RGB GUIDED DEPTH MAP SUPER-RESOLUTION WITH COUPLED U-NET

Yingjie Cui[a], *Qingmin Liao*[*a], *Wenming Yang*[a], *Jing-Hao Xue*[b]

[a]Dept. of E.E./Shenzhen International Graduate School, Tsinghua University, China
cuiyj19@mails.tsinghua.edu.cn; liaoqm@tsinghua.edu.cn; yang.wenming@sz.tsinghua.edu.cn;
[b]Department of Statistical Science, University College London, UK jinghao.xue@ucl.ac.uk

ABSTRACT

The depth maps captured by RGB-D cameras usually are of low resolution, entailing recent efforts to develop depth super-resolution (DSR) methods. However, several problems remain in existing DSR methods. First, conventional DSR methods often suffer from unexpected artifacts. Secondly, high-resolution (HR) RGB features and low-resolution (LR) depth features are often fused in shallow layers only. Thirdly, only the last layer of features is used for reconstruction. To address the above problems, we propose Coupled U-Net (CU-Net), a new color image guided DSR method built on two U-Net branches for HR color images and LR depth maps, respectively. The CU-Net embeds a dual skip connection structure to leverage the feature interaction of the two branches, and a multi-scale fusion to fuse the deeper and multi-scale features of two branch decoders for more effective feature reconstruction. Moreover, a channel attention module is proposed to eliminate artifacts. Extensive experiments show that the proposed CU-Net outperforms state-of-the-art methods.

***Index Terms*—** Guided depth super-resolution, convolutional neural network, feature fusion, U-Net network, feature reconstruction

1. INTRODUCTION

Depth information plays a pivotal role in many applications, such as virtual reality, human-computer interaction and scene reconstruction. However, consumer-level depth cameras generally are of low resolution, which greatly limits their applications in the above fields. Fortunately, although the depth image resolution is 512×424 pixels, the corresponding RGB image is with a high resolution of 1920×1080 pixels for Kinect V2 camera. Therefore, recently many studies have been devoted to exploiting the high-resolution (HR) color image to guide the depth super-resolution (DSR), with the aim of reconstructing a super-resolution (SR) depth map from the input low-resolution (LR) depth map.

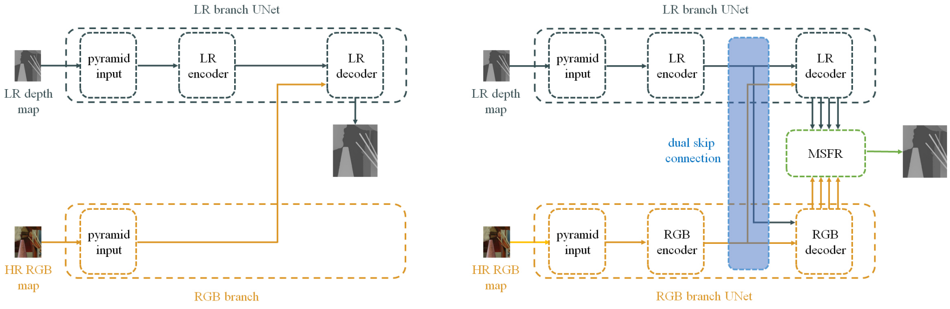
The existing DSR methods can be divided into three categories: regularization-based methods, filtering-based methods and learning-based methods.

In the regularization-based methods, the DSR is formulated as an optimization problem, the objective function of which is often composed of a loss term and a regularization term. The loss term is to ensure the structural consistency between the reconstructed SR features and the input LR features, while the regularization term is to constrain the ill-posed SR problem. Ferstl et al. [1] proposed to use total variation regularization (TGV) to avoid surface flattening and introduced a numerical algorithm based on a primal-dual formulation that can be efficiently parallelized. According to both the LR depth map and the nonlocal similarity in the corresponding high-quality color image, Yang et al. [2] used an adaptive color-guided autoregressive model as the regularization term. The regularization-based methods usually take more time to solve and cannot work well with images containing complex texture.

