IMAGE GUIDED DEPTH ENHANCEMENT VIA DEEP FUSION AND LOCAL LINEAR REGULARIZATION

Jiang Zhu   *Jing Zhang*   *Yang Cao*   *Zengfu Wang*

Department of Automation, University of Science and Technology of China
{zj130129, zjwinner}@mail.ustc.edu.cn, {forrest, zfwang}@ustc.edu.cn

ABSTRACT

Depth maps captured by RGB-D cameras are often noisy and incomplete at edge regions. Most existing methods assume that there is a co-occurrence of edges in depth map and its corresponding color image, and improve the quality of depth map guided by the color image. However, when the color image is noisy or richly detailed, the high frequency artifacts will be introduced into depth map. In this paper, we propose a deep residual network based on deep fusion and local linear regularization for guided depth enhancement. The presented scheme can effectively extract the correlation between depth map and color image in the deep feature space. To reduce the difficulty of training, a specific layer of network which introduces a local linear regularization constraint on the output depth is designed. Experiments on various applications, including depth denoising, super-resolution and inpainting, demonstrate the effectiveness and reliability of our proposed approach.

Index Terms— depth enhancement, deep residual network, local linear regularization, deep feature space

1. INTRODUCTION

High-quality depth maps are very important in computer vision applications, such as 3DTV, 3D reconstruction, hand pose estimation and robot navigation. With the development of consumer depth cameras such as Kinect and Xtion Pro, it is more convenient to capture the depth map of scene. However, depth maps captured by these devices are often noisy and incomplete at edge regions due to the reflection and absorption of structure light or the viewpoint disparity between multi-sensors.

Since depth cameras often provide a pair of color and depth (RGBD) image of a scene, various methods [1] [2] [3] [4] [5] [6] have been proposed to enhance the quality of depth under the guidance of color image. These methods are based on the assumption that there is a co-occurrence of edges in depth map and its corresponding color image [7] [8]. However, this assumption does not always hold in practice. When the color image is noisy or richly detailed, the high frequency component of color image will be introduced into depth map. These generated artifacts will seriously affect the quality of depth map.

Many works introduce deep learning into various image processing applications, such as image denoising [9], image super-resolution [10] [11] [12] [13], and image restoration [14]. Their successes benefit from underlying-feature representative ability of deep neural networks. Very recently, Zhang et al. [15] proposed a deep CNN based method which demonstrates the feasibility of an end-to-end approach to depth enhancement. Their method uses max pooling to widen receptive field. However, the output depth map is blurry. For depth enhancement applications, such as denoising and super-resolution, the sharp edges of depth map are very important.

In this paper, we propose a CNN-based framework for guided depth enhancement. Our goal is to learn the underlying correlation between depth maps and color images, and then to use the correlation to enhance the quality of depth map. To achieve this goal, we need to overcome the following challenges. 1) To preserve the sharp edges in depth map, no pooling can be used in the proposed network. An alternative solution to widen the receptive field is required. 2) One possible solution to improve the performance of deep CNN is to increase the network depth by adding new layers. However, this will introduce more parameters and increase the risk of overfitting. 3) Depth map and color image have different noise level and data distribution. It will not work well for directly using depth map and color image as the input to jointly train the network. A feature level fusion strategy is required for the learning process.

To address the above challenges, we propose a deep residual network for guided depth enhancement. The proposed deep network consists of three components: depth branch, intensity branch and deep feature fusion part. Fig.1 illustrates the details of our proposed network. Our proposed framework has the following advantages:

1) We present a deep residual network for image guided depth enhancement, which can well extract the underlying correlation between depth map and color image in the deep feature space.

