GUIDED DEEP NETWORK FOR DEPTH MAP SUPER-RESOLUTION : HOW MUCH CAN COLOR HELP?

Wentian Zhou *Xin Li* *Daryl Reynolds*

Lane Department of Computer Science and Electrical Engineering
West Virginia University

ABSTRACT

Since the quality of depth maps produced by Time-of-Flight (TOF) cameras is low, color-guided recovery methods have been proposed to increase spatial resolution and suppress unwanted noise. Despite successful applications of deep neural networks in color image super-resolution (SR), their potential for depth map SR is largely unknown. In this paper, we present a deep neural network architecture to learn the end-to-end mapping between low-resolution and high-resolution depth maps. Furthermore, we introduce a novel color-guided deep Fully Convolutional Network (FCN) and propose to jointly learn two nonlinear mapping functions (color-to-depth and LR-to-HR) in the presence of noise. Experimental results on several benchmark data sets show that our method outperforms several existing state-of-the-art depth SR algorithms. Moreover, this work attempts to partially shed some light onto the fundamental question in color-guided depth recovery — how much can color help in depth SR?

Index Terms— Depth map super-resolution, colorguided depth recovery, deep neural network

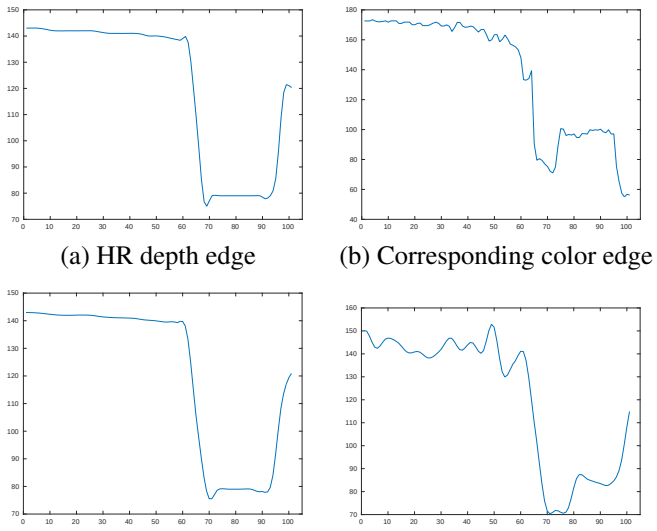
1. INTRODUCTION

Acquiring high-quality depth maps is a fundamental challenge for many vision related tasks, such as intelligent vehicles, gesture recognition, and 3D model rendering. In the past decade several model-based depth map super-resolution (SR) methods have been developed to improve the quality of depth maps. Diebel *et al.* [1] formulated depth map SR as an optimization problem, and integrated low-resolution depth maps with high-resolution color images using Markov Random Field (MRF). Park *et al.* [2] introduced non-local means into MRF to regularize depth maps. They also incorporated an edge-weighting scheme based on color image to preserve fine structural details. Ferstl *et al.* [3] proposed a total generalized variation model to regularize depth maps through an anisotropic diffusion tensor obtained from the color image. Zhou *et al.* [4] formed a dictionary by finding K-nearest neighbors (KNN) for each depth patch under the guidance of its corresponding color image, and iteratively solved a simultaneous sparse coding problem to refine depth details. Despite the success of these color guided approaches, the fundamental question - *how much can color help?* - largely remains open.

Recent breakthroughs in deep learning or deep neural networks have led to state-of-the-art performance in various computer vision applications including both low-level and high-level tasks. Under the context of single image super-resolution (SR) [5] [6] [7], Dong *et al.* [8] proposed a simple end-to-end deep convolutional neural network (CNN) to learn nonlinear mapping between low-resolution (LR) and high-resolution (HR) natural images. This work achieved excellent performance and inspired a deeper CNN architecture proposed by Kim *et al.* [9] as well as another deep recursive neural network [10] for image SR. These methods successfully demonstrate that end-to-end nonlinear mapping can be learned between low-resolution color images and their corresponding high-resolution images. However, these network models cannot be directly applied to depth map SR because depth maps acquired by TOF cameras have different intrinsic properties.

