Channel Attention based Iterative Residual Learning for Depth Map Super-Resolution

Xibin Song[1,2], Yuchao Dai[3]*, Dingfu Zhou[1,2*], Liu Liu[5,6], Wei Li[4], Hongdong Li[5,6] and Ruigang Yang[1,2,7]

[1]Baidu Research [2]National Engineering Laboratory of Deep Learning Technology and Application, China [3] Northwestern Polytechnical University, China [4] Shandong University, China [5] Australian National University, Australia [6]Australian Centre for Robotic Vision, Australia [7]University of Kentucky, Kentucky, USA

{songxibin,zhoudingfu}@baidu.com, daiyuchao@gmail.com

Abstract

Despite the remarkable progresses made in deep-learning based depth map super-resolution (DSR), how to tackle real-world degradation in low-resolution (LR) depth maps remains a major challenge. Existing DSR model is generally trained and tested on synthetic dataset, which is very different from what would get from a real depth sensor. In this paper, we argue that DSR models trained under this setting are restrictive and not effective in dealing with real-world DSR tasks. We make two contributions in tackling real-world degradation of different depth sensors. First, we propose to classify the generation of LR depth maps into two types: non-linear downsampling with noise and interval downsampling, for which DSR models are learned correspondingly. Second, we propose a new framework for real-world DSR, which consists of four modules : 1) An iterative residual learning module with deep supervision to learn effective high-frequency components of depth maps in a coarse-to-fine manner; 2) A channel attention strategy to enhance channels with abundant high-frequency components; 3) A multi-stage fusion module to effectively re-exploit the results in the coarse-to-fine process; and 4) A depth refinement module to improve the depth map by TGV regularization and input loss. Extensive experiments on benchmarking datasets demonstrate the superiority of our method over current state-of-the-art DSR methods.

1. Introduction

Depth maps have been widely embraced as a new technology by providing complementary information in many applications [12][23][41][42][43][44]. However, depth sensors, such as Microsoft Kinect and Lidar, can only provide depth maps of limited resolutions. Hence, depth map super-resolution (DSR) draws more and more attentions. As a fundamental low-level vision problem, DSR aims at super-resolving a high-resolution (HR) depth map from a low-resolution (LR) depth map input [9][18][33][39][40][46], which is a challenging task due to the great information loss in the down-sampling process. Besides, depth maps generally contain less textures and more sharp boundaries, and are usually degraded by noise due to the imprecise consumer depth cameras, which further increase the challenge.

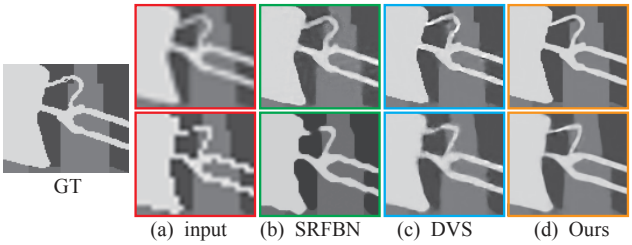


Figure 1. Results of different methods using different types of LR depth maps as input (×4). (a) input, (b) SRFBN [22], (c) DVS [40] and (d) Ours. The first row shows the results under non-linear (bi-cubic) down-sampling degradation, while the second row shows the results under interval down-sampling degradation.

Recently, significant progress has been made in super-resolution by using convolutional neural networks (CNNs) in regression ways, both in color image super-resolution (CSR) and DSR [6][9][16][18][22][24][46]. These methods usually apply bi-cubic downsampling as the degradation model and add noise to simulate the generation of LR images. Besides, [10] and [49] propose to estimate the down-sampling kernels to estimate the degradation of LR

*Corresponding author

images. However, bi-cubic degradation model and degradation kernels are insufficient to describe the process of depth map down-sampling.

Depth map exists in different types in real world, which can be classified into two types: (1). depth maps with smoothed surfaces, such as depth maps generated by stereo matching [1][28][31] and depth maps captured by low-cost sensors (Kinect); (2). depth maps with sharp boundaries, such as depth maps captured by Lidar. For (1), depth maps are always smooth, thus, non-linear downsampling degradation model and down-sampling kernels can be used to simulate the generation of LR depth maps. For (2), depth maps captured by Lidar are generated from 3D points of real world. They are always with sharp boundaries. Imaging the projection process of 3D points onto a 2D image, when two 3D points are projected to a same 2D coordinates in a depth map, it should reserve the 3D point with smaller depth z due to occlusion. Interpolation (bi-cubic or degradation kernel) is not suitable in such process, hence we argue that bi-cubic degradation and blur kernels are not reasonable, and we propose to use interval down-sampling degradation to describe the down-sampling progress. Fig. 1 (a) illustrates the two types of LR depth maps, where interval down-sampling and non-linear degradation have quite different manifestations.

In this paper, to effectively tackle the two types of depth maps (non-linear degradation with noise and interval down-sampling degradation), we adopt an iterative residual learning framework with deep supervision (coarse-to-fine), which guarantees that each sub-module can gradually obtain the high-frequency components of depth maps step by step. Besides, in each sub-module, channel attention strategy is utilized to enhance the channels with more effective information, thus, obtains better results. What's more, the inter-media results obtained by different sub-modules are fused to provide effective information to tackle different types of depth maps. Total Generalized Variation (TGV) term and input loss are utilized to further refine the obtained HR depth maps. Any support of HR color information is not needed and weight sharing between different sub-modules can effective reduce the number of parameters, which makes our proposed approach much more flexible. The proposed framework is trained in an end-to-end manner, and experiments on various benchmarking datasets demonstrate the superiority of our method over state-of-the-art super-resolution methods, including both DSR and CSR methods.

Our main contributions are summarized as:

- To tackle real world degradation in low-resolution depth maps, we propose to classify the generation of LR depth maps into two types: non-linear down-sampling with noise and interval downsampling, for which DSR models are learned correspondingly.
- We propose an iterative residual learning based framework for real world DSR, where channel attention, multi-stage fusion, weight sharing and depth refinement are employed to learn HR depth maps in a coarse-to-fine manner.
- Extensive experiments on various benchmarking datasets demonstrate the superiority of our proposed framework over current state-of-the-art DSR methods.

2. Related work

In this section, we briefly review related work in both color image super-resolution (CSR) and depth map super-resolution (DSR).

