BridgeNet: A Joint Learning Network of Depth Map Super-Resolution and Monocular Depth Estimation

Qi Tang[1,2], Runmin Cong[1,2,4*], Ronghui Sheng[1,2], Lingzhi He[1,2], Dan Zhang[3], Yao Zhao[1,2], Sam Kwong[4]
[1]Institute of Information Science, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China
[3]UISEE Technology (Beijing) Co., Ltd. Beijing, China
[4]City University of Hong Kong, Hong Kong, China
{qitang,rmcong,21120297,lingzhihe,yzhao}@bjtu.edu.cn,dan.zhang@uisee.com,cssamk@cityu.edu.hk

ABSTRACT

Depth map super-resolution is a task with high practical application requirements in the industry. Existing color-guided depth map super-resolution methods usually necessitate an extra branch to extract high-frequency detail information from RGB image to guide the low-resolution depth map reconstruction. However, because there are still some differences between the two modalities, direct information transmission in the feature dimension or edge map dimension cannot achieve satisfactory result, and may even trigger texture copying in areas where the structures of the RGB-D pair are inconsistent. Inspired by the multi-task learning, we propose a joint learning network of depth map super-resolution (DSR) and monocular depth estimation (MDE) without introducing additional supervision labels. For the interaction of two subnetworks, we adopt a differentiated guidance strategy and design two bridges correspondingly. One is the high-frequency attention bridge (HABdg) designed for the feature encoding process, which learns the high-frequency information of the MDE task to guide the DSR task. The other is the content guidance bridge (CGBdg) designed for the depth map reconstruction process, which provides the content guidance learned from DSR task for MDE task. The entire network architecture is highly portable and can provide a paradigm for associating the DSR and MDE tasks. Extensive experiments on benchmark datasets demonstrate that our method achieves competitive performance. Our code and models are available at https://rmcong.github.io/proj_BridgeNet.html.

CCS CONCEPTS

• **Computing methodologies → Computer vision problems**.

KEYWORDS

Depth map, Super-resolution, Monocular Depth Estimation, Multi-task Learning.

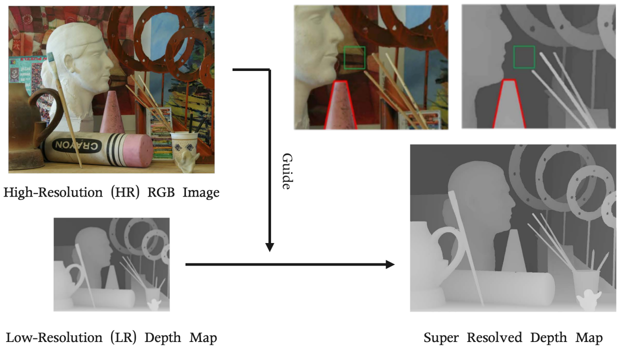

Figure 1: Color-guided depth map super-resolution. In the red trapezoidal region, the two have the same edge structure. While in the green rectangular region, there are more texture changes in the color image, but no corresponding texture structure in the depth map.

ACM Reference Format:
Qi Tang, Runmin Cong, Ronghui Sheng, Lingzhi He, Dan Zhang, Yao Zhao, Sam Kwong. 2021. BridgeNet: A Joint Learning Network of Depth Map Super-Resolution and Monocular Depth Estimation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475373

1 INTRODUCTION

When understanding a scene, people can not only perceive its appearance (*e.g.*, color, texture), but also capture the depth information to generate the stereo perception. Better scene understanding can facilitate many research areas such as autonomous navigation [26], 3D reconstruction [25], *etc*, all of which rely on high-quality depth information. The emergence and popularization of portable consumer-grade depth cameras, such as Microsoft Kinect and Lidar, provide great convenience for quickly acquiring the depth of scene [6]. However, due to the limitation of the imaging capabilities of depth cameras, the resolution of depth map is usually limited, let alone paired with high-resolution color image. Facing with the urgent demand for high-quality depth map in applications [5, 7, 9, 15, 17, 39], depth map super-resolution (SR) technology as a solution has attracted more and more attention.

Depth SR aims to super-resolve a low-resolution (LR) depth map to a high-resolution (HR) depth map. It not only needs to generate

[*] Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475373

the high-frequency counterpart (the high-frequency information of the image is the region where the gray level changes rapidly, such as the object edge) of the depth map from the degraded LR depth map, but also needs to effectively suppress the random noise and blur phenomenon in the imaging process. It is difficult to recover accurate HR depth map using artificially constructed filters or objective functions in the traditional filter-based methods [32, 33] and optimization-based methods [16, 35]. In recent years, with the success and application of deep learning technology, many works have also verified its role in depth SR task [22, 24, 45], which can reconstruct HR depth map by automatically learning stronger representations from data. In practical application scenarios, high-resolution color image is easily obtained and has strong structural similarity with the depth map [8, 10, 11], thus it can provide some guidance information for depth SR. We call this type of method color-guided depth super-resolution. Existing color-guided depth SR methods usually require an extra branch to acquire rich guidance information [12], and then use it to guide the hierarchical feature learning of the depth branch. However, the edges of the color image do not always coincide with the depth map. For example, as shown in Figure 1, the green rectangular region has a complex texture structure on the color image, but this area is displayed as a smooth region in the corresponding depth map. In this way, if we simply pass the RGB features or the extracted RGB edge features to the depth branch, it is easy to trigger issues such as texture copying and depth bleeding [29, 30]. Therefore, how to effectively explore the high-resolution color information is very important to depth SR task.

