Depth Map Reconstruction using Deep Convolutional Neural Networks

Alice Korinevskaya and Ilya Makarov

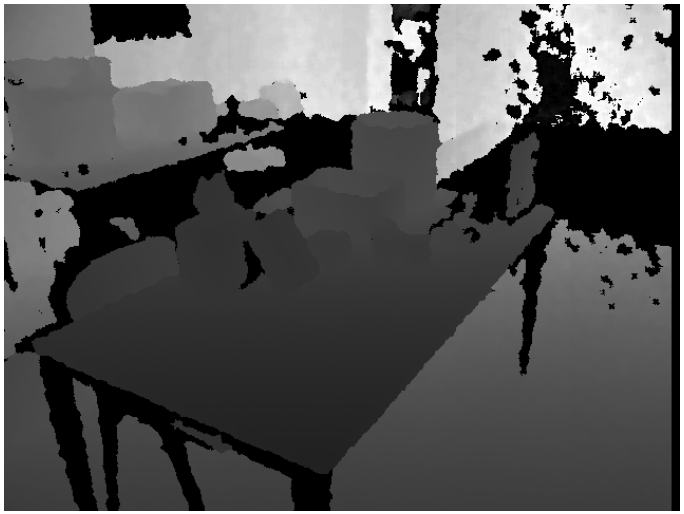
Moscow, 2018

Content

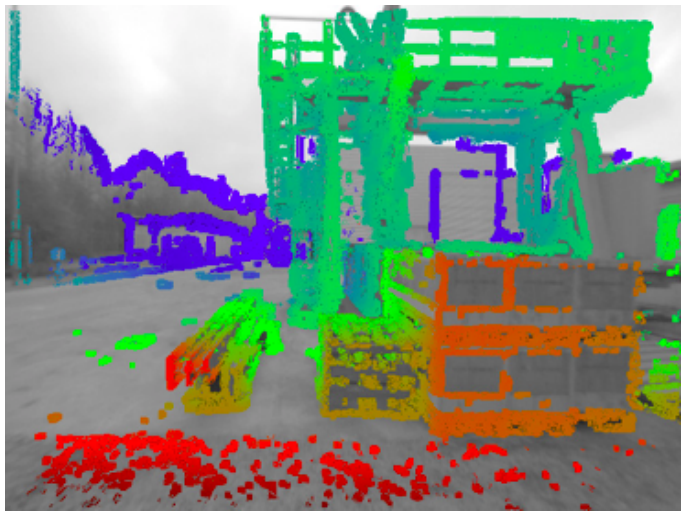
Depth Map

- **Depth Map** is an image containing information about the distance from the observation point to the scene objects.
- Depth could be obtained from **depth sensors**, such as LiDAR, Miscrosoft Kinect, Time of Flight, or calculated from vSLAM models.
- Many sensors allow to get depth of low quality or very sparse depth maps.

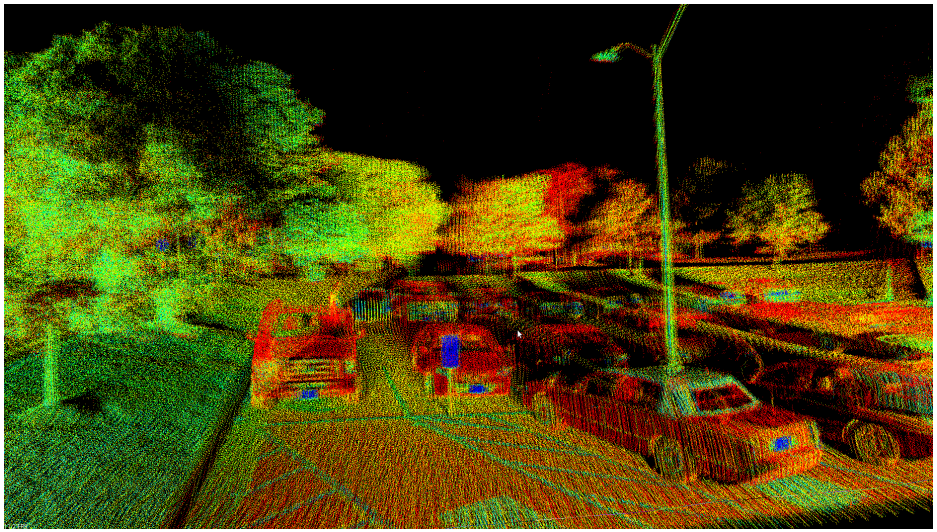
Examples of Depth Maps



(a) Microsoft Kinect



(b) LSD-SLAM

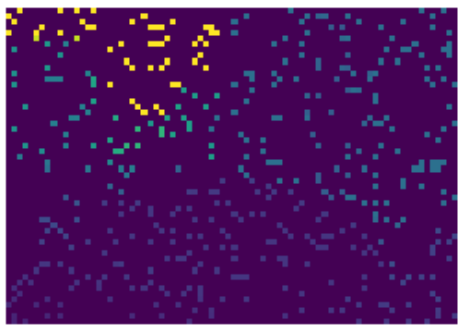


(c) LiDAR

Depth Reconstruction

- **Depth Interpolation**
  Sparse-to-Depth Reconstruction
- **Depth Prediction**
  Depth Estimation based on RGB, etc.
- **Depth Inpainting**
  Filling holes in Depth Map
- **Depth Super-Resolution**
  Upscaling Depth from Interpolation or Depth Sensor

Depth Interpolation



RGB-to-Depth Reconstruction

- noisy

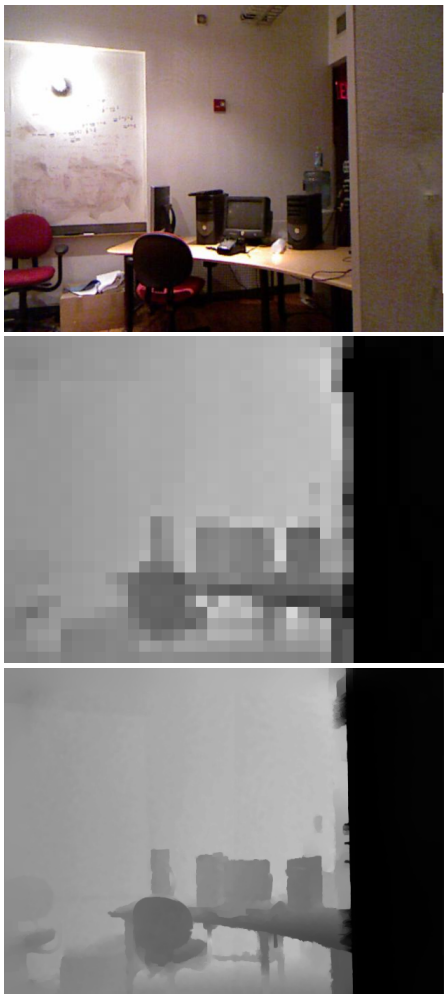
Depth Interpolation with the help of RGB