The filtering-based methods employ filters to replace the center point depth value with a locally weighted depth value. Tomasi et al. [3] proposed bilateral filters, whose weights depend on both the distance and the similarity between pixels. To improve the low-quality and low-illuminance image via a guide reference image, a joint bilateral filter was designed by Eisemann et al. [4]. Liu et al. [5] replaced the Euclidean distance with the geodesic distance to eliminate unexpected artifacts. The filtering-based methods only consider the local information of the image, hence producing only limited super-resolution performance.

The learning-based methods apply additional datasets to train the model for mapping LR depth maps into HR depth maps. These methods can be further divided into two types: DSR without and with color map guidance. The methods without color map guidance take a single image as input, similar to single image super-resolution (SISR) [6]. Riegler et al. [7] combined deep convolutional networks with variational methods to recover accurate HR depth maps. Song et al. [8] proposed a framework based on iterative residual learning to tackle real-world degradation. However, such methods cannot reconstruct accurately the HR edges because the single LR map loses too many edge details. In contrast, the methods with color map guidance can learn edge information easier with the help of the HR color image. Li et al. [9] got a suitable dictionary for generating HR images based on sparse

978-1-6654-3864-3/21/$31.00 ©2021 IEEE



(a) (b)

Fig. 1. **Overviews and comparison of the DepthSRNet and the proposed CU-Net.** (a) DepthSRNet is a single U-Net structure, where the RGB shallow features are directly fused with the LR encoder features. (b) CU-Net is a Coupled U-Net structure, where the LR depth map and the HR RGB image are respectively used as the inputs of two U-Nets. A mechanism of dual skip connection (DSC) is designed to realize the interaction of information between the two U-Nets, and the output of each level of the decoder is fed into a multi-scale feature reconstruction module (MSFR).

representation. Proposed by Hui et al. [10], MSG-Net is the first CNN structure designed for DSR which integrates multi-scale feature maps of the depth image and the color image to ensure the full integration of texture and structure. Then Guo et al. [11] proposed DepthSRNet based on MSG-Net and U-Net, which adopts a pyramid structure to input multi-scale LR depth images and color images into the encoder and the decoder respectively. To estimate structural details and regress filtering results, a deformable kernel network (DKN) was designed by Kim et al. [12].

The learning-based methods generally have better performance than those based on regularization or filtering. However, there are three issues with them. First, when the RGB image is inconsistent with the depth map texture, the DSR result will create unexpected *artifacts*. Secondly, HR RGB features and LR depth features are often fused in shallow layers *only*. Thirdly, *only* the last layer of features is employed for reconstruction, The last two issues result in insufficient exploitation of HR and LR features during the DSR.

To tackle these three issues, we propose Coupled U-Net (CU-Net), a new color image guided DSR method built on two branches of U-Nets to improve the state-of-the-art DepthSRNet. As shown in Fig.1, the CU-Net comprises three parts: an LR branch U-Net, an RGB branch U-Net, and a multi-scale feature reconstruction (MSFR) module. Specifically, our solutions to the three issues are summarized as follows.

First, to address the first issue, we design into the U-Nets a channel attention module, which considers the feature interdependence based on the residual channel attention block (RCAB) proposed in RCAN [13]. This module aims to reduce the channel weights that may cause artifacts. Secondly, to tackle the second issue, the proposed CU-Net fuses deeper feature maps level by level at the two decoders. Unlike the DepthSRNet, which directly employs the pyramid structure to fuse feature maps, the CU-Net can produce a more effective fusion of deeper features. Moreover, in order to let skip connections convey more valuable information, we introduce dual skip connection (DSC), as shown in the blue solid box of Fig.1(b), which divides each skip connection into two parts fed into the two different U-Nets. Thirdly, to attend to the third issue, we propose the MSFR module to integrate each level of the decoders for the final reconstruction.