2.1. DCNN based CSR

In CSR, bi-cubic down-sampling degradation are commonly used down-sampling methods to generate LR color images. Methods, such as [6][20][45][34], have proven that CNN outperformed conventional learning approaches with large margin. These methods regard super-resolution as an LR color image to HR color image regression problem, and generate an end-to-end mapping between LR and HR image. Besides, residual architectures, such as [3][16][32][51] are commonly used in solving CSR. HR color images are generated by learning the residuals between LR images and groundtruth. Recently, back projection strategy, such as [5][13][22][24], are proved to have well performance by representing the LR and HR feature residuals in more efficient ways. Meanwhile, attention based model is also utilized in CSR. To obtain more discriminative representations, [4][50] propose to use attention strategy to enhance feature representation. Kernel based methods [10] [49] [52] are also utilized in CSR, which estimate a blur kernel to simulate the generation of LR color images. Besides, [26] exploits pixel to pixel transfer techonolgy in solving the problem of CSR.

2.2. Depth Map Super-resolution

2.2.1 Conventional Learning based DSR

To solve the problem of DSR, prior information is used as useful guidance to generate HR depth maps from LR depth maps. Using prior information learned from additional depth map datasets, [15][27][48] propose to use MRF method to solve the problem of DSR. Meanwhile, other learning based methods, such as sparse representation and dictionary learning, are utilized in DSR. [8] proposes to exploit sparse coding strategy and Total Generalized Variation (TGV) to effectively generate HR depth edges, which are used as useful guidance in the generation of HR depth maps. Besides, using HR color image as effective guidance, [7] utilizes an anisotropic diffusion tensor to solve the problem of DSR. What's more, a bimodal co-sparse analysis model generated from color images are utilized in [19]

to generate an HR depth map from an LR depth map. Additionally, [30] proposes to compute local tangent planes using HR color images in the process of DSR, since it can provide auxiliary information. Besides, the consistency information between color images and depth maps is used to generate HR depth maps in [25].

2.2.2 DCNN based DSR

The success of DCNN in high-level computer vision tasks has been extended to DSR. Using SRCNN [6] as the mapping unit, [39] proposes an effective DCNN based learning method to generate a HR depth map from an LR depth map. Meanwhile, [33] proposed a ATGV-Net which combines DCNN with total variations to generate HR depth maps. The total variations are expressed by layers with fixed parameters. Besides, a novel DCNN based method is proposed in [9], which combines a DCNN with a non-local variational method. Note that corresponding HR color images and up-sampled LR depth maps are regarded as input to feed into the network in [9][33][39]. What's more, a multi-scale fusion strategy is utilized in [18], which uses a multi-scale guided convolutional network for DSR with and without the guidance of the color images. Besides, [40] proposes a novel framework by using view synthesis to explain the generation of LR depth maps. [46] used rendering of 3D surfaces to measure the quality of obtained depth maps. It demonstrates that a simple visual appearance based loss yields significantly improved 3D shapes.

However, most of conventional learning based DSR and DCNN based DSR exploit bi-cubic degradation to generate LR depth maps, which are not enough to describe the generation of LR depth maps in real world.

3. Our approach

3.1. Overview

To tackle with different types of LR depth maps in DSR, including non-linear degradation with noise and interval down-sampling degradation, we adopt an iterative residual learning framework. As shown in Fig. 2, it contains several sub-modules, and the input of each sub-module is the output the previous sub-module, which guarantees that the residual between the input and groundtruth can be learned step by step. Besides, as the network goes deeper, strong supervision from groundtruth is added in each sub-module to release gradient vanishing. In the last sub-module, high-frequency components obtained by previous sub-modules are fused as input, which re-exploits cause-to-fine high-level information to further improve the performance of the proposed framework. In each sub-module, residual learning strategy is used, and it contains two indispensable blocks: feature extraction block and channel attention based reconstruction block. Besides, except the loss between output and groundtruth, if it is well recovered, the down-sampled version of obtained depth maps should be same with the input $\mathbf{D}^L$, hence, we use such input loss to constrain the framework. What's more, to maintain sharp boundaries, total generalized variation (TGV) term is utilized to further refine the obtained HR depth maps.

3.2. Network structure

As shown in Fig. 2, our network can be unfolded to K sub-modules. We utilize a residual connection for sub-modules and calculate loss between each sub-module's output and groundtruth to alleviate gradient vanishing. The loss function is defined in Sec. 3.5. Each sub-module contains two parts: feature extraction block (FE) and channel attention based reconstruction block (CAR).

For the k-th ($k \in [1, K]$) sub-module, its input and output is defined as $\mathbf{I}^k$ and $\mathbf{O}^k$, respectively. The operation of learning high-frequency component $\mathbf{O}^k_{\text{CAR}}$ is given by:

$$\begin{aligned} \mathbf{O}^k_{\text{FE}} &= \mathcal{F}_{\text{FE}}\left(\mathbf{I}^k\right) \\ \mathbf{O}^k_{\text{CAR}} &= \mathcal{F}_{\text{CAR}}\left(\mathbf{O}^k_{\text{FE}}\right), \end{aligned} \tag{1}$$

where $\mathcal{F}_{\text{FE}}(\cdot)$ and $\mathcal{F}_{\text{CAR}}(\cdot)$ are feature extraction and channel attention based reconstruction operation, respectively.

The output of k-th sub-module $\mathbf{O}^k$ is given by:

$$\mathbf{O}^k = \mathbf{O}^k_{\text{CAR}} + \mathbf{I}^k. \tag{2}$$

Combing Eq. (1) and (2), the operation of k-th sub-module can be summarized as:

$$\mathbf{O}^k = \mathcal{S}_k\left(\mathbf{I}^k\right), \tag{3}$$

where $\mathcal{S}_k(\cdot)$ denotes the operation of k-th sub-module. The output of k-th sub-module is taken as the input of the next sub-module, *i.e.* $\mathbf{O}^k = \mathbf{I}^{k+1}$.