- not always have corrected RGB images
- better

Sparse-to-Depth Interpolation

- rough but fast
- RGB independent

Depth super-resolution



Using RGB

- not always have good corrected RGB images
- works better

Low-res Depth only

- Works faster
- RGB is not required

Why study this

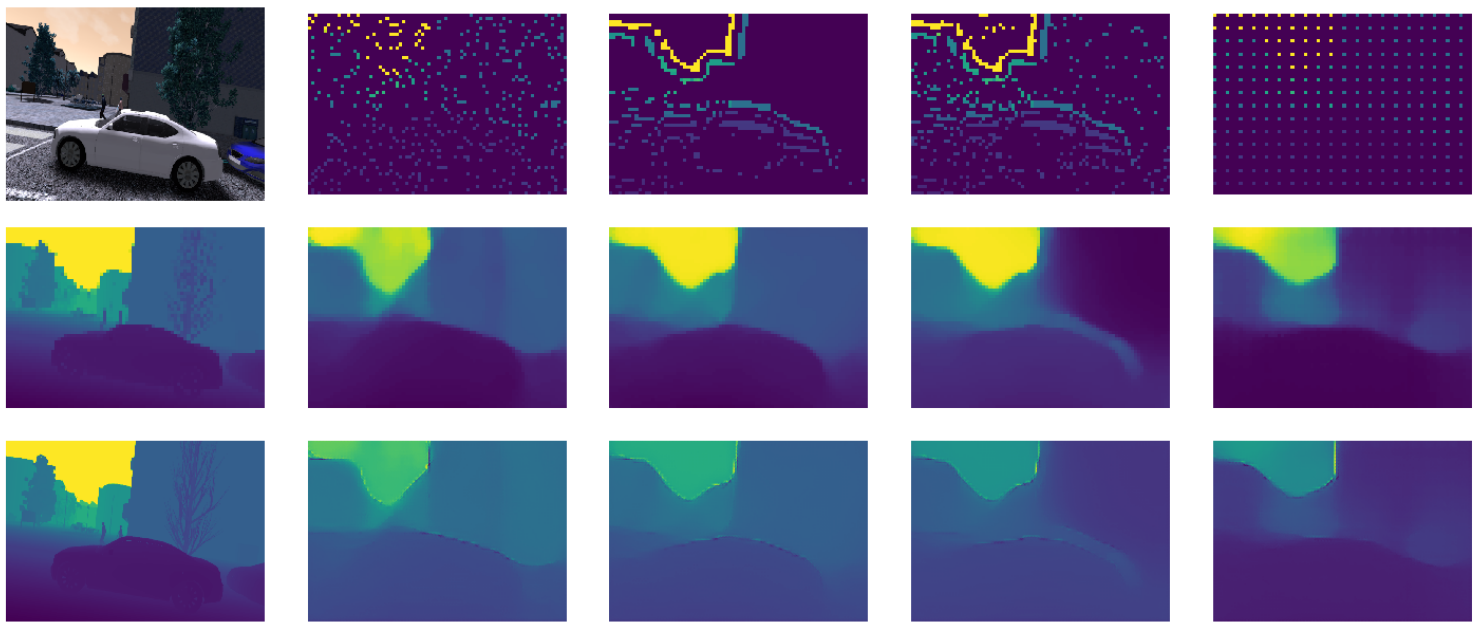
"Semi-Dense Depth Interpolation using Deep Convolutional Neural Networks"
I. Makarov, et al. – the 2017 ACM:

- Semi-dense–to–Dense Depth Reconstruction and Super-resolution
- SotA results.
- Memory consumption, GPU-only real-time.

Experiment 1:
Fast Depth Interpolation

Combined approach:

1. Downscale sparse depth map
2. Interpolate
3. Upscale



Fast Depth Interpolation: Results

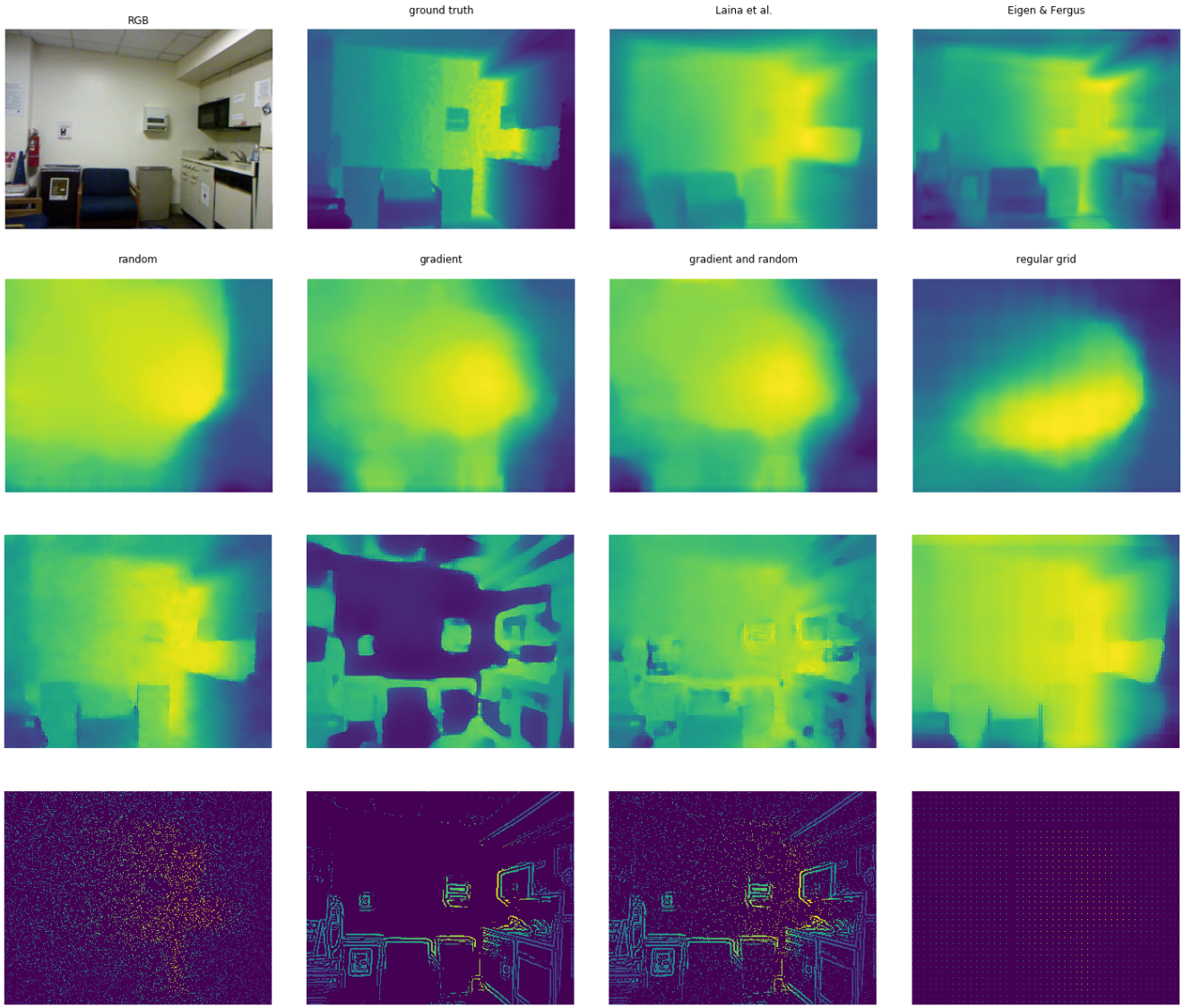
- For SYNTHIA, NYU Depth Dataset v2, KITTI results are worse than Interpolation
- For Matterport 3D results are better due to noisy data
- Quality is independent to sparse depth values distribution
- x5 times faster than interpolation (on average, 200 FPS compared to 40 FPS and 5 FPS for interpolation and RGB-to-Depth models Laina et al. [10], respectively)

