Abstract—RGB-D cameras have been widely used in various applications, such as gesturing or exercise games, consumer healthcare systems, and 3D measurements. However, the resolution of depth maps is lower than that of an RGB image, which significantly limits the potential applications of depth maps. Recently, deep learning-based super-resolution techniques have achieved state-of-the-art results for image resolution enhancement, in which paired high- and low-resolution images are used for training. The challenge with the depth map super resolution for RGB-D cameras is that we do not have high-resolution depth maps for training models (deep networks). We propose a deep self-learning approach for color-guided depth map super resolution. We achieve super resolution using only low-resolution depth maps to train a network, which is comparable to the ideal case (training the network using paired high- and low-resolution images). The high-resolution image is used as a guide to improve the resolution enhancement. Experimental results demonstrate that the proposed method outperforms almost as same as accuracy of trained by external datasets even not using external datasets.

Index Term—super resolution, depth map, color guidance, deep learning, convolutional neural network, self-learning

I. INTRODUCTION

RGB-D camera devices, e.g., Microsoft Kinect [1] have been widely used in various applications, such as gesturing or exercise games, consumer healthcare systems, and 3D measurements. However, the resolution of depth maps is lower than that of color maps; thus, it must be increased up to color map resolution. High-definition depth map generation is essential for precise 3D reconstruction and automatic mapping systems. For this purpose, many depth map super-resolution (SR) methods [2]-[4] have been proposed. For example, DepthSR-Net [4], which is a convolutional neural network (CNN)-based method that uses color maps as guides, provides high-definition depth maps and is the state-of-the-art (SOTA) depth map SR method. In deep learning-based SR methods, paired high- and low-resolution images are used for training. The challenge with the depth map super resolution for RGB-D cameras is that we do not have high-resolution depth images for training the model (i.e., a deep network). Although we can use high-resolution (HR) images generated from an external dataset, such as 3D stereo images, for training, it is difficult to realize a high SR accuracy because of different device parameters or settings, the shooting environment, and the added noise or blur.

An SR method that uses a self-learning (SL) approach [5][6] has been proposed previously and has archived good results in a blind SR task and SR of real-world images. The Generic SR method involves training networks or dictionaries using external training data. The SL SR method involves training ones using internal data which is the image improving the resolution itself to improve the resolution. It is especially effective for images with unique artifacts, noise, or fractal structures.

In this paper, we propose a CNN-based depth map SR method that uses color maps as guides and a deep SL approach to generate optimum networks for each depth map. In addition, to generate high-definition depth maps, we have designed a simple and effective network architecture that adopts residual learning [7], a dense connection [8], and batch normalization layers. As a result, we can construct a network that is optimized for each depth map and generates high-definition depth maps. Fig.1 shows that the proposed method generates smoother, clearer depth maps in comparison to upscaling with bicubic interpolation.

II. RELATED WORK

A. Zero-shot super-resolution (ZSSR)

Zero-shot super-resolution (ZSSR) [6] is a CNN-based SR method that uses an SL approach. It was developed to improve the resolution of low-resolution (LR) images that contain unique noise or artifacts. The ZSSR method is based on a simple CNN that uses only convolution layers. The network is trained by the LR image and its downscaled image as a paired HR- and LR images. After training, the LR image is the fed into the trained network as an input LR image to estimated its high-resolution images. Though this method is called as zero-shot SR, it can be considered as unsupervised SR or internal SL SR.
**B. DepthSR-Net**

DepthSR-Net [4] is the SOTA SR method for depth maps. This method uses a corresponding HR color map for depth map SR. The network is based on U-Net [9] and involves the concatenation of the input depth map to the output feature map of each downscaling layer in an encoder as well as the HR color map to the output feature map of each upscaling layer in a decoder. The network is trained using stereo RGB-D datasets with paired HR and LR depth maps and improves depth map resolution using transfer learning.

**III. PROPOSED METHOD**

**A. Overview**

The flow of the proposed method is shown in Fig.2, which is based on ZSSR. Unlike ZSSR, we do not only have the LR depth map, we also have the corresponding HR color image. We are going to use the HR color image as a guided image to improve the ZSSR accuracy. We used the LR depth map and its downsampled depth image as a paired HR- and LR depth maps to train the network. The downsampled color image is used as a HR guided color image for training. Furthermore, the amount of training data is increased using data augmentation. The CNN, which is described in detail the next section, is trained using the training data, and the original LR depth map resolution is increased using the trained CNN.

**B. Network architecture**

The network architecture (orange blocks in Fig.2) of the proposed method is shown in Fig.3. This network is inspired by DepthSR-Net, which is based on the U-Net architecture. We remove the encoder in DepthSR-Net and add the residual block (ResBlock) from SRResNet [10], which facilitates stable training using residual learning and batch normalization layers to extract useful feature maps from a guide color map. Furthermore, based on the ResBlock, we propose a GuideBlock that includes extracted guide information efficiently to depth map SR. We stack three GuideBlocks and apply the dense connection structure of the output feature maps of each Guideblock. The proposed method generates training data from only a depth map and a color map using data augmentation techniques; thus, the training dataset is small. Thus, to facilitate stable training, we include the residual skip connections at the beginning and the end of the network, as well as before the first GuideBlock and after the final GuideBlock. In addition, batch normalization layers are implemented after all convolution layers (except the final convolution layer).