For the last sub-module K, the input is the concatenation of $\mathbf{O}^1$ to $\mathbf{O}^{K-1}$, dubbed $\mathcal{F}_{\text{concat}}\left(\mathbf{O}^1, \cdots, \mathbf{O}^{K-1}\right)$, where $\mathcal{F}_{\text{concat}}(\cdot)$ is the concatenation operation, and the operation of last sub-module K is given by:

$$\mathbf{O}^k = \mathcal{S}_k\left(\mathcal{F}_{\text{concat}}\left(\mathbf{O}^1, \cdots, \mathbf{O}^{K-1}\right)\right). \tag{4}$$

For the first sub-module, the input is the up-sampled version of LR depth maps $\mathbf{D}^L$ ($\uparrow \lambda$ times, where λ is the up-sampling factor). We use bi-cubic up-sample kernel for simplicity.

3.3. Feature extraction block

A convolutional layer is dubbed as Conv (m, n), where m is the kernel size and n is the number of kernels. In feature extraction block, it contains l convolutional layers with ReLU as activation function. We set $m = 3$, $n = 64$ and $l = 8$ in this paper.

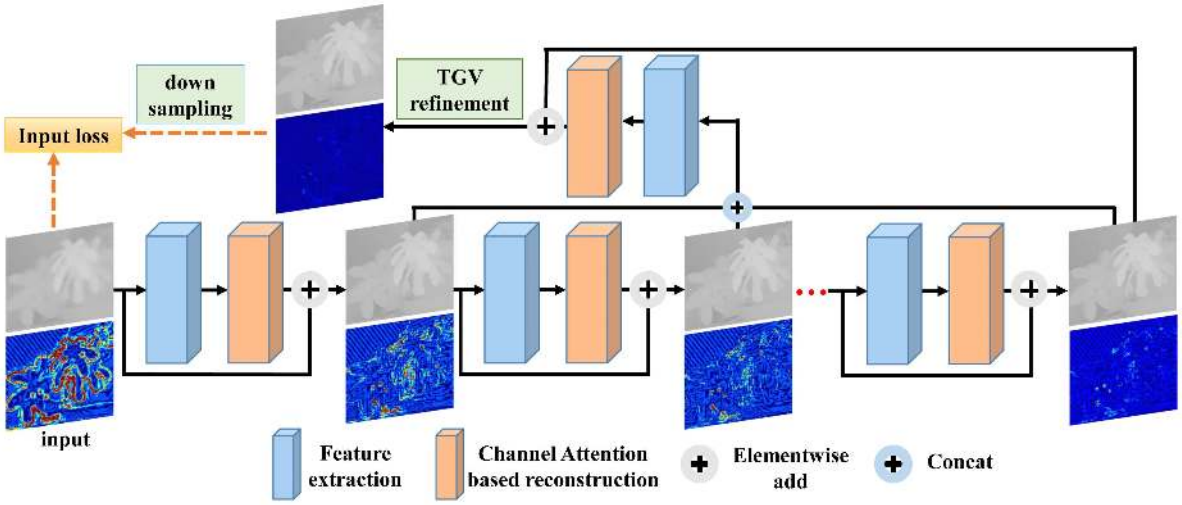


Figure 2. The figure shows the pipeline of the proposed framework. We show the residual of the output between each sub-module and groundtruth. From blue to red means value from 0 to ∞.

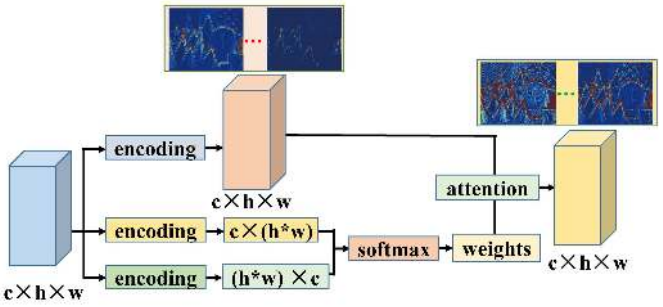


Figure 3. The figure shows the pipeline of channel attention.

3.4. Channel attention based reconstruction

Inspired by [24], we proposed to use attention strategy in DSR, and the proposed channel attention based reconstruction block provides imperative information to learn high-frequency components of depth maps. It contains two steps: channel attention and reconstruction.

Channel attention: Each sub-module $\mathcal{F}_{\text{CAR}}(\cdot)$ takes the output of feature extraction block $\mathbf{O}_{\text{FE}}^k$ as input. For $\mathbf{O}_{\text{FE}}^k$ with tensor size of $(c \times h \times w)$, $\mathcal{F}_{\text{CAR}}(\cdot)$ first converts $\mathbf{O}_{\text{FE}}^k$ to three components $\mathcal{P}\left(\mathbf{O}_{\text{FE}}^k\right)$, $\mathcal{Q}\left(\mathbf{O}_{\text{FE}}^k\right)$ and $\mathcal{V}\left(\mathbf{O}_{\text{FE}}^k\right)$ via encoding operations $\mathcal{P}(\cdot)$, $\mathcal{Q}(\cdot)$ and $\mathcal{V}(\cdot)$, respectively. The tensor size of $\mathcal{P}\left(\mathbf{O}_{\text{FE}}^k\right)$, $\mathcal{Q}\left(\mathbf{O}_{\text{FE}}^k\right)$ and $\mathcal{V}\left(\mathbf{O}_{\text{FE}}^k\right)$ is $(c \times h \times w)$, $(c \times hw)$ and $(hw \times c)$, respectively.

$\mathcal{P}(\cdot)$ is data pre-processing and it contains α convolutional layer. $\mathcal{Q}(\cdot)$ and $\mathcal{V}(\cdot)$ are convolution with reshape operations. The number of convolutional layers are β and γ, respectively. $\mathcal{Q}(\cdot)$ and $\mathcal{V}(\cdot)$ are defined for learning channel attention parameters. $\mathcal{Q}\left(\mathbf{O}_{\text{FE}}^k\right)$ and $\mathcal{V}\left(\mathbf{O}_{\text{FE}}^k\right)$ are dot-producted (elementwise multiplication), and fed to a softmax operation to regress channel attention weights $\boldsymbol{\theta}$. $\mathcal{P}\left(\mathbf{O}_{\text{FE}}^k\right)$ and $\boldsymbol{\theta}$ are dot-producted to obtain the output of channel attention $\mathbf{O}_{\text{CAR}}^k$. Fig. 3 shows the pipeline of the proposed channel attention. The above operations can be defined as following:

$$\begin{aligned} \boldsymbol{\theta} &= \text{softmax}\left(\mathcal{Q}\left(\mathbf{O}_{\text{FE}}^k\right) \odot \mathcal{V}\left(\mathbf{O}_{\text{FE}}^k\right)^{\mathsf{T}}\right), \\ \mathbf{O}_{\text{CAR}}^k &= \boldsymbol{\theta} \odot \mathcal{P}\left(\mathbf{O}_{\text{FE}}^k\right). \end{aligned} \tag{5}$$

The channel attention can be understood as non-local convolution process, which aims to enhance the channels with much more effective information. The non-local operation in the proposed channel attention based reconstruction can obtain effective attention weights for each channel by exploiting all the position information of the feature maps. $\mathcal{Q}\left(\mathbf{O}_{\text{FE}}^k\right) \odot \mathcal{V}\left(\mathbf{O}_{\text{FE}}^k\right)^{\mathsf{T}}$ can be regarded as a form of covariance of the input data. It provides an effective score to describe the tendency of two feature maps at different channels.