Fast Depth Interpolation: NYU Depth (1)

Comparison on NYU Depth Dataset V2

Methods	**RMSE**	**rel**
Laina et al. [10]	0.573	0.127
Eigen & Fergus [3]	0.641	0.158
Sparse-to-Dense [16] (RGB)	0.514	0.143
Liu et al. [14]	0.759	0.213
Interpolation	<u>0.286</u>	<u>0.08</u>
Proposed Approach	0.958	0.37

Fast Depth Interpolation: NYU Depth (2)

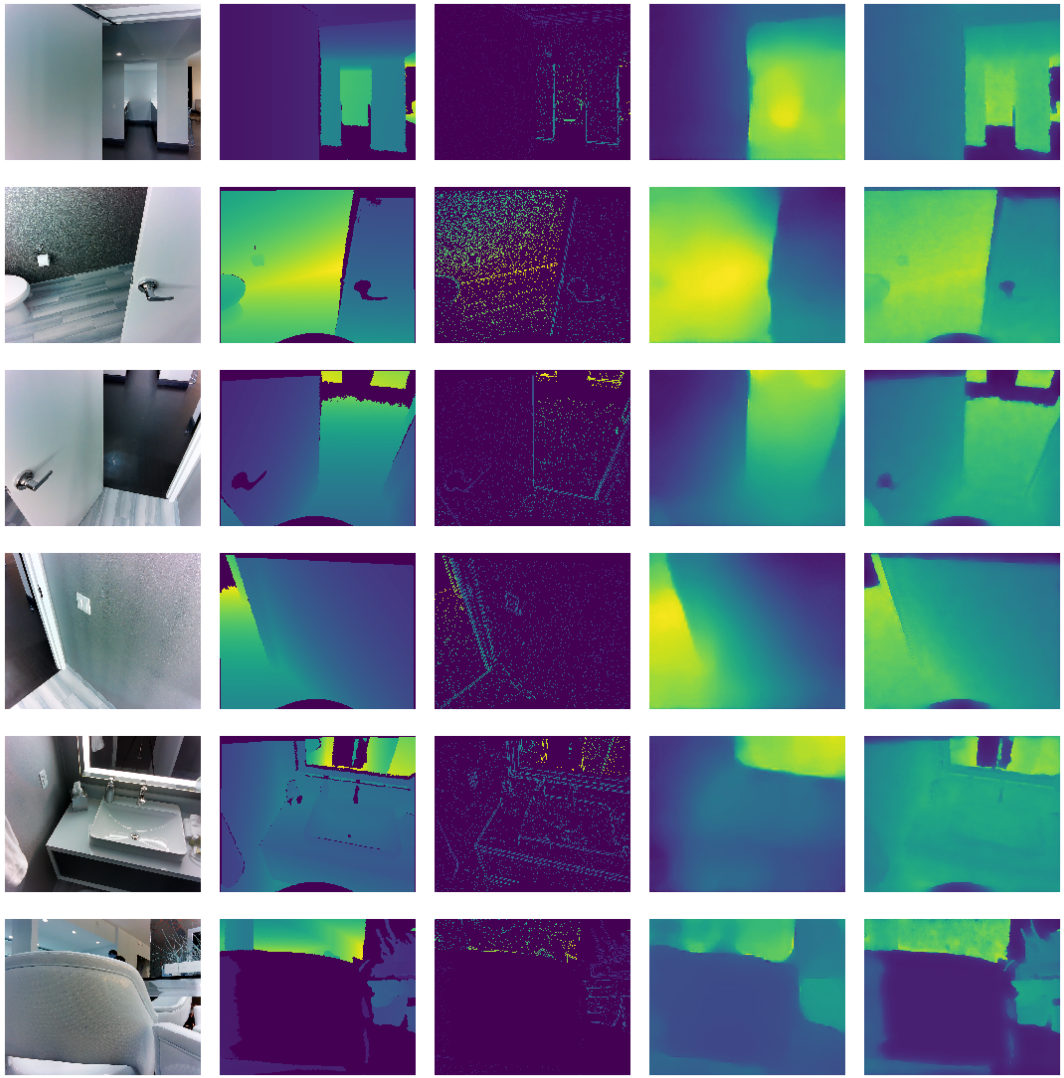


Fast Depth Interpolation: Matterport 3D (1)

	Proposed Model			
Distribution	**uniform**	**gradient**	**unif & grad**	**grid**
RMSE	4.197	4.477	4.225	4.280
rel	0.396	0.447	0.395	0.408
SSIM	0.790	0.694	0.795	0.774

	Interpolation			
Distribbution	**uniform**	**gradient**	**unif & grad**	**grid**
RMSE	4.494	5.530	4.513	4.996
rel	0.638	0.848	0.643	0.760
SSIM	0.777	0.479	0.780	0.771

Fast Depth Interpolation: Matterport 3D (2)