Before searching for a solution to the above problem, let us turn our attention to another depth-related task, namely monocular depth estimation, aiming to map a scene from the photometric representation to the geometrical representation. In layman's terms, the input of the monocular depth estimation is an RGB image, and the output is the estimated depth map. The two tasks are naturally related. The first is that the training datasets for these two tasks can be shared. If the two models are put in a unified framework for training, there is no need to introduce additional supervision labels (*e.g.*, semantic labels). In addition, orientated by the task of monocular depth estimation, the features learned from RGB image are more suitable for guiding the depth SR, because it can realize this kind of cross-modality information transformation in the continuous training and learning process. In summary, the joint learning of depth map super-resolution reconstruction and monocular depth estimation can achieve better color guidance without increasing the supervision information.

Motivated by the above analyses, we propose a joint learning network of depth super-resolution (DSR) and monocular depth estimation (MDE), namely BridgeNet, focusing on achieving better depth SR by effectively bridging the two tasks. To this end, we design two subnetworks based on the encoder-decoder architecture, *i.e.*, DSRNet and MDENet, which work together in a multi-task learning manner. Moreover, two different bridges in the encoder stage and decoder stage are designed to achieve differentiated guidance of two subnetworks. The encoder of MDENet learns the multi-level features oriented towards depth map from RGB image, which is suitable for guiding the depth SR branch. Thus, we propose a high-frequency attention bridge (HABdg) to learn the high-frequency information of MDENet to guide the depth SR branch. The feature decoders of MDENet and DSRNet are used to further extract task-oriented features for depth estimation and depth super-resolution. In contrast, the MDE task is more difficult than the DSR task because of its scale ambiguity. Therefore, following the principle of simple task guiding difficult task, we propose a content guidance bridge (CGBdg) to let DSRNet provide content guidance for MDENet in the depth feature space. In addition to associating the two tasks at the model design level, we also constrain them in terms of loss function, in the hope that the two subnetworks can promote each other.

To sum up, the contributions of this work are as follows:

- This work attempts to associate the depth map super-resolution and monocular depth estimation in a joint learning framework to boost the performance of depth map super-resolution, including a depth super-resolution network (DSRNet), a monocular depth estimation network (MDENet), and two joint learning bridges. Our entire network architecture is highly portable and can provide a paradigm for associating the DSR and MDE tasks. Moreover, different from other multi-task learning, no additional labels are required.
- The high-frequency attention bridge (HABdg) in the feature encoding stage transfers the RGB high-frequency information learned from MDENet to DSRNet, which can provide color guidance information closer to the depth modality. Following the principle of simple task guiding difficult task, we switch the guiding roles of the two tasks in the feature decoding stage, and propose the content guidance bridge (CGBdg) to provide the content guidance learned from DSRNet for MDENet.
- Without the introduction of additional supervision labels, our method achieves competitive performance on benchmark datasets.

2 RELATED WORK

Depth Map Super-Resolution. Due to the structural similarity between color image and depth map, numerous methods have been proposed to use color information to guide the reconstruction of degraded LR depth map. Zhao *et al.* [53] proposed a texture-depth conditional generation confrontation network to learn the structural similarity between texture images and low-resolution depth maps. Hui *et al.* [24] designed a multi-scale guided convolutional network for depth SR, which uses the rich hierarchical texture feature to eliminate the blurring phenomenon after depth map reconstruction. Zuo *et al.* [54] proposed a texture-guided enhanced residual dense network and a multi-scale fusion residual network to explore how to use the multi-scale guidance information provided by texture images to gradually guide the up-sampling and recovery of depth map. Ye *et al.* [50] proposed a progressive multi-branch aggregation network by using the multi-branch fusion method to gradually restore the degraded depth map. Guo *et al.* [19] used the U-Net structure to encode the interpolated depth image, and fused the texture features of the corresponding scale during the decoding process.

Monocular Depth Estimation. Monocular depth estimation is a typical inverse problem, since we are attempting to recover some