2) A deep fusion strategy is adopted. The low-level features for depth and color image are firstly extracted in each

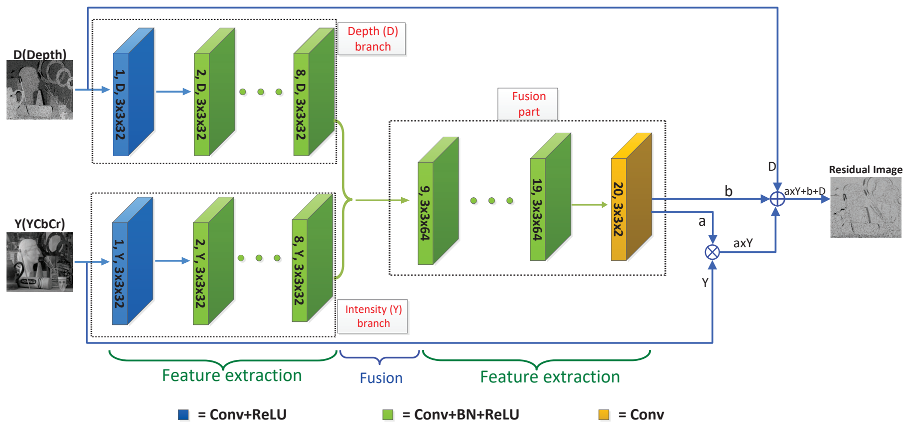


Fig. 1. The architecture of our proposed deep residual convolutional neural network.

branch, respectively. And then the features are combined together and fused into high-lever features. This scheme effectively suppresses the influence of different noises and data distributions in depth map and color image.

3) To avoid overfitting, a specific layer which introduces a local linear regularization constraint on the output depth is presented. It significantly improves convergence speed and accuracy.

2. PROPOSED METHOD

2.1. Proposed architecture

We propose a deep residual network which employs a single residual unit to learn the residual image, as shown in Fig.1. We refer to the proposed deep residual network for guided depth enhancement as DRECNN. Our network consists of three components: depth branch, intensity branch and deep feature fusion part. Moreover, to achieve a wide receptive field, we set the number of convolution layers to 20 for all of the depth enhancement applications.

The Conv, ReLU, BN and SR represent convolution layers, rectified linear units, batch normalization layer and a specific regularized layer, respectively. There are three types of layers in our network. (1) Conv+ReLU: using in the first layer of depth branch (D) and intensity branch (Y). (2) Conv+BN+ReLU: using from the second layer to the nineteenth layer. (3) Conv+SR: using in the last layer to reconstruct the output depth map. The specific regularized layer (SR) will be introduced in detail in Sec. 2.3.

2.2. A deeply fusion strategy for joint-feature extraction

Depth map and color image have different noise level and data distribution. Directly using depth map and intensity image as the input of network will bring great interference to training process. To overcome this problem, we propose a deep fusion strategy for joint-feature extraction.

In our proposed network, the first eight layers extract the low-level features of the depth maps and the intensity images, respectively. Then the next eleven layers combine the two types of low-level features together and fuse into the high-level joint-features. This deep fusion strategy can efficiently extract the correlation between depth map and color image even in the presence of high-level noises.

2.3. A local linear regularization constraint

The image guided filter [8] starts from a local linear model as follows:

$$q_i = a_k I_i + b_k, \forall i \in \omega_k. \tag{1}$$

Here *I* is a guidance image and q is the filter output. The ω_k is the filter window and (a_k, b_k) are the linear coefficients which are assumed to be constant in the window ω_k. The filter output q is a linear transform of the guidance image *I*.

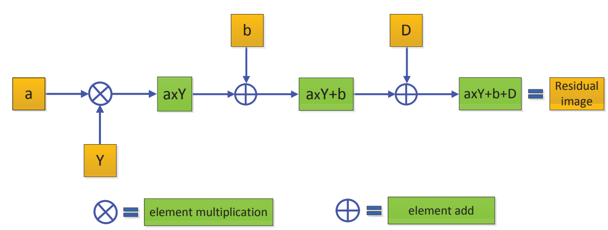


Fig. 2. The structure of the local linear regularization constraint on the output depth.

Inspired by the local linear model, we propose a specific regularization (SR) layer which is shown as Fig.2. Specifically, we first insert a convolution layer to output the linear coefficient map a and b, and then compute $aY + b$ as the results of linear model. Finally, the residual map $aY+b-D$ is computed and supervised by the ground truth label (We use aY+b+D instead in our network. Indeed, it is an equivalent expression as aY+b-D, since the network will learn the opposite linear coefficients adaptively.). The SR layer is exploited in the output part of our proposed network, which serves as a constraint on the output depth map. This method can effectively boost the performance of depth enhancement and reduce the risk of overfitting.