**IV. EXPERIMENT**

**A. Experimental Settings**

We used 37 paired depth and color maps from the 2001 [11], 2003 [12], 2005 [13][14] and 2006 [13][14] datasets from the Middlebury Stereo Datasets. Five paired images (Cones, Laundry, Moebius, Teddy, and Tsukuba) were used for testing, and the remaining 32 data were used as an external dataset for training.
training without the SL approach to facilitate comparison. In the training phase, the training data size was 64, the batch size was 32, ADAM [5] was employed as the optimization method, and L1 was implemented as the loss function. The initial learning rate was 0.001, which was multiplied by 0.1 when the minimum value of loss was not updated over 10 epochs. The number of epochs was 1000; however, once the learning rate was below 0.000001, training was completed. For data augmentation, we used downscaling (from 0.1 to 1 in 0.1 increments), rotation (from 0° to 360° in 15° increments), and flipping. In the preprocessing, input depth and color map were normalized such that the mean and standard deviation to be 0 in the preprocessing. And in the postprocessing, the result depth map is reduced reconstruction errors by iterative back projection method and the rounding in maximum and minimum value of the one before processing, after improving the resolution. We experimented with upscaling factors of 2x and 4x. For the 4x upscaling factor, the depth map was improved via 2x upscaling factor twice. For the second time, the first trained weight of the network was set as the initial value, and the network was retrained using a 2x upscaled depth map as new label data. We compared four pattern methods, i.e., bicubic interpolation, a

![Ground truth (PSNR / SSIM)](image)

![Guide image (-> / -)](image)

![Bicubic interpolation (26.04 / 0.86)](image)

![SL with guided (26.98 / 0.88)](image)

![SL without guided (26.45 / 0.86)](image)

![External HR- and LR-dataset (28.77 / 0.93)](image)

![SL with guided (25.14 / 0.94)](image)

![SL without guided (25.07 / 0.94)](image)

![External HR- and LR-dataset (25.17 / 0.95)](image)

Fig.4. 4x results of Tsukuba and Teddy: (a)-(f) ground truth depth map, paired color map, bicubic interpolation, deep SL approach with guided, deep SL approach without guided, and (d) trained by external dataset, respectively.
deep SL approach with color guidance, a deep SL approach without color guidance, and SR with color guided trained from an external dataset (ideal result). We evaluated the proposed method relative to peak signal-to-noise ratio (PSNR) and structural similarity (SSIM).

B. Results

The results of 2x and 4x upscaling are shown in Tables 1 and Table 2, respectively. The left value of each cell is the PSNR, and the right value is the SSIM. The best results among three methods with LR image only for training are highlighted with bold face.

<table>
<thead>
<tr>
<th>Table I</th>
<th>Comparison of reconstruction accuracy (PSNR / SSIM) for the case of upscaling factor: 2x. The best results among three methods with LR image only for training are highlighted with bold face.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bicubic interpolation</td>
</tr>
<tr>
<td>Cones</td>
<td>30.58 / 0.96</td>
</tr>
<tr>
<td>Laundry</td>
<td>31.97 / 0.97</td>
</tr>
<tr>
<td>Tsukuba</td>
<td>32.35 / 0.94</td>
</tr>
<tr>
<td>Teddy</td>
<td>30.09 / 0.96</td>
</tr>
<tr>
<td>Moebius</td>
<td>30.86 / 0.96</td>
</tr>
<tr>
<td>Average</td>
<td>23.63 / 0.92</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table II</th>
<th>Comparison of reconstruction accuracy (PSNR / SSIM) for the case of upscaling factor: 4x. The best results among three methods with LR image only for training are highlighted with bold face.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bicubic interpolation</td>
</tr>
<tr>
<td>Cones</td>
<td>26.58 / 0.91</td>
</tr>
<tr>
<td>Laundry</td>
<td>29.20 / 0.95</td>
</tr>
<tr>
<td>Tsukuba</td>
<td>26.04 / 0.86</td>
</tr>
<tr>
<td>Teddy</td>
<td>24.74 / 0.92</td>
</tr>
<tr>
<td>Moebius</td>
<td>27.55 / 0.94</td>
</tr>
<tr>
<td>Average</td>
<td>26.88 / 0.91</td>
</tr>
</tbody>
</table>

We have proposed a CNN-based depth map super-resolution method that uses color guidance and a deep SL approach and is usual for RGB-D devices which high resolution depth maps are not exist. The experimental results demonstrate that the proposed method is more effective for depth map super-resolution than upsampling with bicubic interpolation and a method that does not employ color guidance. In future, we plan to further evaluate the proposed method using real depth maps obtained from RGB-D cameras. In addition, we plan to improve the method’s processing speed while sustaining accuracy such that it can be applied to tasks that demand quick processing and high accuracy (relative to small details), such as gesture recognition. For this purpose, we will explore more optimal training parameters and network architectures.

REFERENCE