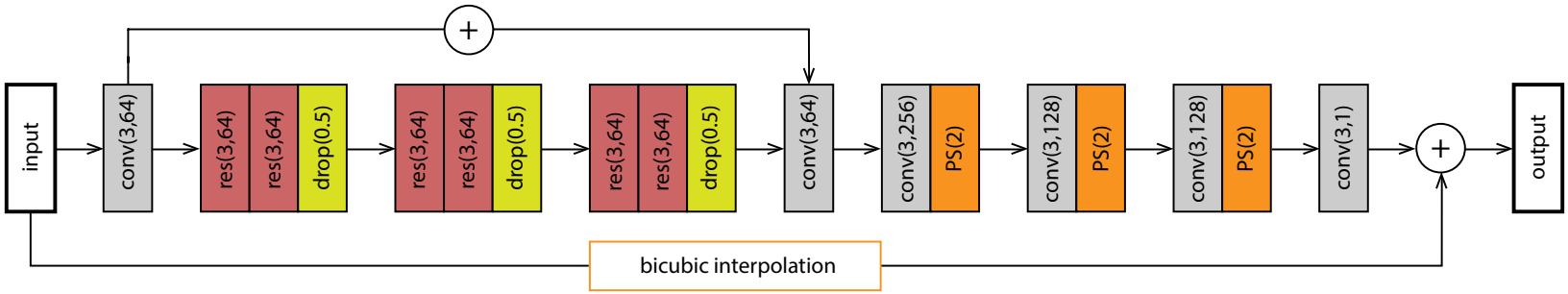


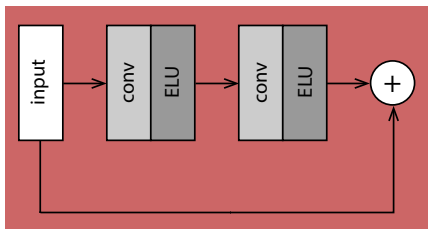
Experiment 2: Depth Map Super-resolution

GAN architecture based on previous idea:

- activations ELU and Leaky ReLU are used in generator and discriminator, correspondingly
- layers for batch normalization were added to generator and discriminator
- generator output activation changes to $\tanh$

Depth Map Super-resolution: Network





residual block

Depth Map Super-resolution: Loss

Generator Loss:

$$L_G = L_{MSE} + \alpha L_{VGG} + \beta L_{GAN} + \gamma L_{TV}$$

Crossentropy for Discriminator:

$$L_D = -\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{data}} \log D(\mathbf{x}) - \frac{1}{2}\mathbb{E}_{\mathbf{z}} \log\{1 - D(G(\mathbf{z}))\}$$

- hyperparameter choice α, β, γ during cross-val
- training on SYNTHIA dataset

Depth Map Super-resolution: Results

- Results on Sintel & Middlebury are better than original network.
- Preserves edges and curves, but produces texture artifacts.
- Compared to SotA values of RMSE.
- x10 faster than RGB SRGAN [11], x4 faster than MS-Net [6].

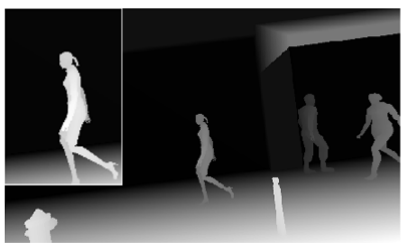
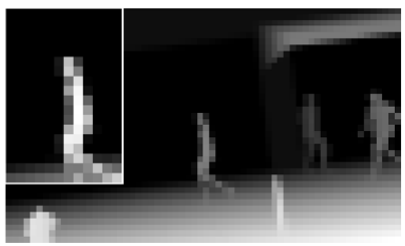
Depth Map Super-resolution: Comparison on Sintel. RMSE and SSIM

Sintel Dataset	Root Mean Square Error (RMSE)				
	Alley	Ambush	Market	Bandage	Cave
Bicubic	3.60	**6.66**	9.59	**6.66**	8.21
NN	8.17	9.05	9.98	9.05	8.98
Proposed Model	**3.57**	6.70	**4.11**	6.70	**5.99**

Sintel Dataset	Structural Similarity (SSIM)				
	Alley	Ambush	Market	Bandage	Cave
Bicubic	0.92	**0.85**	0.83	**0.85**	**0.85**
NN	0.87	0.81	0.68	0.81	0.82
Proposed Model	**0.96**	0.84	**0.84**	**0.85**	0.84

Depth Map Super-resolution: Comparison on Sintel. PSNR measure

Sintel Dataset	Peak Signal-to-Noise Ratio (PSNR)				
	Alley	Ambush	Market	Bandage	Cave
Bicubic	28.09	25.10	23.57	25.10	22.99
NN	25.02	23.09	20.67	23.09	21.77
Proposed Model	**33.21**	**25.37**	**27.41**	**25.37**	**25.36**

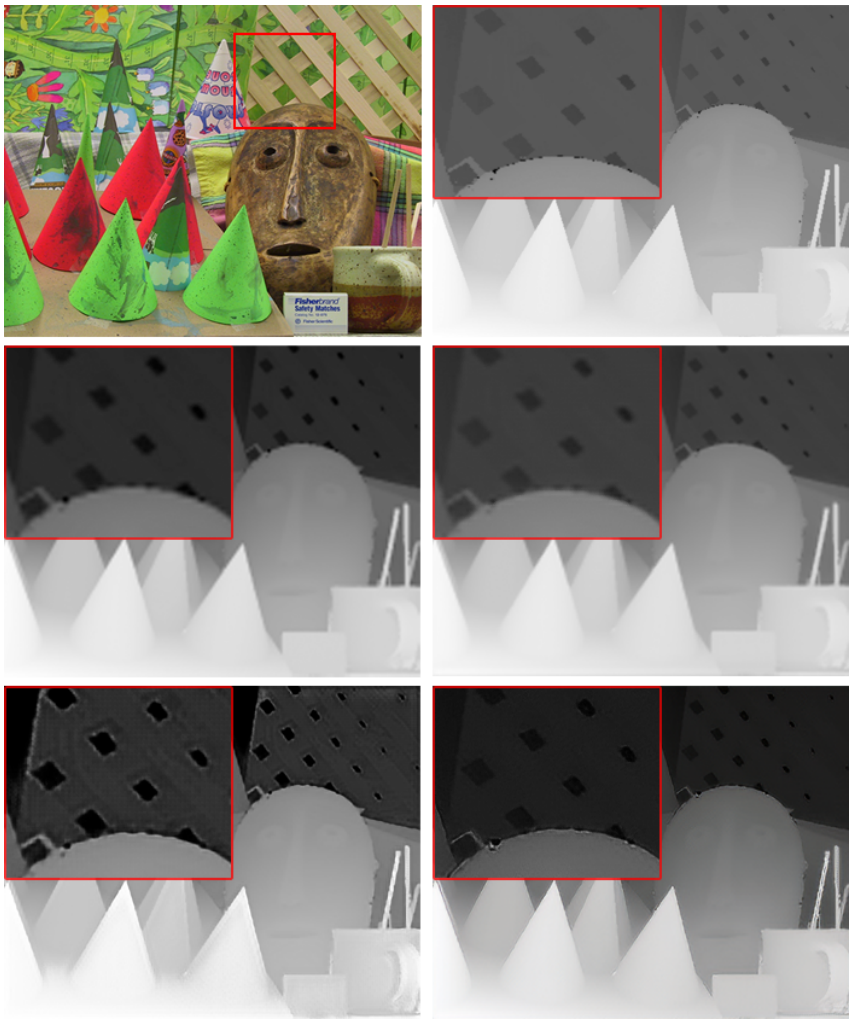



Depth Map Super-resolution: Comparison on Sintel

Comparison:

- Ground Truth
- BiCubic Interpolation
- Original Network
- Proposed Model

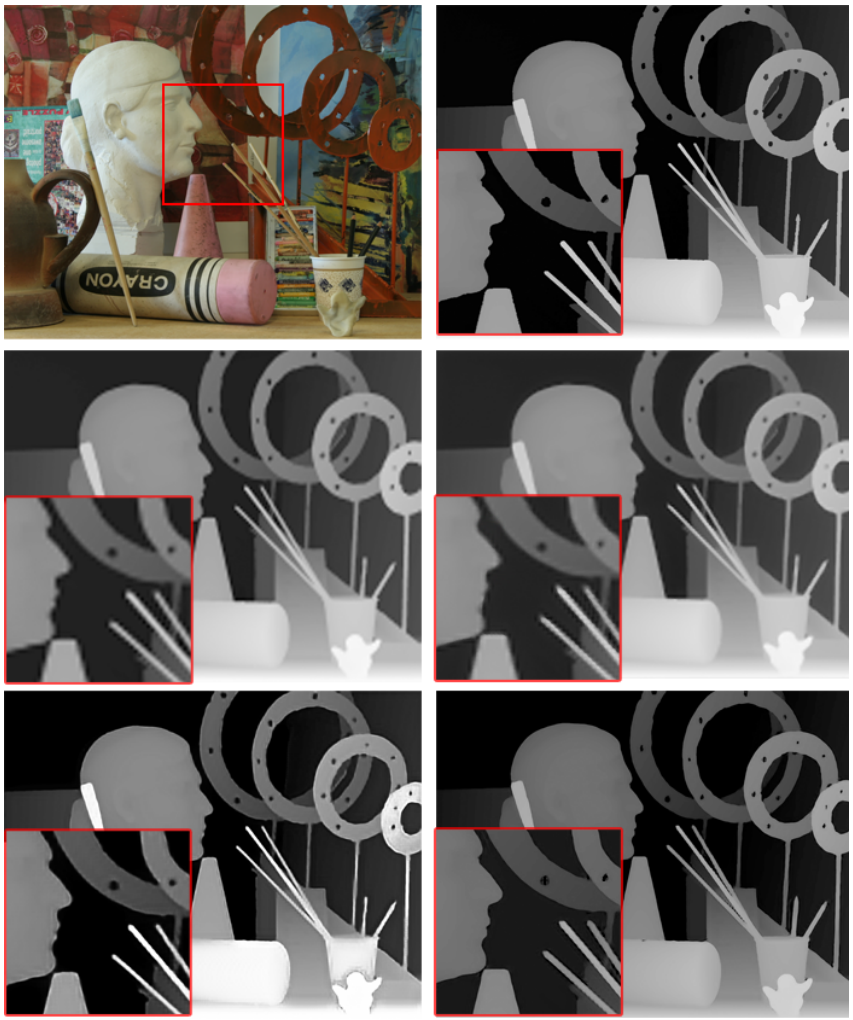
Depth Map Super-resolution: Results on Middlebury (1)



Comparison:

- Ground Truth
- Bicubic Interpolation
- Original Network
- Proposed Model
- RGB SRGAN [11]

Depth Map Super-resolution: Comparison on Middlebury (2)



Comparison:

- Ground Truth
- BiCubic
- Original Network
- Proposed Model
- DEN model (RGB guided)[26]

Conclusion:
Fast Depth Interpolation

- very fast, but very rough results; may be used as a part of user subsystem of depth guidance for naviation
- JtbD: faster input change, complex network, RGB/edge guidance.

Conclusion:
Depth Super-resolution

- Better results but certain artefacts for Depth Maps with high variance in log space.
- Better edges, worse textures.
- JtbD: complex network, change loss (Wasserstein loss, VGG loss for Discriminator).

Directions for further Study

Further development may be done in the following directions:

- More complex architectures, RGB/edge guidance
- Pix2Pix [7] integration
- Application, including depth inpainting methods, for diminished reality in real estate AR apps.

Research Papers

1. **Fast Semi-dense Depth Map Estimation** / Ilya Makarov, Alisa Korinevskaya, Vladimir Aliev // International ACM Conference on Multimedia Retrieval: RETech Workshop. — 2018.
2. [poster] **Super-resolution of Interpolated Downsampled Semi-dense Depth Map** / Ilya Makarov, Alisa Korinevskaya, Vladimir Aliev // International ACM Conference on 3D Web Technology. — 2018.
3. **Sparse Depth Map Interpolation using Deep Convolutional Neural Networks** / Ilya Makarov, Alisa Korinevskaya, Vladimir Aliev // 41st International Conference on Telecommunications and Signal Processing. — 2018.
4. [under review] **Fast Depth Map Super-Resolution using Deep Neural Network** / Ilya Makarov, Alisa Korinevskaya, Vladimir Aliev // International Symposium on Mixed and Augmented Reality. — 2018. (конференция уровня A* по рейтингу Core)

Thank You!

References I

[1] Daniel J. Butler и др. «A Naturalistic Open Source Movie for Optical Flow Evaluation». В: *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*. ECCV'12. Florence, Italy: Springer-Verlag, 2012, с. 611—625. ISBN: 978-3-642-33782-6. DOI: 10.1007/978-3-642-33783-3_44. URL: http://dx.doi.org/10.1007/978-3-642-33783-3_44.

[2] Chao Dong и др. «Image Super-Resolution Using Deep Convolutional Networks». В: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.2 (2016), с. 295—307. ISSN: 01628828. DOI: 10.1109/TPAMI.2015.2439281.

[3] David Eigen и Rob Fergus. «Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture». В: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, с. 2650—2658.

[4] David Ferstl и др. «Image guided depth upsampling using anisotropic total generalized variation». В: *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE. 2013, с. 993—1000.

[5] Jia-Bin Huang, Abhishek Singh и Narendra Ahuja. «Single image super-resolution from transformed self-exemplars». В: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, с. 5197—5206.

[6] Tak Wai Hui, Chen Change Loy и Xiaoou Tang. «Depth map super-resolution by deep multi-scale guidance». В: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9907 LNCS (2016), с. 353—369. ISSN: 16113349. DOI: 10.1007/978-3-319-46487-9{_}22.

[7] Phillip Isola и др. «Image-to-Image Translation with Conditional Adversarial Networks». В: *CoRR* abs/1611.07004 (2016). arXiv: 1611.07004. URL: http://arxiv.org/abs/1611.07004.

References II

[8] Martin Kiechle, Simon Hawe и Martin Kleinsteuber. «A joint intensity and depth co-sparse analysis model for depth map super-resolution». В: *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE. 2013, с. 1545—1552.