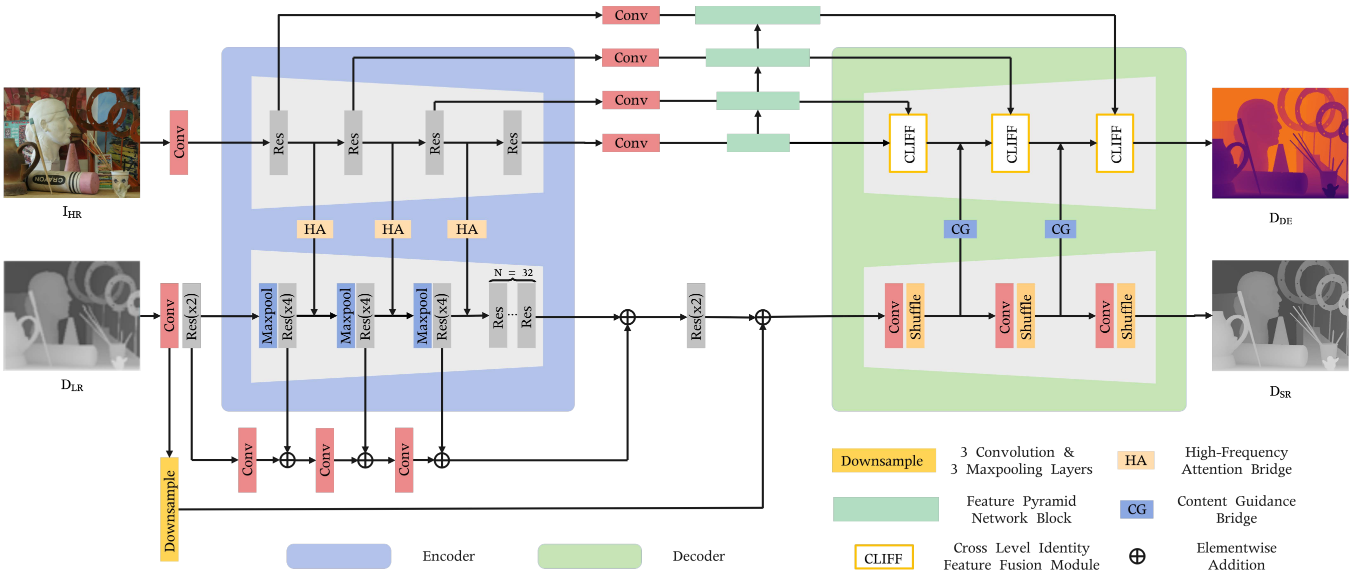

Figure 2: Architecture of the proposed BridgeNet, which consists of a depth super-resolution subnetwork (DSRNet), a monocular depth estimation subnetwork (MDENet), a high-frequency attention bridge (HABdg), and a content guidance bridge (CGBdg). The encoder-decoder structure at the top is the MDENet, and the bottom encoder-decoder structure corresponds to the DSRNet. The HABdg works between the feature encoders of the two subnetworks, focusing on passing the high-frequency color guidance obtained from MDENet to DSRNet. On the contrary, CGBdg works on the decoder side and is used to provide the MDENet with content guidance information learned from the DSRNet.

unknowns given insufficient information to fully specify the solution. Compared with the depth estimation task based on the left and right views, the monocular depth estimation has broader practical application prospects, but the current estimation performance is still very limited. Eigen *et al.* [14] estimated depth from a monocular image by using a convolutional neural network with two scales, which establish a precedent for deep learning based MDE method. To achieve up-sampling, Laina *et al.* [28] used a deeper residual network and small convolutions rather than large convolutions. Cao *et al.* [4] improved performance by discretizing the original continuous depth into a fixed number of ranges of predetermined width and transforming the depth regression task into a classification task.

Depth-Oriented Multi-Task Learning. The purpose of multi-task learning is to boost the performance of a specific task by combining other related tasks. The simplest and most common way is to combine the two tasks through the loss function. But this is often not optimal because of the lack of interaction in network design. At present, many tasks related to depth processing have adopted the multi-task learning strategy. Zhang *et al.* [52] proposed Task-Recursive Learning (TRL) framework recursively refine the results of semantic segmentation and monocular depth estimation tasks based on serializing the problems as a task-alternate time sequence. He *et al.* [21] proposed a SOSD-Net for simultaneous monocular depth estimation and semantic segmentation based on the geometric relationship of these two tasks. However, the correlation between the tasks modeled by these methods is still weak, and additional labels (such as semantic labels) are required for training. Sun *et al.* [42] proposed a knowledge distillation method to enable MDE to help DSR better understand the structure of the scene in the training process, thereby demonstrating effectiveness of MDE in improving DSR performance. By comparing the average pixel error of the two tasks in the training process to determine the interaction method of the two tasks, this does not directly explore the correlation between DSR and MDE.

In this paper, we propose a joint learning network of MDE and DSR to achieve better performance of DSR. Different form the previous work [42], when designing the interaction between the two subnetworks, we adopt a more explicit guidance mode. In the feature encoder, we let the MDE task provide high-frequency guidance information for the DSR task through the HABdg, so that the color guidance provided will be closer to the depth modality. In the feature decoder, we follow the principle of simple task guiding difficult task, and use DSR branch to provide content guidance for depth estimation branch via the CGBdg. Moreover, our method surpasses the work of [42] in performance of multiple evaluation indicators.

3 PROPOSED METHOD

3.1 Network Architecture

Figure 2 depicts the overall architecture of the proposed network, which consists of two subnetworks (*i.e.*, DSRNet and MDENet) and two bridges (*i.e.*, HABdg and CGBdg). The DSRNet and MDENet are equipped into a unified framework to achieve depth super-resolution and monocular depth estimation jointly, and the HABdg and CGBdg are respectively applied to the encoder and decoder stages to link these two tasks together. Given a set of HR RGB-D pairs $\{I_{HR}^{(n)}, D_{HR}^{(n)}\}_{n=1}^{N}$ and the corresponding LR depth maps $\{D_{LR}^{(n)}\}_{n=1}^{N}$ as training data, where N is the number of training images. Moreover, LR depth maps are interpolated to the size of HR depth maps. Our model takes the LR depth map (D_{LR}) and the corresponding HR RGB image (I_{HR}) as inputs to train the DSRNet and MDENet simultaneously. The super-resolved depth map (D_{SR}) is the main output of our network, and we also output the estimated depth map (D_{DE}) as the auxiliary.

