls	
	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE
Yang [10]	0.43	46.21	0.41	33.87	0.41	19.01	0.54	16.42
Wang [9]	0.55	47.07	0.53	32.56	0.54	19.76	0.63	15.16
Konno et al. [3]	0.63	30.01	0.59	**21.31**	0.61	12.09	0.70	**10.86**
Our Method	**0.68**	**28.02**	**0.79**	21.92	**0.87**	**8.95**	**0.86**	26.98

Table: Comparison of Depth Map Superresolution, x8 upsacling. esults obtained from [3].

Results on Depth Map Superresolution



Example of superresolution. From left to right: ground truth, low-resolution depth map, output.

Discussion and Future Work.

- Artifacts for scenes with big depth differences;
- Neural Network compression and optimisation;
- Integration with saprse depth map estimation methods (SLAM, etc.);

Thanks for your attention!

iamakarov@hse.ru

- Butler, D. J., Wulff, J., Stanley, G. B., and Black, M. J.
  A naturalistic open source movie for optical flow evaluation.
  In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI* (Berlin, Heidelberg, 2012), ECCV'12, Springer-Verlag, pp. 611–625.

- Engel, J., Sturm, J., and Cremers, D.
  Semi-dense visual odometry for a monocular camera.
  In *Proceedings of the IEEE international conference on computer vision* (2013), pp. 1449–1456.

- Konno, Y., Tanaka, M., Okutomi, M., Yanagawa, Y., Kinoshita, K., and Kawade, M.
  Depth map upsampling by self-guided residual interpolation.
  In *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR2016)* (New York, NY, USA, 2016), IEEE, pp. 1395–1400.

- Liu, L. K., Chan, S. H., and Nguyen, T. Q.

Depth reconstruction from sparse samples: Representation, algorithm, and sampling.
IEEE Transactions on Image Processing 24, 6 (June 2015), 1983–1996.

Pillai, S., Ramalingam, S., and Leonard, J.
High-performance and tunable stereo reconstruction.
In *Robotics and Automation (ICRA), 2016 IEEE International Conference on* (2016), IEEE.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M.
The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes.
In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (New York, NY, USA, June 2016), IEEE, pp. 3234–3243.

Scharstein, D., and Pal, C.
Learning conditional random fields for stereo.
In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (New York, NY, USA, 2007), IEEE, IEEE, pp. 1–8.

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

Wang, L., Wu, H., and Pan, C.
Fast image upsampling via the displacement field.
IEEE Transactions on Image Processing 23, 12 (Dec 2014), 5123–5135.

Yang, J., Wright, J., Huang, T. S., and Ma, Y.
Image super-resolution via sparse representation.
Trans. Img. Proc. 19, 11 (Nov. 2010), 2861–2873.