[9] HyeokHyen Kwon, Yu-Wing Tai и Stephen Lin. «Data-driven depth map refinement via multi-scale sparse representation». В: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, с. 159—167.

[10] Iro Laina и др. «Deeper depth prediction with fully convolutional residual networks». В: *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016* (2016), с. 239—248. DOI: 10.1109/3DV.2016.32.

[11] Christian Ledig и др. «Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network». В: *arXiv preprint* (2016). ISSN: 0018-5043. DOI: 10.1109/CVPR.2017.19. URL: http://arxiv.org/abs/1609.04802.

[12] Yanjie Li и др. «Joint example-based depth map super-resolution». В: *Multimedia and Expo (ICME), 2012 IEEE International Conference on*. IEEE. 2012, с. 152—157.

[13] Yiyi Liao и др. «Parse geometry from a line: Monocular depth estimation with partial laser observation». В: *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE. 2017, с. 5059—5066.

[14] Wei Liu и др. «An enhanced depth map based rendering method with directional depth filter and image inpainting». В: *Visual Computer* 32.5 (2016), с. 579—589. ISSN: 01782789. DOI: 10.1007/s00371-015-1074-2.

[15] Jiajun Lu и D. Forsyth. «Sparse depth super resolution». В: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. New York, NY, USA: IEEE, июнь 2015, с. 2245—2253. DOI: 10.1109/CVPR.2015.7298837.

References III

[16] Fangchang Ma и Sertac Karaman. «Sparse-to-dense: Depth prediction from sparse depth samples and a single image». В: *arXiv preprint arXiv:1709.07492* (2017).

[17] Ilya Makarov, Vladimir Aliev и Olga Gerasimova. «Semi-Dense Depth Interpolation Using Deep Convolutional Neural Networks». В: *Proceedings of the 2017 ACM on Multimedia Conference* (2017), с. 1407—1415. DOI: 10.1145/3123266.3123360. URL: http://doi.acm.org/10.1145/3123266.3123360.

[18] Jaesik Park и др. «High quality depth map upsampling for 3d-tof cameras». В: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE. 2011, с. 1623—1630.

[19] G. Ros и др. «The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes». В: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. New York, NY, USA: IEEE, июнь 2016, с. 3234—3243. DOI: 10.1109/CVPR.2016.352.

[20] Daniel Scharstein и Chris Pal. «Learning conditional random fields for stereo». В: *IEEE IC on CVPR*. NY, USA: IEEE, 2007, с. 1—8.

[21] Wenzhe Shi и др. «Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network». В: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, с. 1874—1883.

[22] Nathan Silberman и др. «Indoor Segmentation and Support Inference from RGBD Images». В: *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, с. 746—760. ISBN: 978-3-642-33715-4. DOI: 10.1007/978-3-642-33715-4_54. URL: http://dx.doi.org/10.1007/978-3-642-33715-4_54.

References IV

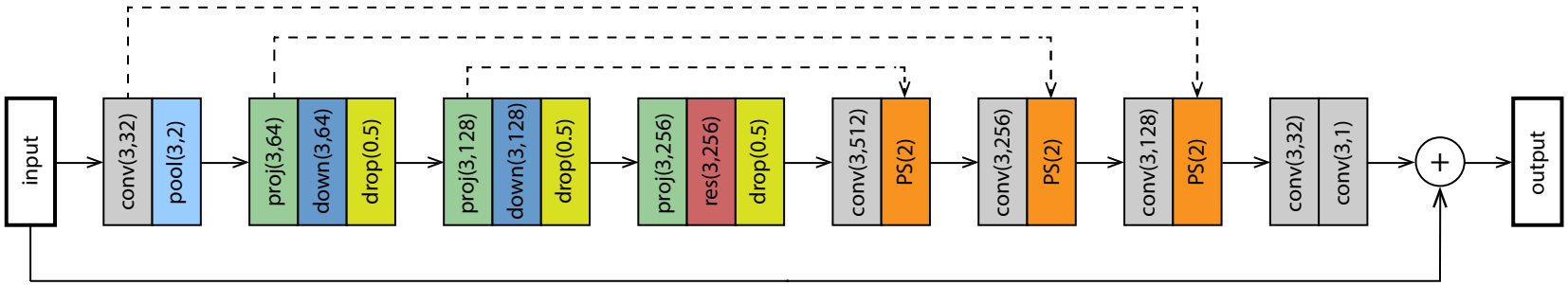
[23] Xibin Song, Yuchao Dai и Xueying Qin. «Deep depth super-resolution: Learning depth super-resolution using deep convolutional neural network». В: *Asian Conference on Computer Vision*. Springer. 2016, с. 360—376.

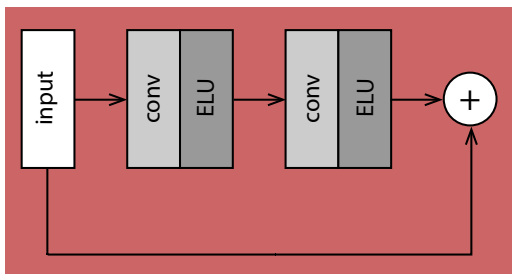
[24] Zhaowen Wang и др. «Deep networks for image super-resolution with sparse prior». В: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, с. 370—378.

[25] Jingyu Yang и др. «Depth recovery using an adaptive color-guided auto-regressive model». В: *European Conference on Computer Vision*. Springer. 2012, с. 158—171.

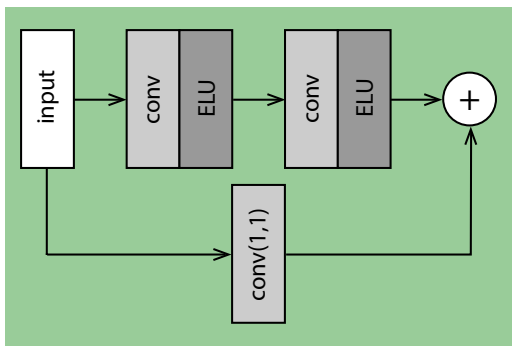
[26] Wentian Zhou, Xin Li и Daryl Reynolds. «Guided deep network for depth map super-resolution: How much can color help?». В: *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE. 2017, с. 1457—1461.