Monocular Depth Estimation Subnetwork (MDENet). The encoder-decoder architecture has achieved great success in the monocular depth estimation. In our model, we follow the encoder-decoder network architecture used in [44] as our MDENet, which is mainly composed of three parts, that is, feature extractor, feature pyramid, and depth prediction. The feature extractor learns the multi-level features from the input RGB image, and the feature pyramid module propagates the high-level features to low-level features and generates the refined multi-level features. During the feature decoding, the cross-level identity feature fusion (CLIFF) module is used to progressively fuse the refined multi-level features and achieve depth estimation, where the interpolated high-level features and low-level features are fed as the inputs. It will refine the low-level features by multiplying them with the high-level features, so that accurate responses in the low-level features are further strengthened. Finally, the high-level features, original and refined low-level features are selected through two convolutional layers, thereby obtaining the most beneficial features to the MDE.

Depth Super-Resolution Subnetwork (DSRNet). Following the architecture of face super-resolution network [51], we also take the encoder-decoder network as our baseline of DSRNet. The encoder stage comprises a convolutional layer, a residual block, and three consecutive transformation modules, each of which module consists of a max-pooling layer and four residual blocks in series. Considering the positive effect of deeper network on super-resolution, we use some stacked residual blocks to further enhance the feature representation. Then, in order to recover the fine structure and tiny objects, we introduce the multi-scale strategy to guide the top-level features by further fusing the encoder features of the middle layers. Moreover, the shallow features in the encoder are fed into the Downsample block to generate the low-frequency features, and are combined with encoder features via a long skip connection, which forces the network to focus on learning high-frequency information in depth SR. The Downsample block consists of three convolution layers, each of which is followed by a max-pooling layer. During feature decoding, we use three identical modules connected sequentially, including a convolutional layer and a pixel shuffle layer. Feature maps of each layer are upsampled twice in resolution, and a 1 × 1 convolutional layer is finally applied to reconstruct the HR depth map.

Joint Learning Strategy. Depth super-resolution and monocular depth estimation have a natural correlation, and they can be trained under the supervision of the same dataset. Therefore, the joint learning of these two tasks is first manifested in the joint optimization of the loss function. Different from other multi-task learning methods whose loss function is a weighted sum of all branches, we assign different optimizers for the loss functions of DSR and MDE, respectively. The reason for this is that the learning difficulty of DSR and MDE is quite different, resulting in different convergence speeds of two tasks, and it is difficult to find a suitable weight setting to ensure that both tasks achieve the best performance. Therefore, in the design of the loss function, we separately optimize the parts related to DSR and MDE. Thus, their losses are defined as follows:

$$\mathcal{L}_{DSR} = ||D_{SR} - D_{HR}||_1 \tag{1}$$

$$\mathcal{L}_{MDE} = ||D_{DE} - D_{HR}||_1 \tag{2}$$

where $\mathcal{L}_{DSR}$ and $\mathcal{L}_{MDE}$ are pixel-wise L_1 loss for the task of DSR and MDE, respectively.

In addition to the loss constraint, we also elaborately design two bridges to associate the two tasks in the encoder and decoder stages respectively to achieve mutual benefit and common progress. One is the high-frequency attention bridge (HABdg) in the encoder, which uses the excellent color feature learning ability of the MDENet to provide guidance for the depth map super-resolution. The other is the content guidance bridge (CGBdg) in the decoder. Considering the difficulty of the two prediction tasks, the decoding features of DSRNet can provide effective content guidance for depth estimation, thereby improving the effect of depth estimation. In the following sections, we will stress the principles and details of these two bridges.

3.2 High-Frequency Attention Bridge

First, let us consider the correlation of the encoder parts of the two subnetworks. Recalling the previous color-guided depth SR methods, the guidance methods of color image mainly include direct corresponding feature guidance [12, 19, 24] or edge detail guidance [43, 49]. Although the color image and depth map have strong structural similarities, the abundant texture and edge of the color image are not always consistent with the depth map, so that the direct feature guidance or edge guidance may result in texture copying and depth bleeding. Looking back at the monocular depth estimation task, its purpose is to start from an RGB image, mapping photometric representation to geometrical representation, and then generate the depth map. Therefore, the features of color image provided by the depth estimation encoder are closer to the feature representation of depth modality, which can avoid the noticeable artifacts when providing high-frequency information guidance for DSR. This is why we use the depth estimation branch to guide the depth super-resolution branch in the encoder stage.

After determining the guiding direction of information transmission, the next question is how to effectively implement it. The simplest and most intuitive way is to pass the corresponding-layer features of the MDENet directly to the DSRNet through concatenation or addition. But this is obviously not a wise way. In the MDENet encoder, as the network deepens, the feature resolution gradually decreases, among which high-level features have rich semantic information, while low-level features have more structural information. Since the LR depth map contains less high-frequency information, it is more important that the HR color image can provide high-frequency information (such as the edge details) rather than the semantic information of the image. Motivated by this, we design a high-frequency attention bridge, which is specially used