2.4. Loss function definition

For the learning process of our proposed networks, we use a Euclidean-based distance function as the loss function:

$$L(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \|F(D_i, Y_i; \Theta) - (D_i - D_i^{GT})\|_F^2. \quad (2)$$

We denote F as the mapping function and Θ as our network parameters, where D_i is the input depth map, Y_i is the input intensity image, D_i^{GT} is the ground truth depth map, and N represents the amount of the training samples.

3. APPLICATIONS AND EXPERIMENTAL RESULTS

In this part, we conducted a series of experiments on various applications including depth denoising, super-resolution and inpainting to demonstrate the effectiveness and reliability of our proposed DRECNN.

3.1. Data preprocessing and augmentation

Training and testing data: We used the same training dataset for all of the depth enhancement applications. We chose 40 RGBD images from MPI Sintel depth dataset [16] and 21 RGBD images from Middlebury dataset [17] [18] [19]. In the training images, 55 images were used for training and the remaining 6 images were used for validation. Then we chose the rest images of Middlebury dataset which are not used in the training as the test images. Note that we used bilateral filter [20] to pre-process the groundtruth depth maps of Middlebury dataset since it still had some missing values.
Data augmentation: For data augmentation, we rotated the original training images with 90°, 180°, 270° and also flipped them upside-down. In this way, we obtained 8× more samples.

3.2. Experimental setting

Parameter setting : We set the same parameter for different applications. The size of the patches was 50×50. We used Adam algorithm to train our DRECNN. The momentum was 0.9, a mini-batch size was 128 and the weight decay was 0.0005. We set the learning rate from $1e-3$ to $1e-6$ which was gradually decayed for $1.8e+5$ iterations. The kernel size of convolution layers was 3×3 and we padded zeros for convolution layers to keep the same size as the input depth.
Network training : Our DRECNN was implemented using caffe [21]. We used a Tesla M40 GPU for the training and testing stage. To reduce training time and memory burden, we converted the input color images from RGB to YCbCr and only used the luminance component (Y) for training.

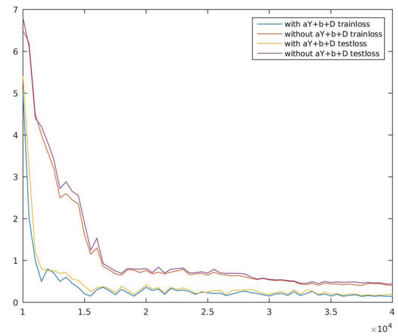


Fig. 3. The training and test loss of two different methods.

3.3. Performance analysis on regularized layer

To demonstrate the significance of the specific regularized (SR) layer, we performed two contrasting experiments. One experiment used the regularized layer to add constraint on the output depth, the other one used the normal output layer. Fig.3 presented the two training results. As can be seen, the regularized layer significantly improved the convergence speed. Both training loss and test loss were much lower compared with using normal output layer.

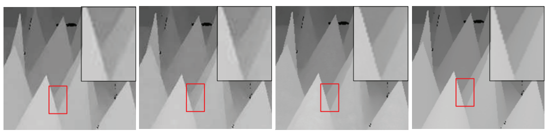
3.4. Experimental results on various applications

A. Depth Blind Denosing
For depth blind denosing, we added different levels of additive white Gaussian noise(AWGN) to the depth map pathes of the training dataset, where the noise level σ was set as 10, 15, 20, 25 and 30. We also added the additive white Gaussian noise(AWGN) to intensity image (Y) with standard variance of 15. The denoising comparisons with the state-of-the-art methods in terms of PSNR are summarized in table 1 (The best results are highlighted in bold font). And the visual insepection results are shown in Fig.4.