Fast Depth Interpolation: Interpolation Network

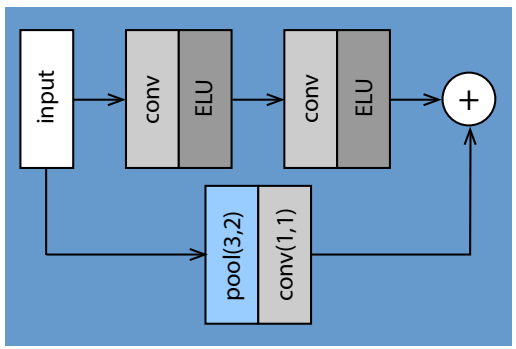




(a) residual block

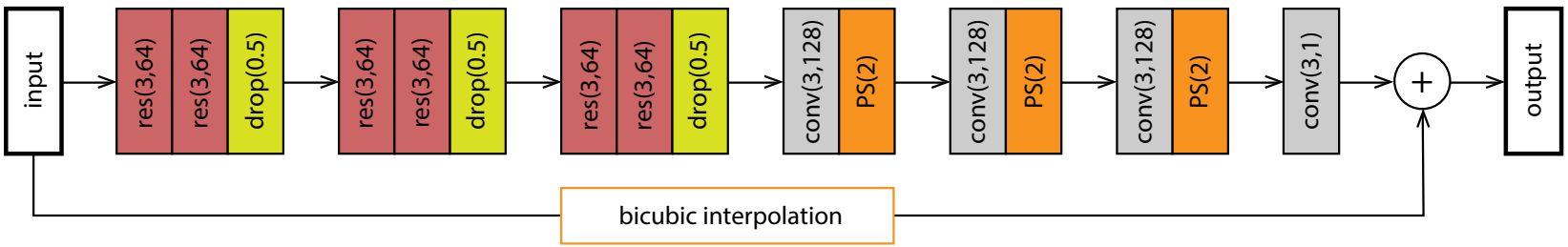


(b) projection block



(c) downsampling block

Fast Depth Interpolation: Super-resolution Network



- Trained on SYNTHIA
- Same loss functions

Fast Depth Interpolation: loss function

- MSE only leads to smoothing
- VGG loss preserves details

$$L = L_{MSE} + \alpha L_{VGG}$$

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

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

$y_{i,j}$ and $\hat{y}_{i,j}$ are ground truth and predicted depth values
$\Phi(\mathbf{y})$ – output from conv 3.1 layer of VGG-16

Fast Depth Interpolation: Results on KITTI

Comparison on KITTI Dataset.

Type	**Methods**	**rms**	**rel**
RGB	Eigen et al. [3]	7.156	0.190
RGB	Sparse-to-Dense [16]	6.266	0.208
RGBd	Sparse-to-Dense [16]	3.378	0.073
RGBd	Liao et al. [13]	4.500	0.113
d	Interpolation [17]	10.098	0.7509
d	Interpolation + SR	22.897	2.841

Fast Depth Interpolation: Results on NYU and SYNTHIA

Proposed Model

	NYU Depth V2		SYNTHIA	
Sampling	rms	rel	rms	rel
Uniform (p = 0.1)	0.944	0.323	251.669	0.585
Along gradient	0.971	0.331	260.428	0.653
Gradient + uniform	0.964	0.332	252.587	0.755
Regular grid	0.973	0.322	252.522	0.576

Interplation only

	NYU Depth V2		SYNTHIA	
Sampling	rms	rel	rms	rel
Unif	0.54	0.058	52.1	0.049
Grad	0.62	0.115	54.9	0.087
Grad + unif	0.35	0.048	46.1	0.038
Grid	0.69	0.237	71.7	0.246

Network Sparse-to-Dense from [16]

	NYU Depth V2		SYNTHIA	
Sampling	rms	rel	rms	rel
Unif	0.285	0.077	46.681	0.122
Grad	1.489	0.343	264.015	0.378
Grad + unif	0.277	0.079	40.980	0.136
Grid	0.445	0.120	64.25	0.156

Depth Super-resolution: loss function

$$L_G = L_{MSE} + \alpha L_{VGG} + \beta L_{GAN} + \gamma L_{TV}$$

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

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

$$L_{GAN} = \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D(G(Z_i))\right)$$

$$L_D = \frac{1}{m} \sum_{i=1}^{m} \left(\log D(X_i) + log(1 - D(G(Z_i))) \right)$$

$y_{i,j}$ and $\hat{y}_{i,j}$ are values the obtained and target depth maps,
Z_k is an input, low-resolution depth map,
X_k is a high-resolution depth map from the training data,
$\Phi(\mathbf{y})$ is the output of the conv 3.1 layer of the VGG-16,
m is the number of training samples.

Depth Super-resolution: Results on Middlebury. RMSE and SSIM

Comparison of Super-resolution (x8).

RMSE	Middlebury			
	Cones	Teddy	Tsukuba	Venus
Bicubic	9.64	5.33	7.52	2.65
NN	9.90	11.13	8.10	10.0
NN+gan_loss	9.35	6.48	5.05	5.58
Park et al [18]	7.73	7.15	15.1	2.99
Li et al [12]	8.90	7.24	15.8	5.76
Ferstl et al.[4]	7.14	5.39	17.2	4.04
Kiechle et al[8]	5.11	2.76	10.9	1.76
Kwon et al[9]	**2.37**	**2.19**	5.67	1.68
MSG-Net [6]	2.75	2.28	**3.64**	**0.76**
Kiechle et al[8]	4.52	2.37	10.0	1.16
Lu et al[15]	5.34	3.47	13.7	2.13
Song et al. [23]	4.59	2.88	11.6	1.71
SRCNN [2]	5.18	3.25	11.3	1.70
Wang et al. [24]	4.87	3.01	9.58	1.78
MS-Net [6]	5.22	2.87	9.99	0.88

SSIM	Middlebury			
	Cones	Teddy	Tsukuba	Venus
Bicubic	0.85	0.94	0.75	**0.98**
NN	0.91	0.87	0.74	0.91
NN+gan_loss	0.74	0.75	**0.91**	0.82
Park et al. [18]	0.92	0.95	0.81	**0.98**
Yang et al. [25]	0.94	0.95	0.86	**0.98**
Huang et al [5]	0.93	0.93	0.90	0.97
Song et al. [23]	**0.95**	**0.97**	**0.91**	**0.98**

Depth Super-resolution: Results on Middlebury. PSNR

PSNR	Middlebury				
	Cones	Teddy	Tsukuba	Venus	Art
Bicubic	19.52	**30.03**	23.92	**34.82**	**21.87**
NN [17]	**23.37**	25.35	22.62	27.14	21.59
GAN NN	18.28	18.50	**30.18**	28.87	19.35

Datasets

- **SYNTHIA** [19] – synthetic road scenes, train 8000 samples, test 4000 samples.
- **NYU Depth** [22] – sensor-based for indoors scenes, 2500 samples.
- **Sintel** [1] – synthetic dataset, 1500 samples.
- **Middlebury** [20] — 12 samples.