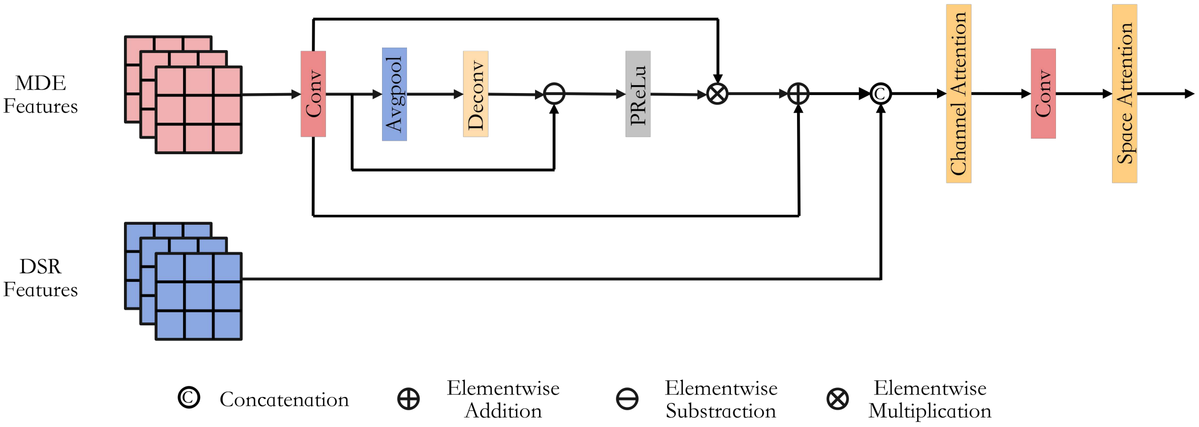

Figure 3: Illustration of HABdg. We first learn the high-frequency attention from the encoder features of the MDENet. Then, it is used to weight the original features to obtain the refined guidance features. After cascading with the features of the corresponding layer of DSRNet, the final output features (*i.e.*, the features of feeding into the next DSRNet encoder layer) are obtained through the CA and SA mechanisms.

to learn the high-frequency information from the MDENet to guide the depth SR branch. The pipeline of the HABdg is shown in Figure 3.

Specifically, we first use the average-pooling and deconvolution operations to blur the original features of the MDENet, which is formulated as:

$$F^i_{blurred} = deconv(avgpool(F^i_{MDE})), \tag{3}$$

where $F^i_{blurred}$ is the obtained blurred features of the i^{th} layer, F^i_{MDE} denotes the encoder features of the i^{th} layer in MDENet, $avgpool(\cdot)$ and $deconv(\cdot)$ are the average-pooling and deconvolution operations, respectively.

Then, we calculate the the high-frequency information by subtracting between the original features and the blurred ones, thereby generating the high-frequency attention weight:

$$A^i_{hf} = PRelu(F^i_{MDE} - F^i_{blurred}), \tag{4}$$

where A^i_{hf} denotes the high-frequency attention of the i^{th} layer, and $PRelu(\cdot)$ represents the parametric rectified linear unit. Next, we use the high-frequency attention to refine the original features F^i_{MDE} through residual connection and obtain the refined guidance features:

$$F^i_{hg} = F^i_{MDE} + A^i_{hf} \cdot F^i_{MDE}, \tag{5}$$

where F^i_{hg} represents the refined guidance features of the i^{th} layer. The intuition is to highlight the high-frequency in the original features, so that the LR depth map can better yield HR counterpart during feature fusion.

With the refined encoder guidance features of MDENet, we first concatenate them with the corresponding encoder features of DSRNet to generate the composition features F^i_{comp}. Such a simple feature fusion will have a lot of redundancy in the spatial dimension and channel dimension, thus we introduce an attention block including a channel attention [46] and a spatial attention [40] to enhance the fused features. The channel attention learns the importance of each feature channel, and spatial attention highlights the important spatial locations in the feature map. These processes can be formulated as:

$$F^i_{comp} = [F^i_{DSR}, F^i_{hg}], \tag{6}$$

$$F^i_{ha} = SA(conv_{1\times1}(CA(F^i_{comp}))), \tag{7}$$

where F^i_{DSR} denotes the encoder features of the i^{th} layer in DSRNet, CA and SA are channel attention and spatial attention blocks, respectively, $conv_{1\times1}$ is the convolutional layer with the kernel size of 1×1, and $[\cdot, \cdot]$ denotes the channel-wise concatenation operation. The output features F^i_{ha} will be used as the input of the next layer in DSRNet.

3.3 Content Guidance Bridge

For the feature decoding stages of MDENet and DSRNet, their roles are to further extract the task-oriented features for depth estimation and depth super-resolution. In this way, we can obtain the corresponding depth maps of the two subnetworks, either estimated or super-resolved. Compared with these two tasks, monocular depth estimation is known as the ill-posed inverse problem because of the scale ambiguity [14]. What it means is that many 3D scenes observed in the world can indeed correspond to the same 2D plane, that is, they are not in a one-to-one correspondence. As a result, training a model that maps well from RGB to depth is a very difficult task. Although the depth map super-resolution reconstruction is also an ill-posed problem, it learns the mapping in the same domain and focuses on restoring details, which is relatively simpler than monocular depth estimation. Therefore, due to the large gap between the performance of the two tasks, the decoder features of MDENet is no longer appropriate to providing guidance information for the decoding stage of DSRNet. Following the principle of simple task guiding difficult task, we exchange the guiding roles of the two subnetworks in the decoding stage, that is, let DSRNet provide content guidance for MDENet in the depth feature space. The detailed structure is illustrated in Figure 4.