As can be seen in the Table 1, our proposed network produced superior denoising results than the state-of-the-art methods such as BM3D [22] and NLGBT [23] in terms of PSNR. As shown in Fig.4, our proposed method reduced artifacts effectively while preserving more details. Moreover, our DRECNN achieved better performance in presence of high-level noises.
B. Depth Super-resolution

Table 1. Denosing comparison in terms of PSNR (dB)

Images	Methods	σ=10	σ=15	σ=20	σ=25	σ=30
	BM3D [22]	40.56	37.49	35.28	33.81	32.75
Cones	NLGBT [23]	42.84	39.18	36.53	34.43	32.97
	DRECNN	**43.15**	**40.53**	**39.02**	**37.46**	**36.02**
	BM3D [22]	41.36	38.33	36.12	34.45	33.25
Teddy	NLGBT [23]	42.29	39.38	36.71	34.62	33.42
	DRECNN	**42.84**	**41.22**	**39.64**	**38.31**	**36.90**
	BM3D [22]	46.04	43.51	41.84	40.16	39.13
Sawtooth	NLGBT [23]	48.41	45.30	43.22	41.71	40.01
	DRECNN	**49.05**	**46.28**	**45.14**	**43.21**	**42.06**



(a)BM3D / 40.56dB (b)NLGBT / 42.84dB (c)DRECNN / **43.15**dB (d)Ground Truth

Fig. 4. Denoising results of the Cones with noise level 10.

For depth map super-resolution, the high-resolution depth maps were down-sampled and up-sampled with the scaling factors 4 and 8 using the method of Mandal et al. [24]. Additive white Gaussian noise(AWGN) with standard variance of 5 was added to the low-resolution depth maps. We also added the additive white Gaussian noise(AWGN) to intensity image (Y) with standard variance of 15.

The comparison results with the state-of-the-art methods in terms of the root mean squared errors (RMSE) are summarized in Table 2, where we refer to Mandal et al.'s method [24] and Aodha et al.'s method [25] to PS and EB. The visual inspection results are shown in Fig.5. As can be seen, our proposed method achieved better performance with the larger scaling factors. Moreover, our method suppressed the noise effectively while preserving the sharp edge.

Table 2. Depth super-resolution comparison in terms of RMSE for σ= 5

	Scale-4			Scale-8		
Images	EB [25]	PS [24]	DRECNN	EB [25]	PS [24]	DRECNN
Aloe	8.09	5.73	**5.09**	13.05	9.05	**7.59**
Baby	5.06	3.78	**3.05**	8.43	5.64	**4.83**
Cones	6.11	4.49	**4.23**	10.00	6.94	**5.67**
Plastic	4.47	3.19	**2.42**	8.32	4.65	**3.25**
Teddy	5.04	3.77	**3.50**	8.38	5.50	**4.32**
Venus	2.66	2.63	**2.18**	4.35	3.43	**2.96**

C. Depth Inpainting

For depth inpainting, we used the method of LRMC [26] to pre-process the depth maps as training data. Moreover, we added additive white Gaussian noise(AWGN) with standard variance of 25 to the intensity image (Y). The comparison results with the state-of-the-art methods in term of the PSNR are summarized in Table 3 (The best results are highlighted in bold font). The visual inspection results are shown in Fig.6.

In most cases, our DRECNN achieved higher PSNR than

(a)EB / 8.38 (b)PS / 5.50 (c)DRECNN / **4.32** (d)Ground Truth

Fig. 5. Depth map super-resolution results of the Teddy with upscaling factor 8.

the state-of-the-art methods. As can be seen in Fig.6, DE-CNN [15] tends to generate blurry edges because of max pooling in their network. Our DRECNN produced sharper edges and preserved more details than the other methods, e.g. LRMC [26] and Shen et al.'s method [27].

Table 3. Depth inpainting comparison in terms of PSNR (dB)

Methods	Art	Dolls	Reindeer	Laundry	Baby	Wood1	Teddy
DE-CNN [15]	33.84	40.67	35.36	38.11	40.28	40.84	39.77
LRMC [26]	33.10	41.86	36.13	38.59	39.00	39.96	40.77
Shen et al. [27]	33.42	**42.89**	37.19	38.71	39.32	41.54	40.89
DRECNN	**34.02**	41.59	**38.18**	**39.17**	**41.13**	**41.74**	**41.21**

(d)LRMC / 39.00dB (e)Shen et al. / 39.32dB (f)DRECNN / **41.13**dB

Fig. 6. Depth map inpainting results of the Baby.