In short, the main contributions of this work are:

- We propose Coupled U-Net (CU-Net), a new color image guided DSR method built on two branches of U-Nets, to improve the state-of-the-art DepthSRNet.
- We design DSC to ensure the feature interaction of the two branches, which can fully fuse LR depth feature maps and deep color feature maps during decoding. We propose the MSFR module to exploit deeper and multi-scale features to improve the final reconstruction. We leverage channel attention to eliminate artifacts created by the DepthSRNet.
- Extensive experiments show that the proposed CU-Net outperforms state-of-the-art methods.

2. THE PROPOSED METHOD: CU-NET

2.1. Overall structure of CU-Net

As shown in Fig.1(b), the proposed CU-Net consists of three parts: an LR branch U-Net (the black dashed frame), an RGB branch U-Net (the yellow dashed frame), and a multi-scale feature reconstruction (MSFR) module (the green dashed frame). In addition, there is a DSC structure (the blue frame) between the two branch U-Nets. The LR branch U-Net takes the LR depth map as input and the multi-scale deep features of the decoder as output. The RGB branch U-Net takes the Y channel of the YUV image as input and aims to get the deep features of a color image. The output features of each level

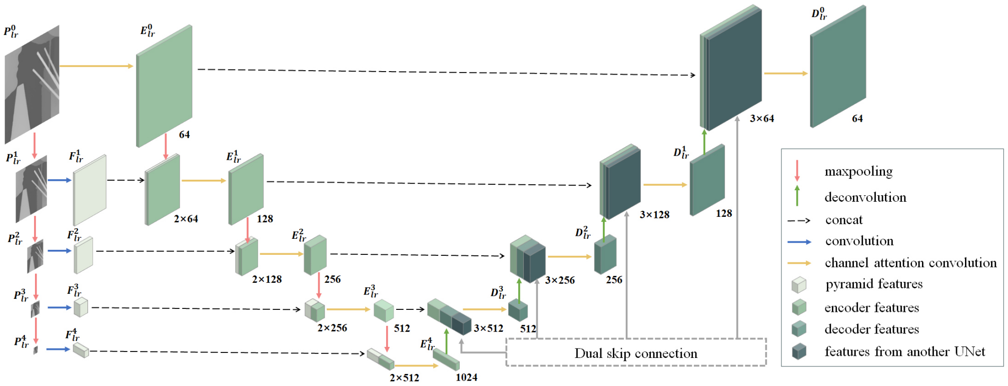


Fig. 2. **Architecture of the LR branch U-Net.** Different colored blocks indicate different features, and different colored arrows indicate different operations.

of the decoders are fed into the MSFR module for integrating the features and producing the reconstructed SR depth map.

2.2. Dual skip connection (DSC)

In the CU-Net, DSC executes the first fusion of LR depth features and HR color features and enables features to be transferred from encoders to decoders. The DSC can achieve two main purposes. First, it directly transfers the features from the encoder to the corresponding decoder in each branch, which prevents the loss of discriminative features during the downsampling process of the encoder. Secondly, the features of the LR depth map and the RGB image are integrated with each other, enriching the feature diversity of the two branches without introducing more parameters.

2.3. LR branch U-Net

Even though the LR branch U-Net and the RGB branch U-Net have different inputs, their structures are the same, so only the LR branch U-Net is introduced here, as shown in Fig.2.

Pyramid features F_{lr}^i. As with that in the DepthSRNet [11], a pyramid strategy is used to explore data input, which turns the LR depth map into multiple scales and attains multi-level receptive fields. Specifically, the maxpooling layer is applied N times to the original LR depth map (N is the number of times for U-Net downsampling; e.g., $N = 4$ in Fig.2), where each maxpooling layer reduces the length and width of the LR map to 1/2 of the original, and then the features of each scale are extracted by convolution layers:

$$P_{lr}^i = \text{maxpool}(P_{lr}^{i-1}), \tag{1}$$

$$F_{lr}^i = \sigma(W_{lr}^i \, P_{lr}^i + b_{lr}^i), \tag{2}$$

where P_{lr}^0 is the input interpolated LR map; P_{lr}^i, $i \in \{1, \ldots, N\}$, is the LR map after i times of maxpooling; F_{lr}^i is the shallow features extracted from P_{lr}^i; and W_{lr}^i and b_{lr}^i are the weight and the offset in the i^{th} convolution operation, respectively. Such a pyramid branch can make full use of input features of various scales, preventing the encoder from losing key texture information.