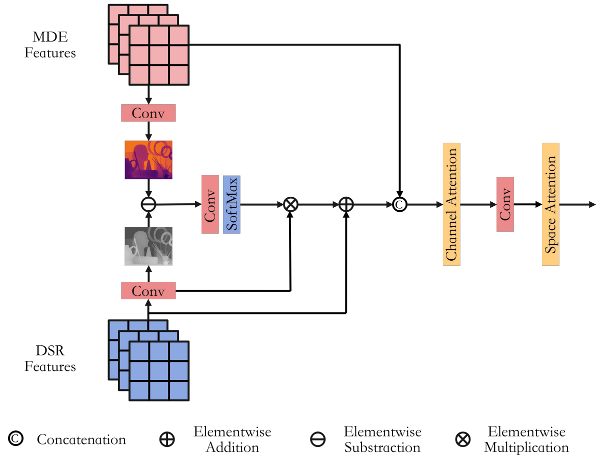

Figure 4: Illustration of CGBdg. We first calculate the difference map between estimated depth map and super-resolved depth map, and then learn the difference weight through a convolution operation and softmax activation. Applying the difference weight to the depth SR encoder features to generate the content guidance for the depth estimation branch. Finally, the depth estimation features and content guidance are concatenated, and fed into the channel attention and spatial attention modules to produce the output features.

As mentioned earlier, we can obtain two depth maps according to the decoder features of the two subnetworks. Specifically, we apply a convolutional layer with the kernel size of 1 × 1 to the decoder features, obtaining a super-resolved depth map and a estimated depth map:

$$M^i_{DSR} = conv_{1\times1}(Fd^i_{DSR}) \tag{8}$$

$$M^i_{MDE} = conv_{1\times1}(Fd^i_{MDE}) \tag{9}$$

where the Fd^i_{DSR} and Fd^i_{MDE} are the decoder features of the i^{th} layer in the DSRNet and MDENet respectively, M^i_{DSR} and M^i_{MDE} denote the depth maps of the i^{th} layer predicted by the DSRNet and MDENet respectively.

Then, we calculate the difference map between the estimated depth map M^i_{MDE} and super-resolved depth map M^i_{DSR}. The difference map highlights those positions in the estimated depth map that need to be further optimized relative to the super-resolved depth map. As the network is trained, we hope that this difference will become smaller and smaller. Based on this, we learn the difference weight by applying a convolution operation and softmax activation to the difference map, and further generate the content guidance for the depth estimation branch. The above procedures are formulated as:

$$W^i_{diff} = softmax(conv_{1\times1}(M^i_{DSR} - M^i_{MDE})) \tag{10}$$

$$F^i_{cg} = Fd^i_{DSR} + W^i_{diff} * Fd^i_{DSR} \tag{11}$$

where W^i_{diff} denotes the difference weight, F^i_{cg} is the content guidance of the i^{th} layer, and $softmax$ represents the softmax activation. Finally, similar to the HABdg, we still use the attention block to optimize the concatenation features (*i.e.*, $F^i_{con} = [Fd^i_{MDE}, F^i_{cg}]$), so as to obtain the features that are fed to the next layer of decoding block in MDENet.

4 EXPERIMENTS

4.1 Training and Implementation Details

We collect 36 RGB-D pairs from Middlebury dataset (6, 21, 9 images from 2001 [3], 2006 [23], and 2014 [37] datasets, respectively) for training, and 6 RGB-D pairs of the Middlebury 2005 [38] dataset for testing. Another training and testing dataset is NYU v2 dataset [39]. Following the common splitting method [31], we use the first 1000 pairs as training data, and evaluate on the last 449 pairs. The RGB-D pairs from training and testing are all normalized to the range of [0, 1].

Following the previous method [24], sufficient patches are cropped by dividing each HR image into a regular grid of small overlapping patches. This training tactic does not weaken the performance of network but it leads to lessen the training time. The HR patches are cropped into the squared size of 64, 128, and 256 according to the up-scaling factors of 4, 8, and 16, respectively. To produce the LR depth patches, we down-sample the HR depth patches to the fixed size of 16 × 16 by using the Bicubic interpolation. The metric of Mean Absolute Difference (MAD) and Root Mean Square Error (RMSE) are introduced for quantitative evaluation.

We implement our network with PyTorch and train with an NVIDIA 2080Ti GPU. We also implement our network by using the MindSpore Lite tool[1]. During training, a batch size of 8 is applied. We use ADAM with momentum of 0.9, $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-8}$ for network optimization. The learning rate is initiated to $1e^{-4}$, which will be decreased by multiplying 0.1 for every 100 epochs. Under ×8 Depth SR, the inference time for an image with the size of 256 × 256 is 0.052 second via the aforementioned GPU.

4.2 Performance Comparison

Middlebury Dataset. We compare with some state-of-the-art DSR methods under different up-sampling factors (× 4, × 8, and × 16), including six traditional depth SR methods (*i.e.*, CLMF [33], JGF [32], TGV [16], CDLLC [47], PB [2], and EG [48]) and seven deep learning based methods (*i.e.*, SRCNN [13], ATGVNet [36], MSG [24], DGDIE [18], DEIN [49], CCFN [45], GSRPT [12], and CTKT [42]).