Reconstruction: Based on $\mathbf{O}_{\text{CAR}}^k$, we can obtain its reconstruction result $\mathbf{O}^k$ by using η convolutional layers. In this paper, we set $\alpha = \beta = \gamma = \eta = 1$, and use Conv $(3, 64)$ in channel attention stage and Conv $(3, 1)$ in reconstruction stage. The effectiveness of the proposed channel attention based reconstruction block will be demonstrated in the experiment section.

3.5. Loss Function

We use the L_1 loss to optimize the proposed framework.

3.5.1 Sub-module loss

For the k-th sub-module, the loss is defined as:

$$L_k = ||\mathbf{O}^k - \mathbf{D}^G||_1, \tag{6}$$

where L_k is the loss between the output of k-th sub-module and the groundtruth.

Our framework can obtain K HR depth maps with LR depth maps as input. Generally, one will pay more attention on the output of the last sub-module, hence different weights are set for losses at different sub-modules, and the loss weight increases as the network goes deeper. The final loss for sub-modules is defined as following:

$$L_s = \sum_{k=1}^{K} \frac{k}{N} L_k, \tag{7}$$

where $N = \sum_{k=1}^{K} k = K(K+1)/2$.

3.5.2 Input loss and TGV term

The HR depth map is well recovered, the down-sampled version (same degradation model) of the finally obtained depth maps should be the same as the original LR input $\mathbf{D}^L$. Hence, we use the input loss to further constrain the obtained HR depth map, which is defined as:

$$L_{\text{input}} = ||\mathcal{F}_{\text{down}}\left(\mathbf{O}^K\right) - \mathbf{D}^L||_1, \tag{8}$$

where $\mathcal{F}_{\text{down}}(\cdot)$ is the degradation model, with output tensor of the same size as $\mathbf{D}^L$.

Besides, depth maps usually contain sharp boundaries, hence, the total generalized variation $\mathsf{TGV}\left(\mathbf{O}^K\right)$ is exploited to refine the final obtained HR depth maps.

The final loss of our proposed framework is defined as:

$$L = L_s + \xi_1 L_{\text{input}} + \xi_2 \mathsf{TGV}(\mathbf{O}^K), \tag{9}$$

where ξ_1 and ξ_2 are weights for input loss and total variation term. We set $\xi_1 = 0.1$ and $\xi_2 = 0.05$ in this paper.

3.6. Implementation

We employed Adam [21] as the optimizer to optimize the parameters, and the learning rate varies from 0.1 to 0.0001 by multiplying 0.1 for every 25 epochs. Adjustable gradient clipping strategy [20] is used. The proposed framework converged after 100 epochs.

4. Experiment

In this section, we evaluate the performance our method against different state-of-the-art (SOTA) methods on diverse publicly available datasets using different types of LR depth maps as input, including non-linear degradation with noise and interval down-sampling degradation.

4.1. Datasets

We used 6 different datasets in this paper: (1). Middlebury dataset [14][35][36][37], which provides high-quality depth maps for complex real-world scenes; (2). The Laserscan dataset [27], which is captured by laser sensors and provides accurate depth measurements; (3). Sintel dataset [2], ICL dataset [11] and synthetic New Tsukuba dataset [29], which are synthesized datasets and contain lots of depth details and high quality depth maps; (4). SUN RGBD dataset [38], which contains images captured with different consumer-level RGBD cameras, such as Microsoft Kinect, and provides low-quality depth maps for complex real-world scenes; and (5). Apolloscape dataset [17][47], which contains high-quality depth maps captured by Lidar in real traffic scenes.

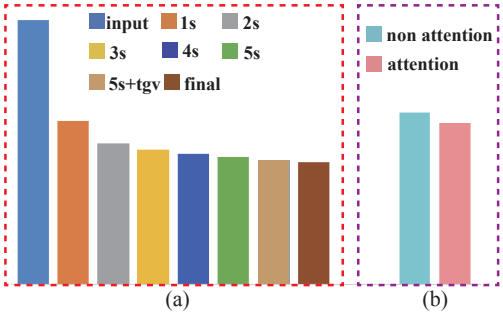


Figure 4. (a) shows the results of average $RMSE$ of the proposed framework with different number of sub-modules. $1s$ to $5s$ are number of sub-modules from 1 to 5 respectively. $5s + tgv$ means 5 sub-modules with TGV refinement and $final$ means 5 sub-modules with input loss and TGV refinement. (b) shows the average $RMSE$ results of attention and non-attention respectively.

Training dataset: To effectively training the proposed framework, we follow DVS [40] to prepare our training dataset. 115 depth maps are collected from the Middlebury dataset [14][36][37], the Sintel dataset [2] and the synthetic New Tsukuba dataset [29]. Using these depth maps, the input LR depth maps $\mathbf{D}^L$ are obtained by $\mathbf{D}^L = \downarrow_\lambda \mathbf{D}^G$, where λ is the down-sampling factor. To simulate the generation of real LR depth maps, non-linear degradation with noise and interval down-sampling degradation are used. Besides, bi-cubic downsampling is commonly used in DSR [39][40][48], hence, we use bi-cubic degradation to evaluate the performance of non-linear degradation.