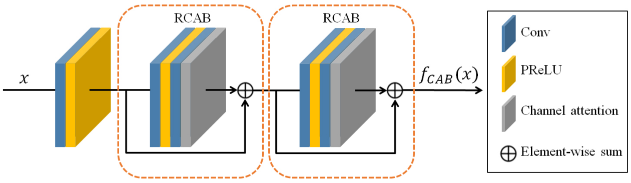


Fig. 3. **Channel attention module.** The channel attention module consists of a convolution layer and two residual channel attention blocks (RCABs). An RCAB contains a residual block and a channel attention block, which can extract more information between channels.

Encoder features E_{lr}^j. The encoder produces a set of hierarchical features for the decoder. The encoder features are obtained by applying a channel attention module to the concatenation of the pyramid features and the lower-resolution encoder features. As shown in Fig.3, a channel attention module is designed and embedded in the U-Net for exploiting the inter-dependencies between channels and eliminate artifacts. Let $f_{CAM}(\cdot)$ denote CAM operations. Then the encoder features can be obtained through

$$E_{lr}^0 = f_{CAM}(P_{lr}^0), \tag{3}$$

$$E_{lr}^j = f_{CAM}(C(F_{lr}^j, E_{lr}^{j-1})), \tag{4}$$

where $C(F_{lr}^j, E_{lr}^{j-1})$, $j \in \{1, \ldots, N\}$, represents the concatenation of features F_{lr}^j and E_{lr}^{j-1}.

Decoder features D_{lr}^k. With the help of DSC, the decoder can fuse the decoder features with the features from both the RGB encoder and the LR depth encoder:

$$D_{lr}^3 = f_{CAM}(C(E_{lr}^3, \text{deconv}\left(E_{lr}^4\right), E_{rgb}^3)), \tag{5}$$

$$D_{lr}^k = f_{CAM}(C(E_{lr}^k, \text{deconv}\left(D_{lr}^{k+1}\right), E_{rgb}^k)), \tag{6}$$

where $k \in \{N-2, \ldots, 0\}$; $\text{deconv}(\cdot)$ is the deconvolution operation with upsampling scale 2; and E_{rgb}^k denotes the encoder features from the RGB branch U-Net. It is worth noting that D^k_{lr} and E^j_{lr} have the same dimension when k is equal to j. Similarly, the decoder features in the RGB branch U-Net can be represented as D^m_{rgb}, where $m \in \{0, \ldots, N-1\}$.

2.4. Multi-scale feature reconstruction (MSFR) module

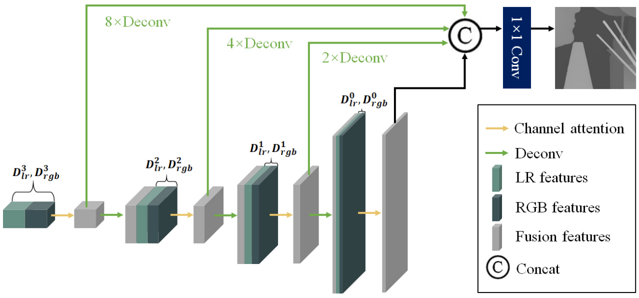


Fig. 4. **MSFR module.** The input of MSFR is the decoder features from two U-Nets, and the final SR image is obtained after multi-scale features fusion and reconstruction.