Figure 5 demonstrates the visual comparisons of different methods under the factor of ×8. As visible, our method can recover more fine-grained depth details, *e.g.*, less artifacts around the stick in the Art image, more accurate shape of the toy head in the Dolls image. The comparison methods may generate some artifacts, blurred boundaries, or shape changes. Those phenomena are related to the severe damages on fine structure and tiny objects during down-sampling degradation, which brings more difficulties on these regions. In contrast, our method has advantages in accurately recovering the depth boundaries of these tiny objects. The quantitative comparisons are reported in Table 1, we can see that our network obtains the competitive result against other comparison methods, even under challenging scaling factors of ×8 and ×16. Taking the ×16 SR as an example, for the Books image, compared with the

[1] https://www.mindspore.cn/

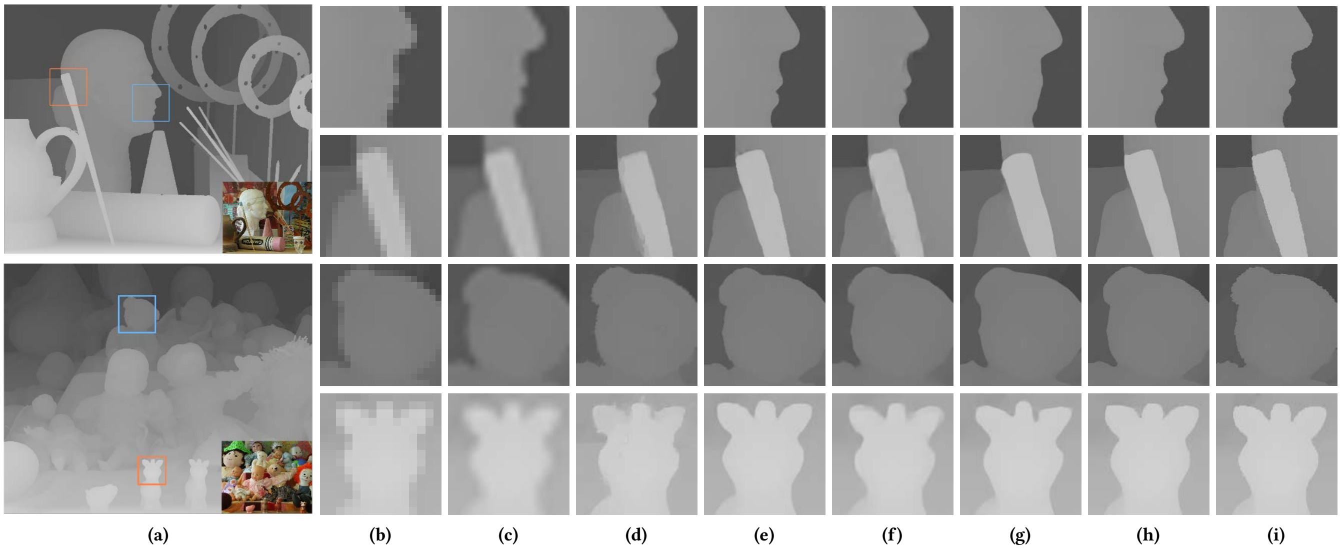

Figure 5: Visual comparisons of ×8 up-sampling results on two examples (*i.e.*, Art in the first row and Dolls in the second row). (a) Ground truth depth maps and color images; (b) LR depth patches; (c)-(h) The super-resolved depth maps generated by Bicubic, TGV [16], MSG [24], DGDIE [18], CTKT [42], and BridgeNet, respectively. (i) Ground truth. Depth patches are enlarged for clear visualization.

Table 1: Quantitative depth SR results (in MAD) on Middlebury 2005 dataset. The best performance is displayed in bold, and the second best performance is marked in underline.

	Art			Books			Dolls			Laundry			Mobius			Reindeer		
	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16
CLMF [33]	0.76	1.44	2.87	0.28	0.51	1.02	0.34	0.60	1.01	0.50	0.80	1.67	0.29	0.51	0.97	0.51	0.84	1.55
JGF [32]	0.47	0.78	1.54	0.24	0.43	0.81	0.33	0.59	1.06	0.36	0.64	1.20	0.25	0.46	0.80	0.38	0.64	1.09
TGV [16]	0.65	1.17	2.30	0.27	0.42	0.82	0.33	0.70	2.20	0.55	1.22	3.37	0.29	0.49	0.90	0.49	1.03	3.05
CDLLC [47]	0.53	0.76	1.41	0.19	0.46	0.75	0.31	0.53	0.79	0.30	0.48	0.96	0.27	0.46	0.79	0.43	0.55	0.98
PB [2]	0.79	0.93	1.98	0.16	0.43	0.79	0.53	0.83	0.99	1.13	1.89	2.87	0.17	0.47	0.82	0.56	0.97	1.89
EG [48]	0.48	0.71	1.35	0.15	0.36	0.70	0.27	0.49	0.74	0.28	0.45	0.92	0.23	0.42	0.75	0.36	0.51	0.95
SRCNN [13]	0.63	1.21	2.34	0.25	0.52	0.97	0.29	0.58	1.03	0.40	0.87	1.74	0.25	0.43	0.87	0.35	0.75	1.47
ATGVNet [36]	0.65	0.81	1.42	0.43	0.51	0.79	0.41	0.52	**0.56**	0.37	0.89	0.94	0.38	0.45	0.80	0.41	0.58	1.01
MSG [24]	0.46	0.76	1.53	0.15	0.41	0.76	0.25	0.51	0.87	0.30	0.46	1.12	0.21	0.43	0.76	0.31	0.52	0.99
DGDIE [18]	0.48	1.20	2.44	0.30	0.58	1.02	0.34	0.63	0.93	0.35	0.86	1.56	0.28	0.58	0.98	0.35	0.73	1.29
DEIN [49]	0.40	0.64	**1.34**	0.22	0.37	0.78	0.22	0.38	0.73	0.23	0.36	0.81	0.20	0.35	0.73	0.26	0.40	0.80
CCFN [45]	0.43	0.72	1.50	0.17	0.36	0.69	0.25	0.46	0.75	0.24	0.41	**0.71**	0.23	0.39	0.73	0.29	0.46	0.95
GSRPT [12]	0.48	0.74	1.48	0.21	0.38	0.76	0.28	0.48	0.79	0.33	0.56	1.24	0.24	0.49	0.80	0.31	0.61	1.07
CTKT [42]	**0.25**	**0.53**	1.44	**0.11**	0.26	0.67	**0.16**	0.36	0.65	**0.16**	0.36	0.76	**0.13**	0.27	0.69	**0.17**	0.35	0.77
BridgeNet (Ours)	0.30	0.58	1.49	0.14	**0.24**	**0.51**	0.19	**0.34**	0.64	0.17	**0.34**	**0.71**	0.15	**0.26**	**0.54**	0.19	**0.31**	**0.70**