Evaluation: To effectively evaluate the performance of our proposed framework, depth maps ($Motorcycle$, $Playtable$, $Flowers$ and $Jadeplant$) from Middlebury 2014 dataset [35] are chosen as the testing depth maps. Besides, to further evaluate the generalization performance of the proposed framework, we also evaluate depth maps chosen from ICL dataset [11] ($Plant$ and $Room$), Laser-Scan dataset [27] ($ls21$, $ls30$ and $ls42$), SUN RGBD dataset [38] (0100 and 0400) and Apolloscape dataset [17][47]($road01$, $road05$, $road10$ and $road17$). Note that models are trained with depth maps from Middlebury dataset, sintel dataset and synthetic New Tsukuba dataset.

Baseline Methods: Our propose method is compared with the following three categories of methods: (1). Standard interpolation approaches: Bi-cubic and Nearest Neighbour ($Nearest$); (2). State-of-the-art CNN based DSR approaches: EG [48], MS-Net [18], DVS *et*

$\times 4$	$Plant$	$Room$	0100	0400	$Motorcycle$	$Playtable$	$Flowers$	$Jadeplant$	$ls21$	$ls30$	$ls42$
Bi-cubic	1.2340	1.5448	3.0922	1.3039	4.9046	2.9967	4.6655	4.1660	2.8441	2.6544	5.4735
Nearest	1.4102	1.7558	3.4003	1.5271	5.6645	3.4443	5.4189	4.8238	3.3306	3.1039	6.3581
ABPN [24]	1.1588	1.2605	2.8357	1.1167	4.6597	2.7904	4.4472	4.0635	2.5961	2.5063	5.1318
SRFBN [22]	1.1039	1.3029	2.8254	1.0808	4.3934	2.5663	4.0677	3.6864	2.4516	2.2565	4.9645
SAN [4]	1.2297	1.3813	2.9248	1.1987	4.4938	2.6558	4.2388	3.9288	2.5027	2.3519	5.0777
IKC [10]	1.2048	1.3240	2.9011	1.1324	4.4215	2.6078	4.1846	3.8026	2.4865	2.3028	4.9981
EG [48]	1.4253	1.7250	3.3987	1.5038	5.4685	3.3261	5.2067	4.6162	3.2764	6.0576	6.4288
MS-Net [18]	1.1952	1.5116	3.1302	3.6576	5.0119	2.9683	4.7982	4.2426	2.7356	2.6127	5.8623
DVS [40]	0.8494	1.1682	2.8914	0.9601	3.2553	2.0168	3.0409	2.9407	1.8188	1.8079	3.2001
$AIR_{ws}(ours)$	0.7300	1.0952	2.8028	0.7531	3.1025	1.9024	2.9520	2.8004	1.6252	1.5930	2.9528
$AIR(ours)$	**0.7278**	**1.0639**	**2.7800**	**0.7611**	**3.0968**	**1.8626**	**2.8873**	**2.7740**	**1.6048**	**1.5668**	**2.9332**

Table 1. Comparison of $RMSE$ results under up-sampling factor of $\times 4$ (interval down-sampling degradation). AIR_{ws} means results obtained by weights sharing among different sub-modules and AIR means non weights sharing. The best result is highlighted and the second best is underlined.

$\times 4$	$Plant$	$Room$	0100	0400	$Motorcycle$	$Playtable$	$Flowers$	$Jadeplant$	$ls21$	$ls30$	$ls42$
Bi-cubic	0.7300	0.9613	2.2028	0.8189	3.0434	1.8108	2.9092	2.6154	1.6278	1.5801	3.3351
Nearest	0.8841	1.1528	2.3911	1.0324	3.7067	2.2046	3.5688	3.1795	2.1406	2.0200	4.2166
ABPN [24]	0.6561	0.8476	1.5899	0.8005	2.0048	1.2335	1.7998	1.7204	1.3236	1.1814	1.7379
SRFBN [22]	0.7068	0.8727	1.4063	0.7885	1.8300	1.1699	1.6949	1.6797	1.2107	1.1710	1.6175
SAN [4]	0.6651	0.7238	1.7139	0.7588	2.0501	1.4477	1.9034	1.8425	1.4803	1.3692	1.8232
Meta-SR [16]	0.6467	0.8026	1.5319	0.8005	1.9938	1.2517	1.7581	1.7065	1.3016	1.1645	1.7158
IKC [10]	0.6815	0.7523	1.4652	0.7630	2.0812	1.3420	1.8351	1.8092	1.3521	1.2021	1.7956
EG [48]	0.7740	0.9972	2.1337	0.8874	2.9183	1.6414	2.6186	2.5365	1.7593	1.6318	3.3086
MS-Net [18]	0.4675	0.6453	1.0524	0.5760	2.0554	1.3518	1.9564	1.9218	1.4324	1.4087	1.7569
DVS [40]	0.4565	0.5903	0.9826	0.5387	1.9718	1.2588	1.8532	1.8458	1.3800	1.3424	1.7212
$AIR_{ws}(ours)$	0.4101	**0.5196**	0.9692	0.4996	1.7923	1.1655	1.7324	1.6781	1.1456	1.0788	**1.4875**
$AIR(ours)$	**0.4004**	0.5351	**0.9588**	**0.4986**	**1.7764**	**1.1622**	**1.7005**	**1.6765**	**1.1393**	**1.0633**	1.4877

Table 2. Comparison of the $RMSE$ results under up-sampling factor of $\times 4$ (bi-cubic degradation with noise). AIR_{ws} means weights sharing among different sub-modules and AIR means non weights sharing. The best result is highlighted and the second best is underlined.

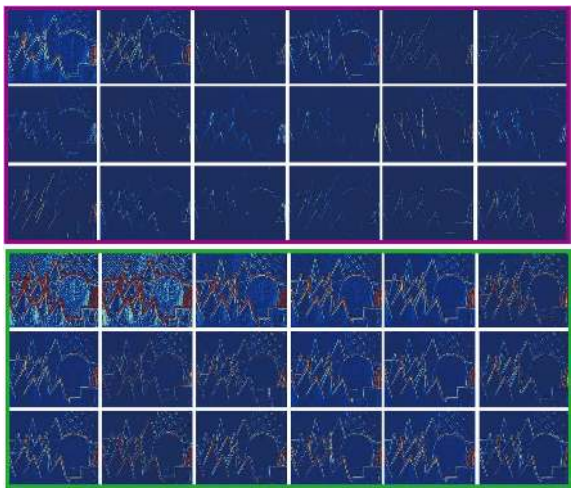


Figure 5. The feature maps of high-frequency component before and after channel attention. Purple and green areas shows the feature maps before and after channel attention block respectively. From blue to red means value from to 0 to ∞. Best viewed on screen.

al. [40]; (3). State-of-the-art CNN based color image super-resolution approaches: SRFBN [22], Meta-SR [16], SAN [4], ABPN [24] and IKC [10]. Besides, all the methods are retrained with the same depth maps.