The LR decoder features (D^3_{lr} D^2_{lr} D^1_{lr} D^0_{lr}) and the RGB decoder features (D^3_{rgb} D^2_{rgb} D^1_{rgb} D^0_{rgb}) are the input of MSFR module. The MSFR fuses the various scales decoder features of two U-Nets for the final reconstruction. The structure of MSFR is shown in Fig.4, the fusion features can be expressed as

$$F^3_{fus} = f_{CAM}(C(D^3_{lr}, D^3_{RGB})), \quad (7)$$

$$F^n_{fus} = f_{CAM}(C(\text{deconv}(F^{n+1}_{fus}), D^n_{lr}, D^n_{RGB})), \quad (8)$$

where F^3_{fus} is obtained by fusing the features of two decoders; and F^n_{fus}, $n \in \{N-2, \ldots, 0\}$ for $N = 4$, is fused from the features of two decoders and the upper-level fusion features. When the upsampling scale is small (2×, 4×), we replace the channel attention with ordinary convolution.

The proposed MSFR has four deconvolution layers with different upsampling scales and one 1×1 convolution layer. The deconvolution layers are to up-sample the fusion features to make their sizes the same as the output image. In order to increase the scale diversity of features and consider more scale features to avoid the loss of low-scale information, we adopt 1×1 convolution for the final reconstruction.

3. EXPERIMENTS AND ANALYSIS

3.1. Implementation details

Network setting. The kernel size of all Conv layers is set to 3×3 (zero-padding) except in the channel attention module and during reconstruction. The scale level of the U-Net is set to $N = 4$. The 2× Deconv kernel size is set to 2×2 (with stride 2); the 4× Deconv kernel size is 4×4 (with stride 4); and the 8× Deconv kernel size is 8×8 (with stride 8).

Dataset. As with that for MSG-Net [10], the dataset used contains 92 RGB-D images obtained from the MPI Sintel depth dataset [14] and the Middlebury dataset [15–17], with 82 for training and 10 for validation. For fair comparison, no other datasets were used and no pre-training.

Training setting. For the color images, only the Y channel from the YUV format is retained; and the LR depth images are obtained by bicubic downsampling to the HR images. Before training, all inputs and labels are split into patches. When upscaling factors are [2, 4, 8, 16], the sizes of patches are [96, 96, 128, 128] and overlapping pixels are [48, 48, 64, 64]. After useless patches are excluded, the number of patches are [48,901, 48,901, 26,245, 26,245]. To make full use of the dataset, the training data are augmented with random horizontal (vertical) flips and 90°, 180°, 270° rotations, and the intensity of patches is normalized to the range [0, 1]. During training, the batch size is set to 64, and Adam [18] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) is chosen as the network optimizer. The initial learning rate is set to 10^{-4} and decays to 0.9 every 20 epochs. The mean squared error (MSE) is chosen as the loss function of the network. After 200 epochs of training, the loss of the validation set no longer decreases. In addition, all network parameters are initialized by the "Xavier" method [19]; all experiments are performed on PC with 2080Ti GPUs.

3.2. Ablation studies

Table 1. **The ablation study results for 4× SR**

Structure	M1	M2	M3	M4
CAM	×	√	√	√
MSFR	√	×	√	√
DSC	√	√	×	√
Average RMSE	0.75	0.67	0.66	0.65

We perform ablation studies on the effectiveness of three key CU-Net modules: channel attention modules (CAMs), multi-scale feature reconstruction module (MSFR) and Dual skip connection (DSC). The results are listed in Table 1.

In M1, the CAM is replaced with two 3×3 convolutions, which leads to 15% relative drop in performance from M4 (RMSE increased from 0.65 to 0.75). In M2, the MSFR module is removed and only the feature of the last layer of LR decoder (D^0_{lr}) is kept for reconstruction. Due to the lack of decoder features of various scales, the performance of M2 drops by 3% relatively from M4 (RMSE increased from 0.65 to 0.67). In M3, the feature interaction (DSC) is canceled between the LR branch and the RGB branch, which results in a relative performance drop of 1.5%. Therefore, it can be concluded that each CU-Net module is helpful to improve the network performance.