In this paper, we propose a novel CNN architecture to tackle *learning-based* color-guided depth map SR problem. For depth maps distorted by noise, denoising and SR are treated jointly in our problem formulation. The first part of our network consists of a series of fully convolutional layers to estimate missing high-frequency and noise components simultaneously. We call this part of the network "Depth Enhancement Network" (DEN). The second part of our network is designed to explicitly exploit the structural correlation between color images and depth maps. Inspired by [11], which attempts to predict a depth map from a single color image, we propose to utilize the HR color image as a prediction network. This part of the network is called "Color-based Prediction Network" (CBPN). With two independent and competing networks, we also address the issue of auto-merging before the final reconstruction of depth maps. The proposed network architecture enables us to at least partially shed some light onto the aforementioned fundamental question.

Under the framework of deep CNN, the amount of reliable information (e.g., high-frequency components) in train-



(a) HR depth edge (b) Corresponding color edge

(c) Interpolated, noise free (d) Interpolated with noise

Fig. 1. Comparison of a typical depth edge with its counterparts. High-frequency components in (c) is the most reliable; including (b) when SR (c) only confuses the network. On the contrary, including (b) when SR (d) provides relatively more reliable high-frequency components.

ing images directly impacts the learning outcomes especially for low-level vision tasks such as SR. To see this, we have shown a visual comparison of depth-color pair along with the interpolated depth profiles with and without noise contamination in Fig.1. It can be observed that when noise is absent, only small amount of high-frequency components are missing in the interpolated depth map (Fig.1c); therefore the help from the corresponding color image (Fig.1b) is limited. Moreover, depth values in the noisy interpolated depth map (Fig.1d) are only partially preserved - precise locations of depth discontinuities are severely distorted. In this situation, the high-frequency components contained in the supplementary color image could be highly useful. Apparently, how much can color help is question that is difficult to answer in traditional model-based color guided depth recovery framework; deep CNN at least suggests an alternative learning-based approach to joint depth SR and noise suppression.

Our contributions are summarized as follows:

1. We propose to directly learn end-to-end nonlinear mapping for the depth map super-resolution problem under the framework of Convolutional-Neural-Network.

2. By analyzing the characteristics of low-resolution depth maps and color images, we observe that color images are more helpful when noise is present and/or the scale factor is large.

3. We propose a color-based prediction network for the learning processes that need color guidance. The combined network leverages the contribution from depth maps and color images automatically to finalize the nonlinear model.

2. DEPTH SUPER-RESOLUTION NETWORK

The goal of our proposed network is to learn a nonlinear mapping that describes the relationship between LR and HR depth maps. Our network consists of four components. The first component is a depth enhancement network (DEN), which estimates the missing high-frequency components from the LR depth map. The importance of estimating missing high-frequency components has been discussed in [9] and [12]. The second component is a color-based prediction network (CBPN), which predicts the high-frequency components for the HR depth map. The third component is an auto-merging part, in which feature maps produced by DEN and CBPN are automatically combined into a set of new feature maps. The last component of our network reconstructs the HR depth map from the merged feature maps. Note that CBPN and auto-merging become more important when the LR depth map becomes less reliable (e.g., due to presence of noise). A graphical illustration of the proposed network is shown in Fig. 2.

Formally, depth SR is formulated as a problem of estimating a HR depth map $D_{\mathrm{HR}} \in \mathbb{R}^{sM \times sN}$ from its LR counterpart $D_{\mathrm{LR}} \in \mathbb{R}^{M \times N}$. Scaling factors s of $2, 4$, or 8 were commonly used in previous studies. Instead of supplying D_{LR} directly as the network input, we pre-process it with bicubic interpolation to reach the target resolution $D_{\mathrm{bic}} \in \mathbb{R}^{sM \times sN}$. Such a step of preprocessing dramatically reduces the computational burden for network training and helps relax the constraint on the input size. When D_{LR} is unreliable, we transform the RGB color image to a YCrCb image and only feed the Y channel Y_{HR} into the network. The final objective of network learning is to find a set of optimal parameters $\Theta = \{\mathbf{W}, \mathbf{B}\}$ such that the following loss function

$$L(\Theta) = \frac{1}{2n} \sum_{i=1}^{n} \| f(\Theta, D^i_{\mathrm{bic}}, Y^i_{\mathrm{HR}}) - D^i_{\mathrm{HR}} \|_2^2, \qquad (1)$$

is minimized, where $f(\cdot)$ estimate the HR depth map based on $\mathbf{W}$, $\mathbf{B}$, *i.e.,* weights and biases, and n is the total number of the training samples.

