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, color-guided 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-
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.

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 $D_{HR} \in \mathbb{R}^{M \times N}$ from its LR counterpart $D_{LR} \in \mathbb{R}^{M \times N}$. 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 $D_{bic} \in \mathbb{R}^{sM \times sN}$. 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 $\Theta = \{W, B\}$ such that the following loss function

$$L(\Theta) = \frac{1}{2n} \sum_{i=1}^{n} \|f(\Theta, D_{bic}^i, Y_{HR}^i) - D_{HR}^i\|_2^2,$$  

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

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 $D_{HR} \in \mathbb{R}^{M \times N}$ from its LR counterpart $D_{LR} \in \mathbb{R}^{M \times N}$. 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 $D_{bic} \in \mathbb{R}^{sM \times sN}$. 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 $\Theta = \{W, B\}$ such that the following loss function

$$L(\Theta) = \frac{1}{2n} \sum_{i=1}^{n} \|f(\Theta, D_{bic}^i, Y_{HR}^i) - D_{HR}^i\|_2^2,$$  

is minimized, where $f(\cdot)$ estimate the HR depth map based on $W, B$, 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
convolutional layers can be used to boost performance, at the price of increased computational complexity.

For each convolutional layer, we use an array of $3 \times 3$ 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 $21 \times 21$ 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- interpolation forces the network to learn the missing high-frequency components and suppress the unwanted noise.

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 $3 \times 3$ 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 $\ell_2$ penalty. Thus the total loss function becomes

$$ L = \frac{1}{2n} \sum_{i=1}^{n} \|f(\Theta, D_{bic}, Y_{HR}) - D_{bic}\|^2 + \sum_{t=1}^{T} \lambda \|u^t\|_2, \quad (2) $$

where $T$ is the total number of filters and the regularization parameter $\lambda = 1e^{-4}$.
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 $44 \times 44$ 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. 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 inference 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. 

Please refer to our supplemental material for visual comparison. https://anvoy.github.io/publication.html
5. REFERENCES