3.3. Comparison to state-of-the-art methods

We compare our CU-Net with the simple Bicubic method and the following state-of-art methods:

Table 2. **SR results (in RMSE) on Middlebury.** The lowest RMSE results among all methods are in bold.

method	Art				Book				Dolls				Laundry				Moebius				Reindeer			
	2×	4×	8×	16×	2×	4×	8×	16×	2×	4×	8×	16×	2×	4×	8×	16×	2×	4×	8×	16×	2×	4×	8×	16×
Bicubic	2.63	3.87	5.46	8.16	1.04	1.60	2.33	3.34	0.91	1.31	1.86	2.63	1.60	2.40	3.45	5.10	0.87	1.33	1.97	2.86	1.92	2.80	3.98	5.85
JBU	2.69	4.04	5.23	7.13	1.11	1.88	2.49	3.96	1.02	1.49	1.85	2.52	1.60	2.64	3.44	5.96	0.91	1.52	2.18	3.08	1.87	2.86	3.60	4.38
GF	3.47	4.76	6.78	9.95	1.43	1.97	2.77	4.15	1.22	1.64	2.26	3.45	2.08	2.88	4.08	6.29	1.23	1.73	2.46	4.16	2.50	3.45	4.81	7.18
TGV	3.17	3.71	7.02	11.55	1.32	1.65	2.08	3.80	1.17	1.42	2.05	4.44	1.84	2.20	3.92	6.75	1.14	1.45	2.41	5.41	2.41	2.67	4.29	8.80
MRFs	2.75	3.80	5.24	7.92	1.16	1.70	2.32	3.54	1.04	1.39	1.82	2.69	1.67	2.34	3.39	6.15	0.92	1.37	2.07	3.27	1.93	2.64	3.71	5.57
JID	1.25	2.01	3.23	5.74	0.65	0.92	1.27	1.93	0.70	0.92	1.26	1.74	0.75	1.21	2.08	3.62	0.64	0.89	1.27	2.13	0.92	1.56	2.58	4.64
MSG-Net	0.56	1.40	2.42	4.17	0.26	0.46	0.88	1.70	0.37	0.73	1.10	1.63	0.37	0.79	1.51	2.63	0.31	0.58	0.94	1.69	0.42	0.98	1.76	2.92
FDKN	1.53	2.10	3.16	9.46	0.50	0.73	1.21	3.93	0.70	0.93	1.30	3.21	0.88	1.26	2.00	5.95	0.60	0.79	1.24	4.26	1.09	1.50	2.27	6.53
RCAN	1.01	1.51	**2.20**	-	0.37	0.56	0.85	-	0.47	0.70	1.02	-	0.55	0.86	1.33	-	0.43	0.60	0.88	-	0.72	1.08	1.57	-
DepthSRNet	0.53	1.22	2.27	3.91	0.43	0.61	0.90	1.54	0.49	0.81	1.11	1.54	0.44	0.79	1.31	2.34	0.44	0.68	0.96	1.56	0.52	0.97	1.57	2.44
CU-Net	**0.27**	**1.05**	2.27	**3.67**	**0.16**	**0.35**	**0.73**	**1.45**	**0.22**	**0.61**	**0.97**	**1.43**	**0.19**	**0.59**	**1.15**	**2.25**	**0.20**	**0.48**	**0.77**	**1.31**	**0.24**	**0.82**	**1.51**	**2.38**

- Three regularization-based methods: total generalized variation (TGV) [1], MRFs [20], joint intensity and depth co-sparse coding (JID) [21].
- Two filtering-based methods: guided filter (GF) [22], joint bilateral upsampling (JBU) [23].
- Four learning-based methods: RCAN [13], MSG-Net [10], DepthSRNet [11], FDKN [12].

We evaluate our methods on the Middlebury dataset, just as other learning-based methods. Specifically, the noise-free dataset B in MSG-Net is used for the experiments, and the results for scaling factors of 2×, 4×, 8×, 16× are listed in Table 2. Experimental results show that the learning-based methods have better performance than both traditional regularization and filtering based methods. It can be also observed that the proposed CU-Net has achieved the best results in most images. Moreover, the CU-Net achieves the remarkable performance when the scaling factor is low or the texture of the test image is complicated (like the textures at panes and cubes in images Laundry and Moebius).