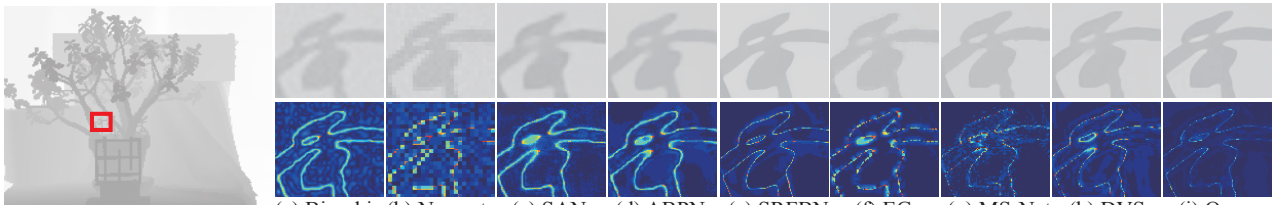
Error metrics: Root Mean Squared Error ($RMSE$) is used to evaluate the performance obtained by our method and other state-of-the-art methods. Specifically, $RMSE = \sqrt{\sum_{i=1}^{N}(O_i - D_i^G)^2/N}$, where O and D^G are the obtained HR depth map and ground truth respectively, N is the number of pixels in the HR depth map.

$\times 4$	$road01$	$road05$	$road10$	$road17$
Bi-cubic	18.5311	33.6010	20.5177	19.1795
Nearest	21.2863	38.6045	23.4327	22.1853
ABPN [24]	16.9054	28.4027	18.2029	17.3051
SRFBN [22]	15.9080	27.9377	17.1010	16.0810
SAN [4]	16.1057	29.2059	18.5678	17.2564
IKC [10]	16.0557	28.3142	17.5034	16.4750
EG [48]	27.7714	49.5420	31.0133	34.4618
MS-Net [18]	19.0029	31.8750	21.1448	20.1059
DVS [40]	16.0110	26.4482	17.0613	16.0500
$AIR_{ws}(ours)$	15.6305	25.9282	16.6152	15.8423
$AIR(ours)$	**15.6239**	**25.9109**	**16.5906**	**15.7792**

Table 3. Comparison of $RMSE$ results under up-sampling factor of $\times 4$ (interval down-sampling degradation). AIR_{ws} means weights sharing among different sub-modules and AIR means non weights sharing. The best result is highlighted and the second best is underlined.

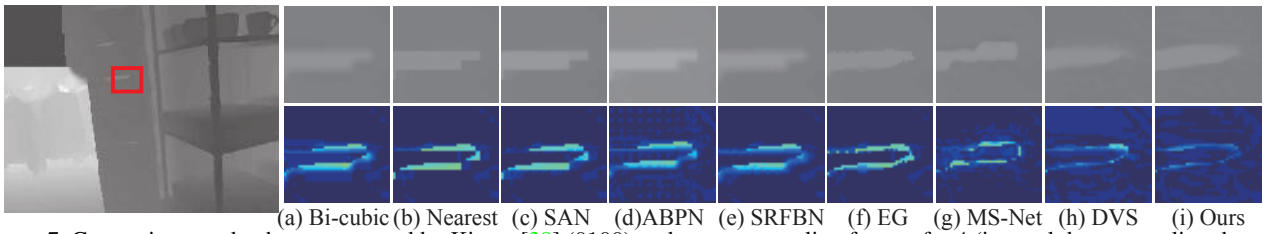
4.2. Ablation analysis

Fig. 4 (a) demonstrates the average $RMSE$ results of the proposed method with different number of sub-modules on Middlebury dataset ($Cones$, $Teddy$, $Tsukuba$ and $Venus$) under up-sampling factor of $\times 4$ (bi-cubic degradation). It can be observed that the $RMSE$ loss drops with the number of sub-module increases, which proves that deeper network with more sub-modules can obtain the residual component effectively, from which high-frequency component



(a) Bi-cubic (b) Nearest (c) SAN (d) ABPN (e) SRFBN (f) EG (g) MS-Net (h) DVS (i) Ours

Figure 6. Comparison on Middlebury 2014 dataset [35] ($Jadeplant$) under up-upsampling factor of $\times 4$ (bi-cubic degradation with noise). (a) Bi-cubic, (b) Nearest Neighbor, (c) SAN [4], (d) ABPN [24], (e) SRFBN [22], (f) EG [48], (g) MS-Net [18], (h) DVS [40] and (i) Our results. The second row shows the residual between the results and groundtruth. From blue to red means 0 to ∞. Best viewed on screen.



(a) Bi-cubic (b) Nearest (c) SAN (d)ABPN (e) SRFBN (f) EG (g) MS-Net (h) DVS (i) Ours

Figure 7. Comparison on depth map captured by Kinect [38] (0100) under up-upsampling factor of $\times 4$ (interval down-sampling degradation) on depth map captured by Kinect. (a) Bi-cubic, (b) Nearest Neighbor, (c) SAN [4], (d) ABPN [24], (e) SRFBN [22], (f) EG [48], (g) MS-Net [18], (h) DVS [40] and (i) Our results. The second row shows the residual between the results and groundtruth. From blue to red means 0 to ∞. Best viewed on screen.

can be recovered step by step. Generally, any number of sub-module can be used in the proposed framework, and as shown in Fig. 4 (a), as the number of sub-module increasing, the whole framework becomes convergent. Hence, we set the number of sub-module $K = 5$ in this paper. What's more, HR depth maps obtained by input loss and TGV refinement get smaller $RMSE$, which demonstrates that these operations can recover more useful high-frequency information and further refine the obtained HR depth maps. Therefore, we can conclude that the all the components utilized in our framework contribute positively toward the final success of our approach.