2.1. Depth Enhancement Network

Deep CNN is one of the most commonly used architectures in the literature of deep learning. It has shown state-of-the-art performance in various vision tasks including image SR. In Fig.2 we have shown the construction of our own DEN using 10 convolutional layers along with a rectified linear unit (ReLU)[13] after each layer. We opt to keep our network as a 10-layer network for the purpose of balancing output accuracy and training resources. It is worth noting that more

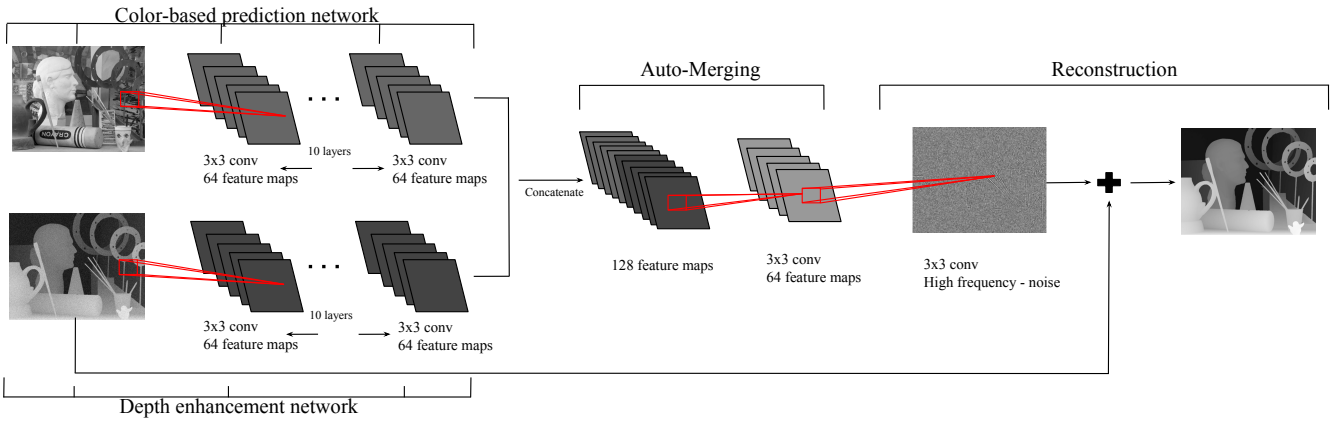


Fig. 2. Network illustrations. LR depth map and HR color image are the inputs of DEN and CBPN. Auto-merging takes the outputs of these networks to estimate the missing high-frequency component and the noise. The final output of this network is the reconstructed HR depth map.

convolutional layers can be used to boost performance, at the price of increased computational complexity.

For each convolutional layer, we use an array of 3×3 kernels to generate 64 feature maps. We then pass the feature maps through the ReLU activation function. In order to keep the feature maps the same size as the inputs, zero-padding is employed at each layer. Therefore, each pixel in the feature maps generated by the 10th layer has a receptive field of 21×21 pixels in the input depth map. After the network construction, bicubic-interpolated and ground-truth depth maps are used as inputs and outputs to train the nonlinear mapping. Note that in DEN, only high-frequency components are learned from the LR depth map because the bicubic-interpolated depth map will be fed forward to the final reconstruction step, as shown in Fig.2.

2.2. Color-based Prediction Network

By comparing edges in Fig.1b with Fig.1d, we can use the differences between them complementarily. Depth values in Fig.1d are much more reliable than blind guesses from a color image. On the other hand, edge locations in Fig.1b are relatively more reliable than the corrupted depth edge. Such mutual complementary characteristics certainly can enhance accuracy and sharpness around depth edges.

Inspired by [11], we utilize color images to predict the locations of depth discontinuities. To fairly evaluate the contribution of color images, we employ a network structure that is identical to DEN. As shown in Fig.2, the only difference between CBPN and DEN is the input. During the experiments in Sec.3, Y_{HR} is chosen as the input to CBPN to eliminate the effect of color variation. A set of 64-feature maps is generated for Y_{HR}, which contains the predicted high-frequency content. Note that the feature maps have the same spatial size as Y_{HR} and D_{bic}. This allows for an easy auto-merging process to fuse the feature maps.