4. CONCLUSION

In this paper, we proposed a deep residual network-based framework for image guided depth enhancement, in which a deep feature fusion strategy and a specific regularization layer are adopted. The deep feature fusion strategy could suppress the influence of different noises distributions in depth map and color image. And the local linear regularization constraint on the output depth could improve the training speed and precision. The experimental results on various applications, including depth denoising, super-resolution and inpainting, demonstrated that our network could produce favorable depth enhancement results even with high-level noises.
Acknowledgments. This work is supported by the Natural Science Foundation of China(61472380).

5. REFERENCES

[1] Jing Zhang, Yang Cao, Zheng Jun Zha, Zhigang Zheng, Chang Wen Chen, and Zengfu Wang, "A unified scheme for super-resolution and depth estimation from asymmetric stereoscopic video," *IEEE Transactions on Circuits & Systems for Video Technology*, vol. 26, no. 3, pp. 479–493, 2016.

[2] Jing Zhang, Yang Cao, and Zengfu Wang, "A new image filtering method: Nonlocal image guided averaging," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2460–2464.

[3] Yuxiang Yang, Mingyu Gao, Jing Zhang, Zhengjun Zha, and Zengfu Wang, "Depth map super-resolution using stereo-vision-assisted model," *Neurocomputing*, vol. 149, pp. 1396–1406, 2015.

[4] Xiaowei Deng and Xiaolin Wu, "Sparsity-based depth image restoration using surface priors and rgb-d correlations," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3881–3885.

[5] Wei Liu, Xiaogang Chen, Jie Yang, and Qiang Wu, "Robust color guided depth map restoration," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 315–327, 2017.

[6] Weisheng Dong, Guangming Shi, Xin Li, Kefan Peng, Jinjian Wu, and Zhenhua Guo, "Color-guided depth recovery via joint local structural and nonlocal low-rank regularization," *IEEE Transactions on Multimedia*, vol. 19, no. 2, pp. 293–301, 2017.

[7] Jing Zhang, Yang Cao, Shuai Fang, Yu Kang, and Chang Wen Chen, "Fast haze removal for nighttime image using maximum reflectance prior," in *IEEE CVPR*, 2017.

[8] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," in *European conference on computer vision*. Springer, 2010, pp. 1–14.

[9] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. PP, no. 99, pp. 1–1, 2017.

[10] Ying Tai, Jian Yang, and Xiaoming Liu, "Image super-resolution via deep recursive residual network," in *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Honolulu, HI, July 2017.

[11] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *IEEE Conferene on Computer Vision and Pattern Recognition*, 2017.

[12] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1637–1645.

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in Neural Information Processing Systems*, 2016, pp. 2802–2810.

[15] Xin Zhang and Ruiyuan Wu, "Fast depth image denoising and enhancement using a deep convolutional network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2499–2503.

[16] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black, "A naturalistic open source movie for optical flow evaluation," in *European Conference on Computer Vision*. Springer, 2012, pp. 611–625.

[17] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[18] Daniel Scharstein and Chris Pal, "Learning conditional random fields for stereo," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[19] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on Pattern Recognition*. Springer, 2014, pp. 31–42.

[20] Carlo Tomasi and Roberto Manduchi, "Bilateral filtering for gray and color images," in *Computer Vision, 1998. Sixth International Conference on*. IEEE, 1998, pp. 839–846.

[21] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[22] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian, "Image denoising with block-matching and 3d filtering," in *Electronic Imaging 2006*. International Society for Optics and Photonics, 2006, pp. 606414–606414.

[23] Wei Hu, Xin Li, Gene Cheung, and Oscar Au, "Depth map denoising using graph-based transform and group sparsity," in *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*. IEEE, 2013, pp. 001–006.

[24] Srimanta Mandal, Arnav Bhavsar, and Anil Kumar Sao, "Depth map restoration from undersampled data," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 119–134, 2017.

[25] Oisin Mac Aodha, Neill DF Campbell, Arun Nair, and Gabriel J Brostow, "Patch based synthesis for single depth image super-resolution," in *European Conference on Computer Vision*. Springer, 2012, pp. 71–84.

[26] Si Lu, Xiaofeng Ren, and Feng Liu, "Depth enhancement via low-rank matrix completion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3390–3397.

[27] Xiaoyong Shen, Chao Zhou, Li Xu, and Jiaya Jia, "Mutual-structure for joint filtering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3406–3414.