4.3. Attention analysis

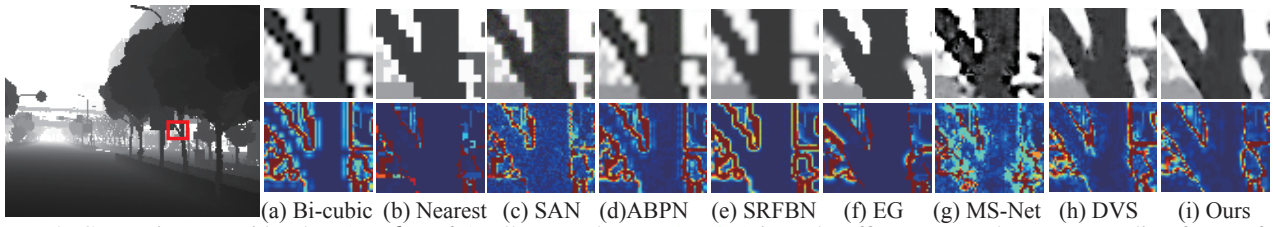
Fig. 5 shows the feature maps obtained before and after channel attention strategy. The top 18 feature maps with high-frequency component are shown in Fig. 5, purple and green areas demonstrate the feature maps before and after channel attention, respectively. And from blue to red means value from 0 to ∞. According to Fig. 5, we can see that effective high-frequency component, such as edges, are efficiently enhanced by channel attention, which can be utilized to reconstruct better HR depth maps. Fig. 4 (b) shows the average $RMSE$ results of the proposed method with and with channel attention strategy. Middlebury dataset ($cones$, $teddy$, $tsukuba$ and $venus$) under up-sampling factor of $\times 4$ are used as input. It is obviously to find that smaller $RMSE$ can be obtained using channel attention strategy, which proves that the proposed channel attention strategy works positively in super-resolving DSR problem.

According to Fig. 5 and Fig. 4 (b), we can conclude that channel attention can enhance the channels with useful high-frequency information and improve the ability of each sub-module to obtain the residual, thus, recover high quality depth maps effectively.

4.4. Interval degradation

We first evaluate the performance of the proposed approach on depth maps with interval down-sampling degradation. The quantitative results in terms of $RMSE$ of up-sampling factors of $\times 4$ are reported in Table. 1 and Table. 3. As indicated in Table. 1 and Table. 3, AIR_{ws} and AIR demonstrate the results of the proposed method with and without weights-sharing among different sub-modules, respectively. It can be observed that the performances of state-of-the-art on interval down-sampled LR depth maps are not good enough (both CSR and DSR methods), and the proposed method outperforms other DCNN based methods with smaller $RMSE$. Besides, Table. 3 shows the results on dense depth maps captured by Lidar on real traffic scenes, which proves that the proposed framework can tackle with real LR Lidar data effectively. Besides, we can see that results of weights-sharing outperforms other state-of-the-art methods, and results of non-weight-sharing obtain better $RMSE$ results because it contains more parameters, thus have stronger non-linear mapping abilities to recover better HR depth maps.

Qualitative results are illustrated in Fig. 7 (0100 extracted from SUN RGBD dataset [38]) and Fig. 8 ($road01$ from Apolloscape dataset [17][47]) for an up-sampling factor $\times 4$ under interval down-sampling degradation. As shown in Fig. 7 and Fig. 8, 0100 and $road01$ are depth maps captured by Kinect and Lidar, which represent the depth maps captured in indoor and outdoor scenes of real



(a) Bi-cubic (b) Nearest (c) SAN (d)ABPN (e) SRFBN (f) EG (g) MS-Net (h) DVS (i) Ours

Figure 8. Comparison on Lidar data ($road01$ of Apolloscape dataset [17][47]) in real traffic scenes under up-upsampling factor of $\times 4$ (interval down-sampling degradation). (a) Bi-cubic, (b) Nearest Neighbor, (c) SAN [4], (d) ABPN [24], (e) SRFBN [22], (f) EG [48], (g) MS-Net [18], (h) DVS [40] and (i) Our results. The second row shows the residual between the results and groundtruth. From blue to red means 0 to ∞. Best viewed on screen.

world. Obviously, the proposed method produces more visually appealing results with smaller residual compared with groundtruth. Boundaries generated by the proposed method are sharper and more accurate, which demonstrate that the structure and high-frequency component of high-resolution depth maps can be well recovered.

4.5. Bi-cubic degradation

In this section, we evaluate the proposed framework on noisy depth maps. Following [33][40], depth dependent Gaussian noise is added to LR depth maps $\mathbf{D}^{LR}$ in the form $\theta(d) = \mathcal{N}(0, \delta/d)$, where $\delta = 651$ and d denotes the depth value of each pixel in $\mathbf{D}^{L}$. Besides, to evaluate the ability of noise handling, we also add noise on depth maps captured by Kinect (0100 and 0400 from SUN RGBD dataset [38]).

Table 2 reports the quantitative results in terms of $RMSE$ for the up-sampling factor of $\times 4$ with bi-cubic degradation and noise as input, from which, we can clearly see that the proposed method outperforms others, even on raw depth maps with additional added noise (0100 and 0400). The proposed method can well eliminate the influence of noise, thus depth maps with smaller $RMSE$ can be obtained.

Fig. 6 illustrates the qualitative results of the proposed method ($Jadeplant$ from Middlebury 2014 dataset [35]) under up-sampling factor $\times 4$ with bi-cubic degradation and noise as input. As shown in Fig. 6, $Jadeplant$ contains complex textures and luxuriant details, which is hard to recover a HR depth map from a LR depth map. Obviously, the proposed method produces more visually appealing results with sharper and more accurate boundaries, which proves that the proposed method can effectively recover the structure of HR depth maps.

4.6. Generalization ability

As discussed in section 4.4 and section 4.5, depth maps of ICL dataset [11], SUN RGBD dataset [38], Laserscan dataset [27] and Apolloscape dataset [17][47] are not included in the training data. Based on Table. 1, Table. 2, Table. 3, Fig. 7 and Fig. 8, we can find that the proposed approach outperforms other methods on all testing depth maps with smaller $RMSE$ results under non-linear (bi-cubic) degradation with noise and interval down-sampling degradation, which demonstrates the excellent generalization ability of the proposed framework on both synthesis and raw depth maps.

4.7. Weight sharing

As reported in Table. 1, Table. 2 and Table. 3, the proposed framework with weight-sharing among different sub-modules outperforms state-of-the-art methods, while it gets similar results with non-weight-sharing strategy. The last sub-module combines the outputs of previous sub-modules as input, hence, we use weight-sharing in other sub-modules except the last one. And the parameters of weight-sharing are only 40% of parameters of non-weight-sharing ($K = 5$), which makes the proposed framework lightweight and more flexible in comparison with other state-of-the-art methods.