2.3. Auto-merging and reconstruction

Instead of designing a switching network that selects from two sets of feature maps (DEN and CBPN), we argue that it is better to allow the network to learn an automatic merging scheme. As shown in Fig.2, we concatenate the feature maps generated by DEN and CBPN into a new set of 128-feature maps. One convolutional layer is applied with filter size of 3×3 to merge the feature maps. The number of feature maps after this layer is reduced by half.

The last component in our network is to reconstruct the HR depth map from concatenated feature maps. One convolutional layer is employed to project feature maps onto the depth domain and generate the final depth map. D_{bic} is then added with this depth image to finalize the result. By adding D_{bic} and minimizing the loss function Eqn.(1), we explicitly force the network to learn the missing high-frequency components and suppress the unwanted noise.

2.4. Training

During the training process, we learn nonlinear mapping using *stochastic gradient decent* [15] with momentum set to 0.9. We randomly initialize the weights and train the model from scratch. The learning rate is set to be 0.01 and reduces to $1e^{-3}$ after 20 epochs. Gradient clipping is especially required for CBPN to avoid gradient explosion during back-propagation. Finally, to ensure low-sparsity constraint on the filters, we penalize all weights with an ℓ_2 penalty. Thus the total loss function becomes

$$L = \frac{1}{2n}\sum_{i=1}^{n}\|f(\Theta, D^i_{\mathrm{bic}}, Y^i_{\mathrm{HR}}) - D^i_{\mathrm{HR}}\|_2^2 + \sum_{t=1}^{T}\lambda\|w^t\|_2, \quad (2)$$

where T is the total number of filters and the regularization parameter λ is $1e^{-4}$.

	W/O Noise									W/ Noise								
	Art			Books			Moebius			Art			Books			Moebius		
	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$
Bicubic	6.62	14.80	30.45	1.02	2.43	5.06	0.82	1.90	4.16	30.18	38.52	54.19	24.69	26.43	28.79	24.38	25.71	27.80
Park[2]	8.03	12.24	17.42	1.43	2.24	3.91	1.13	1.82	3.25	14.13	20.83	35.10	3.79	6.80	10.97	3.82	6.29	10.34
Ferstl[3]	9.19	14.33	22.92	1.66	2.57	3.97	1.28	2.13	3.66	10.16	16.50	25.84	2.38	4.90	**6.08**	2.17	4.12	**6.65**
Diebel[1]	9.73	13.40	30.28	1.45	2.39	4.88	1.41	2.07	4.22	12.17	20.37	40.84	4.26	9.01	16.43	4.52	9.64	17.43
Yang[14]	16.53	16.45	22.21	2.60	2.90	3.80	1.14	1.92	3.31	9.03	16.17	24.89	3.51	5.68	8.30	3.67	5.85	8.87
DEN	**0.57**	**3.68**	16.92	**0.20**	**0.75**	**1.83**	**0.15**	**0.66**	1.82	4.13	10.22	28.45	**1.38**	3.30	7.15	1.56	3.63	7.73
DEN + CBPN	0.58	4.15	**12.94**	0.22	1.12	1.88	0.15	0.68	**1.65**	**4.13**	**9.44**	**23.31**	1.42	**3.26**	7.05	**1.43**	**3.25**	6.71

Table 1. Quantitative evaluation. We compare DEN and DEN + CBPN with several state-of-the-art methods. The mean-square-error (**MSE**) results on Middlebury data sets are compared. The best result is bolded and the second best is underlined.

3. EXPERIMENT

We evaluate the performance of the proposed network on widely-used Middlebury Stereo data sets [16] [17] [18]. Thirty-five subjects are obtained from the Middlebury 2001 - 2006 data sets and 32 of them are used as the training set. There are two depth maps provided for each subject along with their corresponding color images. We extract sub-images with size 44×44 from these 64 depth maps. Data augmentation (e.g., flip and rotate) is used to expand our training set. To test the performance of our method, trained nonlinear mapping is applied to three test images: *Art*, *Books*, and *Moebius*.

3.1. Benchmark Comparison

First, we present a quantitative evaluation of the proposed method. In Table.1, nonlinear mapping learned by DEN and DEN+CBPN are compared with several state-of-the-art methods [2] [1] [3] [14]. SR factors of $\times 2$, $\times 4$, and $\times 8$ are considered, and mean-square-error (MSE) is adopted as the performance metric. We set up two experiments to demonstrate the ability of our proposed method. Noise-free LR depth maps are generated by down-sampling the HR depth maps with bicubic interpolation, and noisy LR depth maps are created by adding Gaussian noise after down-sampling.