For intuitive visual comparison, the SR results of different methods for 8× SR are shown in Fig.5. The MSG-Net and DepthSRNet create unexpected artifacts when the depth of the test image changes, while our CU-Net has no such issue. This is because the CU-Net uses deeper features and applies the CAM after feature fusion to reduce the channel weights that may cause artifacts. Compared with the MRFs, RCAN and FDKN, our CU-Net only has errors at individual points on edge, attributed to multi-scale input and reconstruction.

4. CONCLUSIONS

In this paper, we propose Coupled U-Net (CU-Net) for color guided depth map super-resolution. The CU-Net has two U-Net branches to process the LR depth features and the HR color features, respectively. A channel attention module is designed to exploit the dependency between channels and amplify important features. To make the dual-branch features more tightly integrated, we introduce a multi-scale feature reconstruction module (MSFR) and a dual skip connection (DSC). Ablation studies and extensive comparative experiments verify the effectiveness and superiority of CU-Net.

5. ACKNOWLEDGEMENT

This work was supported by the Natural Science Foundation of China(No.61771276), the Natural Science Foundation of Guangdong Province(No.2020A1515010711) and the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen(Nos. JCYJ20200109143010272and JCYJ20200109143035495).

6. REFERENCES

[1] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *ICCV*, 2013, pp. 993–1000.

[2] Jingyu Yang, Xinchen Ye, Kun Li, Chunping Hou, and Yao Wang, "Color-guided depth recovery from RGB-D data using an adaptive autoregressive model," *TIP*, vol. 23, no. 8, pp. 3443–3458, 2014.

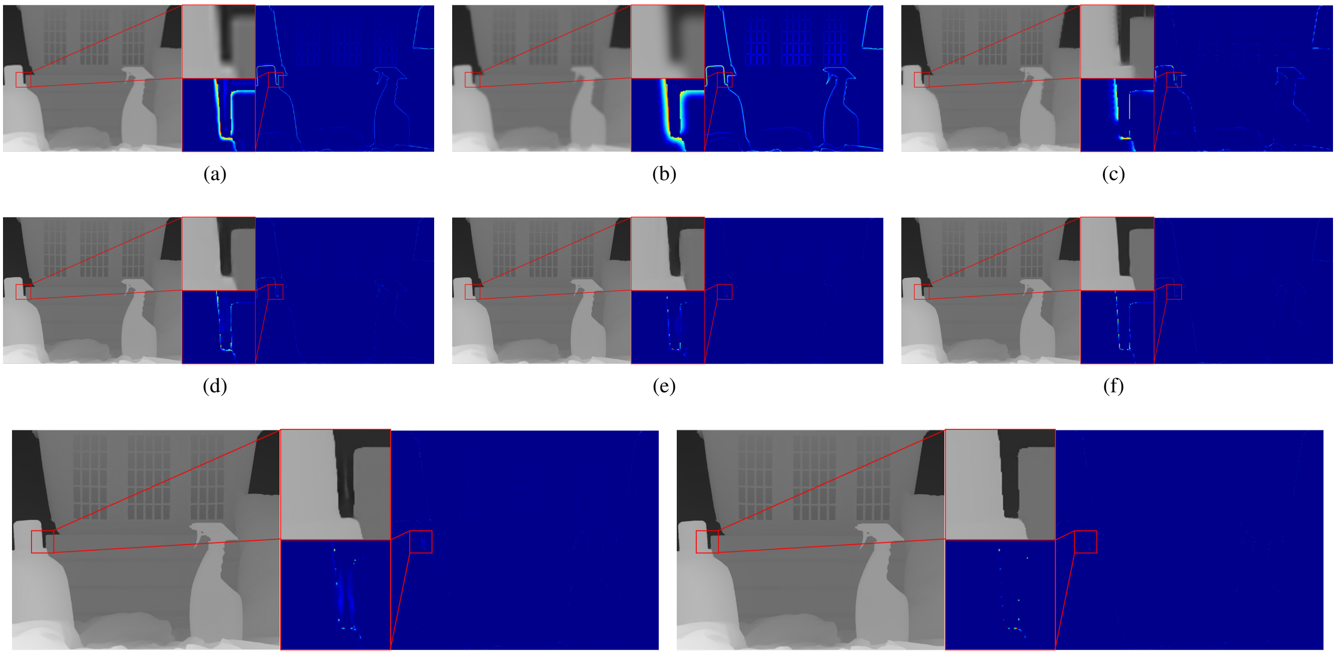
[3] Carlo Tomasi and Roberto Manduchi, "Bilateral filtering for gray and color images," in *ICCV*, 1998, pp. 839–846.