5. Conclusions

In this paper, we have proposed an effective depth map super-resolution method that accounts for real-world degradation processes of different types of physical depth sensors. We have envisaged the employment of our new method to super-resolve depth maps captured by commodity depth sensors such as Microsoft Kinect and Lidar. We analyze two different LR depth map simulation schemes: non-linear downsampling and interval downsampling. Furthermore, we have devised a channel attention based iterative residual learning framework to address real world depth map super-resolution. Extensive experiments across different benchmarks have demonstrated the superiority of our proposed approach over the state-of-the-art.

Acknowledgment. The work is supported by Baidu Research. Yuchao Dai's research is supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102803 and Natural Science Foundation of China grants (61871325, 61420106007, 61671387), and Hongdong Li's research is supported in part by the ARC Centre of Excellence for Robotics Vision (CE140100016) AND ARC-Discovery (DP 190102261), ARC-LIEF (190100080) grants. The authors of ANU gratefully acknowledge the GPUs donated by NVIDIA Corporation. We thank all anonymous reviewers and ACs for their constructive comments.

References

[1] Stephen T Barnard and Martin A Fischler. Computational stereo. Technical report, SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER, 1982.

[2] 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*, pages 611–625. Springer, 2012.

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3086–3095, 2019.

[4] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019.

[5] Xin Deng, Ren Yang, Mai Xu, and Pier Luigi Dragotti. Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3076–3085, 2019.

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. 2014.

[7] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 993–1000, 2013.

[8] David Ferstl, Matthias Ruther, and Horst Bischof. Variational depth superresolution using example-based edge representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 513–521, 2015.

[9] Matthias Rüther Gernot Riegler, David Ferstl and Horst Bischof. A deep primal-dual network for guided depth super-resolution. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 7.1–7.14. BMVA Press, September 2016.

[10] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1604–1613, 2019.

[11] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *IEEE international conference on Robotics and automation*, pages 1524–1531, 2014.

[12] Richard Hartley and Hongdong Li. An efficient hidden variable approach to minimal-case camera motion estimation. *IEEE transactions on pattern analysis and machine intelligence*, 34(12):2303–2314, 2012.

[13] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1732–1741, 2019.

[14] Heiko Hirschmuller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[15] Michael Hornacek, Christoph Rhemann, Margrit Gelautz, and Carsten Rother. Depth super resolution by rigid body self-similarity in 3d. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1123–1130, 2013.

[16] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: a magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1575–1584, 2019.

[17] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 954–960, 2018.

[18] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *European conference on computer vision*, pages 353–369. Springer, 2016.

[19] Martin Kiechle, Simon Hawe, and Martin Kleinsteuber. A joint intensity and depth co-sparse analysis model for depth map super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1545–1552, 2013.

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[22] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, 2019.

[23] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2372–2381, 2017.

[24] Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li, and Wan-Chi Siu. Image super-resolution via attention based back projection networks. In *IEEE International Conference on Computer Vision Workshop(ICCVW)*, October 2019.

[25] Jiajun Lu and David Forsyth. Sparse depth super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2245–2253, 2015.

[26] Riccardo de Lutio, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. Guided super-resolution as pixel-to-pixel

transformation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8829–8837, 2019.

[27] 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*, pages 71–84. 2012.

[28] David Marr and Tomaso Poggio. Cooperative computation of stereo disparity. *Science*, 194(4262):283–287, 1976.

[29] Sarah Martull, Martin Peris, and Kazuhiro Fukui. Realistic cg stereo image dataset with ground truth disparity maps. In *ICPR workshop TrakMark2012*, volume 111, pages 117–118, 2012.

[30] Kiyoshi Matsuo and Yoshimitsu Aoki. Depth image enhancement using local tangent plane approximations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3574–3583, 2015.

[31] Masatoshi Okutomi and Takeo Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (4):353–363, 1993.

[32] Yajun Qiu, Ruxin Wang, Dapeng Tao, and Jun Cheng. Embedded block residual network: A recursive restoration model for single-image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4180–4189, 2019.

[33] Gernot Riegler, Matthias Rüther, and Horst Bischof. Atgv-net: Accurate depth super-resolution. In *European conference on computer vision*, pages 268–284. Springer, 2016.

[34] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[35] 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*, pages 31–42. Springer, 2014.

[36] Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[37] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.

[38] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[39] Xibin Song, Yuchao Dai, and Xueying Qin. Deep depth super-resolution: Learning depth super-resolution using deep convolutional neural network. In *Asian conference on computer vision*, pages 360–376. Springer, 2016.

[40] Xibin Song, Yuchao Dai, and Xueying Qin. Deeply supervised depth map super-resolution as novel view synthesis. *IEEE Transactions on circuits and systems for video technology*, 29(8):2323–2336, 2018.

[41] Xibin Song, Haiyang Huang, Fan Zhong, Xin Ma, and Xueying Qin. Edge-guided depth map enhancement. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2758–2763. IEEE, 2016.

[42] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang Yang. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5452–5462, 2019.

[43] Xibin Song, Jianmin Zheng, Fan Zhong, and Xueying Qin. Modeling deviations of rgb-d cameras for accurate depth map and color image registration. *Multimedia Tools and Applications*, 77(12):14951–14977, 2018.

[44] Xibin Song, Fan Zhong, Yanke Wang, and Xueying Qin. Estimation of kinect depth confidence through self-training. *The Visual Computer*, 30(6-8):855–865, 2014.

[45] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[46] Oleg Voynov, Alexey Artemov, Vage Egiazarian, Alexander Notchenko, Gleb Bobrovskikh, Evgeny Burnaev, and Denis Zorin. Perceptual deep depth super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5653–5663, 2019.

[47] Peng Wang, Xinyu Huang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[48] Jun Xie, Rogerio Schmidt Feris, and Ming-Ting Sun. Edge-guided single depth image super resolution. *IEEE Transactions on Image Processing*, 25(1):428–438, 2015.

[49] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Deep plug-and-play super-resolution for arbitrary blur kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1671–1681, 2019.

[50] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[51] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018.

[52] Ruofan Zhou and Sabine Susstrunk. Kernel modeling super-resolution on real low-resolution images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2433–2443, 2019.