To demonstrate the strength of deep neural networks, we train the nonlinear mapping solely with DEN. The SR depth maps generated by DEN are compared with aforementioned depth SR methods. Note that all compared methods are color-guided and that ours is the only learning-based approach. We can easily observe that DEN outperforms previous methods by a large margin when the LR depth maps are noise-free. This implies that the nonlinear mapping learned by DEN is capable of restoring most of the missing high-frequency components accurately. In the scenario that LR depth maps are corrupted by noise, DEN still outperforms [2] [1] [14] when SR factors are $\times 2$ and $\times 4$, and achieves the state-of-the-art performance when SR factor is $\times 8$.

3.2. How much can color help?

Second, we evaluate the performance of color-guided neural network, *i.e.*, DEN + CBPN. We mainly compare the performance between DEN and DEN + CBPN to illustrate the benefits of including color images in depth map SR[1]. We note that when training DEN + CBPN jointly, parameters learned in DEN are not completely equivalent to those learned in solo DEN. This is mainly due to the fact that two networks can cooperate with each other to achieve optimal accuracy.

For smaller SR factors, $\times 2$ and $\times 4$, including color image contributes effects benefit when LR depth maps are noise-free. This is consistent with our analysis on Fig.1. When the interference from noise is absent, color images are relatively unreliable. As a result, including color images at these scenarios only"confuses" the network and reduces the learning accuracy of the network. For larger SR factors, $\times 8$, depth edges are significantly distorted. Thus, depth edges predicted by color images are more reliable, leading to a superior performance. Similar observations can be made when LR depth maps are corrupted with noise. When the depth edges become unreliable, our network tends to rely on CBPN for restoring more accurate depth edges. Therefore, contribution of color image increases when the reliability of the LR depth map decreases (e.g., as noise gets stronger).

4. CONCLUSION

In this paper, we have taken a learning-based approach for color-guided depth map super-resolution. We adopt the popular deep CNN to learn non-linear mapping between LR and HR depth maps. Furthermore, a novel color-based prediction network is proposed to properly exploit supplementary color information in addition to the depth enhancement network. In our experiments, we have shown that deep neural network based approach is superior to several existing state-of-the-art methods. Further comparisons are reported to confirm our analysis that the contributions of color image vary significantly depending on the reliability of LR depth maps.

[1]Please refer to our supplemental material for visual comparison. *https://anvoy.github.io/publication.html*

5. REFERENCES

[1] James Diebel and Sebastian Thrun, "An application of markov random fields to range sensing," in *NIPS*, 2005, vol. 5, pp. 291–298.

[2] J. Park, H. Kim, Y-W Tai, M.S. Brown, and I. Kweon, "High quality depth map upsampling for 3d-tof cameras," in *International Conference on Computer Vision*. IEEE, 2011, pp. 1623–1630.

[3] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *International Conference on Computer Vision*. IEEE, 2013, pp. 993–1000.

[4] W. Zhou, X. Li, and D. Reynolds, "Image assisted upsampling of depth map via nonlocal similarity," in *48th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2014, pp. 683–687.

[5] J. Yang, J. Wright, T.S. Huang, and Yi Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[6] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian Conference on Computer Vision*, 2014, pp. 111–126.

[7] Kwang In Kim and Younghee Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010.

[8] Chen Dong, C.C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*, 2014, pp. 184–199.

[9] J. Kim, J.K. Lee, and K.M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.

[10] J. Kim, J.K. Lee, and K.M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Computer Vision and Pattern Recognition*. IEEE, June 2016.

[11] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

[12] W.T. Freeman, T.R. Jones, and E.C. Pasztor, "Example-based super-resolution," *IEEE Computer graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

[13] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *IEEE International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[14] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[15] Y.A. LeCun, L. Bottou, G.B. Orr, and K-R Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*, pp. 546–546. 1998.

[16] 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.

[17] Daniel Scharstein and Richard Szeliski, "High-accuracy stereo depth maps using structured light," in *Computer Vision and Pattern Recognition, 2003*. IEEE, 2003, vol. 1, pp. I–195.

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