[4] Elmar Eisemann and Fredo Durand, "Flash photography enhancement via intrinsic relighting," *ACM Transactions on Graphics*, vol. 23, 2004.

[5] Ming-Yu Liu, Oncel Tuzel, and Yuichi Taguchi, "Joint geodesic upsampling of depth images," in *CVPR*, 2013, pp. 169–176.

[6] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.

[7] Gernot Riegler, Matthias Rüther, and Horst Bischof, "ATGV-Net: Accurate depth super-resolution," in *ECCV*, 2016, pp. 268–284.



(a) (b) (c)

(d) (e) (f)

(g) (h)

Fig. 5. **The visual comparison for 8× SR on "Laundry".** (a) Bicubic, (b) GF, (c) MRFs, (d) FDKN, (e) MSG-Net, (f) RCAN, (g) DepthSRNet, and (h) CU-Net. In each sub-figure, the left-hand panel is for the SR result; the right-hand panel is for the absolute error color map between the SR results and the ground truth; the middle panel is for two zoomed maps.

[8] Xibin Song, Yuchao Dai, Dingfu Zhou, Liu Liu, Wei Li, Hongdong Li, and Ruigang Yang, "Channel attention based iterative residual learning for depth map super-resolution," in *CVPR*, 2020, pp. 5631–5640.

[9] Yanjie Li, Tianfan Xue, Lifeng Sun, and Jianzhuang Liu, "Joint example-based depth map super-resolution," in *ICME*, 2012, pp. 152–157.

[10] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang, "Depth map super-resolution by deep multi-scale guidance," in *ECCV*, 2016, pp. 353–369.

[11] Chunle Guo, Chongyi Li, Jichang Guo, Runmin Cong, Huazhu Fu, and Ping Han, "Hierarchical features driven residual learning for depth map super-resolution," *TIP*, vol. 28, no. 5, pp. 2545–2557, 2019.

[12] Beomjun Kim, Jean Ponce, and Bumsub Ham, "Deformable kernel networks for joint image filtering," *IJCV*, 2020.

[13] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018, pp. 294–310.

[14] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black, "A naturalistic open source movie for optical flow evaluation," in *ECCV*, 2012, pp. 611–625.

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

[16] Daniel Scharstein and Chris Pal, "Learning conditional random fields for stereo," in *CVPR*, 2007, pp. 1–8.

[17] 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 *GCPR*, 2014, pp. 31–42.

[18] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010, pp. 249–256.

[20] James Diebel and Sebastian Thrun, "An application of Markov random fields to range sensing," *NIPS*, vol. 18, pp. 291–298, 2005.

[21] Martin Kiechle, Simon Hawe, and Martin Kleinsteuber, "A joint intensity and depth co-sparse analysis model for depth map super-resolution," in *ICCV*, 2013, pp. 1545–1552.

[22] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," *TPAMI*, vol. 35, pp. 1397–1409, 2013.

[23] Johannes Kopf, Michael F Cohen, Dani Lischinski, and Matt Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics (ToG)*, vol. 26, no. 3, pp. 96–es, 2007.