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

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

VGGNet (2014)

VGG-16

Subpixel Convolution (Pixel-Shuffle Layer)

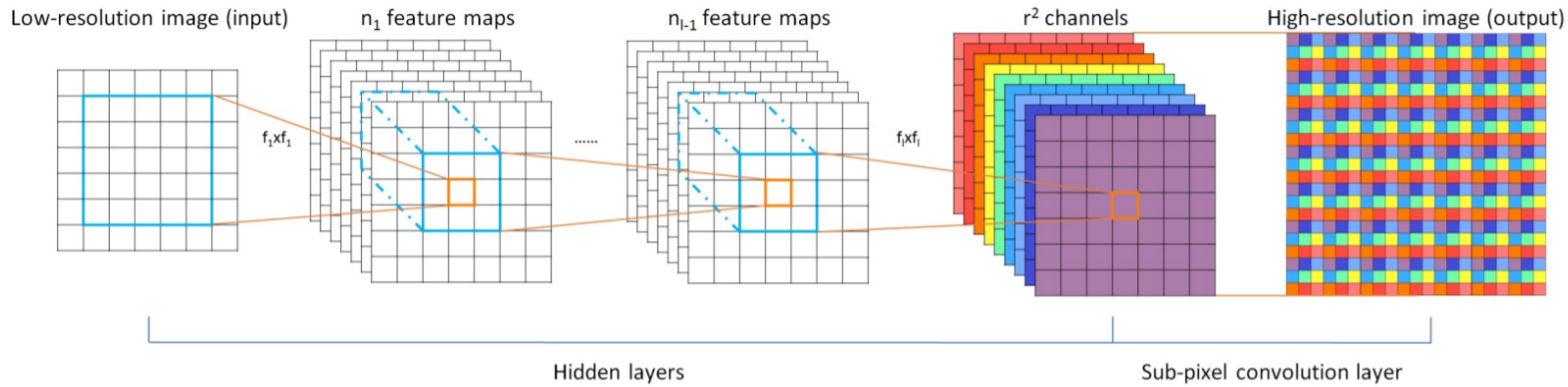
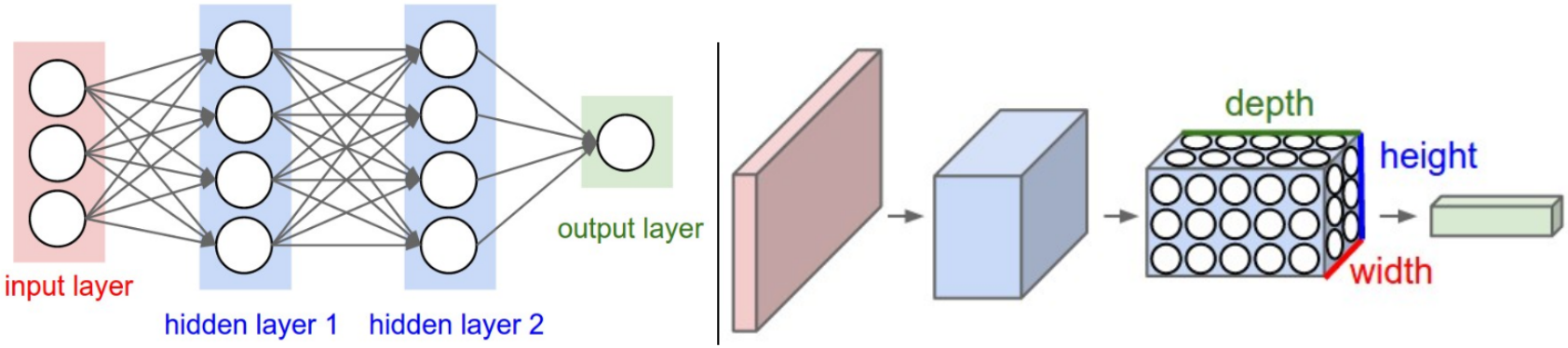
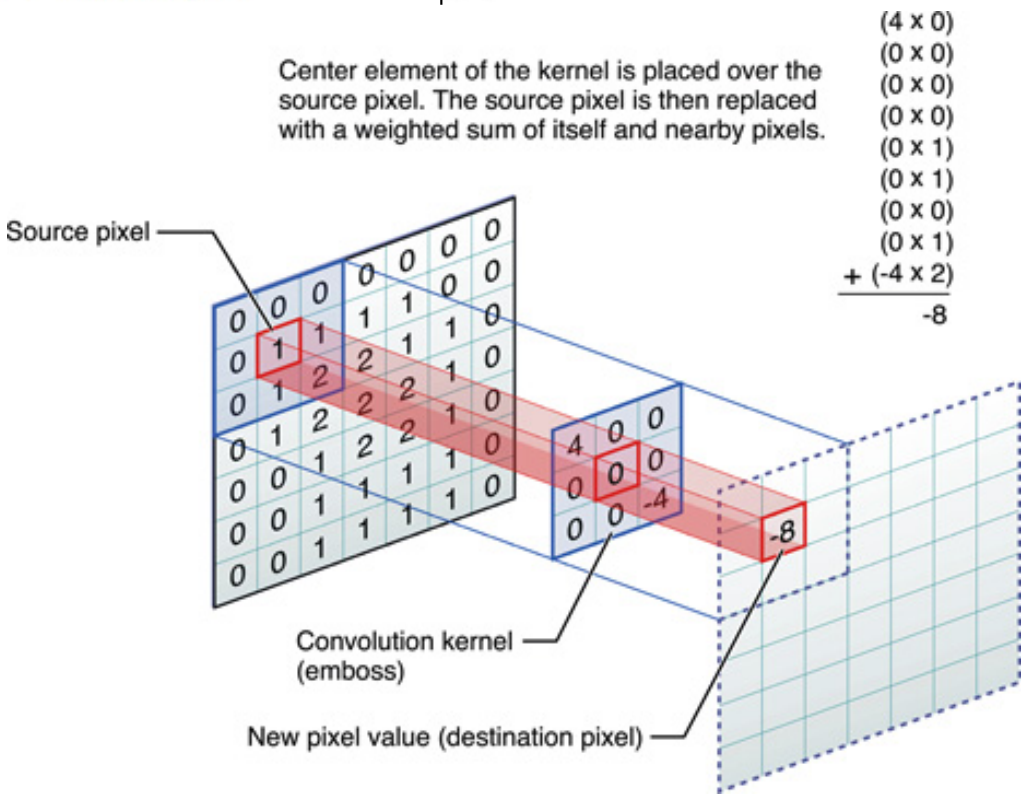


Figure 1. The proposed efficient sub-pixel convolutional neural network (ESPCN), with two convolution layers for feature maps extraction, and a sub-pixel convolution layer that aggregates the feature maps from LR space and builds the SR image in a single step.

Схема из Shi et al. [21].

Convolutional Neural Networks





TV regularization




Scheme from Estrela et al.

Total Variation (TV) –

$$TV(f) = \sup_{\mathcal{P}} \sum_{i=0}^{n_p-1} |f(x_{i+1}) - f(x_i)|$$

f defined on $[a, b] \subset \mathbb{R}$

$$\mathcal{P} = \Big\{ P = \{x_0, ..., x_{n_p}\} | \text{ P – partition of [a,b] } \Big\}$$