second best algorithm, our method refreshes the MAD from 0.67 to 0.51, with the percentage gain of 23.9%.

NYU v2 Dataset. We also evaluate our method on the NYU v2 dataset and compare it with other SOTA methods, including Bicubic, TGV [16], EDGE [35], DJF [31], SDF [20], DGDIE [18], GbFT [1], PAC [41], SVLRM [34], DKN [27], and CTKT [42]. Figure 6 demonstrates the visual results of our method under the ×8 up-sampling. Whether in the red or yellow rectangular area, our model can accurately reconstruct the depth information and edges of the small objects. As reported in Table 2, our method achieves the best performance under the ×8 and ×16 up-sampling cases. Compared with the ***second best*** algorithm, the RMSE of our method reaches 2.63 under the scaling factor of ×8, with an improvement of 3.7%.

4.3 Ablation Study

In this section, we conduct comprehensive ablation studies to verify the designs in our BridgeNet. We report the ×8 depth SR results on Middlebury dataset under different experimental settings in Table 3. The 1^{st} and 2^{nd} rows are the results of DSRNet and MDENet baseline trained separately without any interaction under the same settings. The performance of MDENet is inferior to the DSRNet, which coincides with the previous analysis, that is, MDE task is more difficult than DSR. For the multi-task interaction, we first combine the two task through the simple loss function constraint, and the result is shown in the third row of Table 3. Compared with the DSRNet alone, the MAD of depth SR is improved to 0.363. Then, we gradually integrate HABdg and CGBdg into the framework to discuss their roles. When only using HABdg or CGBdg in the joint learning framework, results obtained are better than only using loss

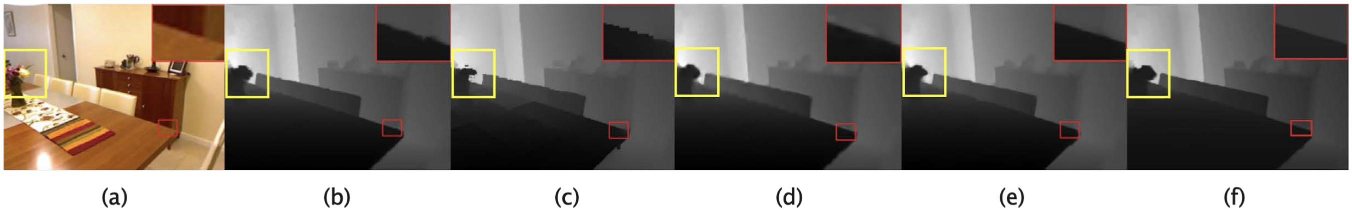

Figure 6: Visual comparisons of different method under ×8 up-sampling on the NYU v2 dataset. (a) Color image. (b) Ground truth. (c) SDF [20]. (d) DJF [31]. (e) SVLRM [34]. (f) BridgeNet.

Table 2: Quantitative depth SR results (in RMSE) on NYU v2 dataset. The best performance is displayed in bold, and the second best performance is marked in underline.

	Bicubic	TGV [16]	EDGE [35]	DJF [31]	SDF [20]	DGDIE [18]	GbFT [1]	PAC [41]	SVLRM [34]	DKN [27]	CTKT [42]	BridgeNet (Ours)
×4	8.16	6.98	5.21	3.54	3.04	1.56	3.35	2.39	1.74	1.62	**1.49**	1.54
×8	14.22	11.23	9.56	6.20	5.67	2.99	5.73	4.59	5.59	3.26	2.73	**2.63**
×16	22.32	28.13	18.10	10.21	9.97	5.24	9.01	8.09	7.23	6.51	5.11	**4.98**

Table 3: Ablation studies (in MAD) of our BridgeNet on the Middlebury 2005 dataset (×8 case).

	DSRNet	MDENet	HABdg	CGBdg	Middlebury
1	✓				0.366
2		✓			0.472
3	✓	✓			0.363
4	✓	✓	✓		0.355
5	✓	✓		✓	0.361
6	✓	✓	✓	✓	0.343

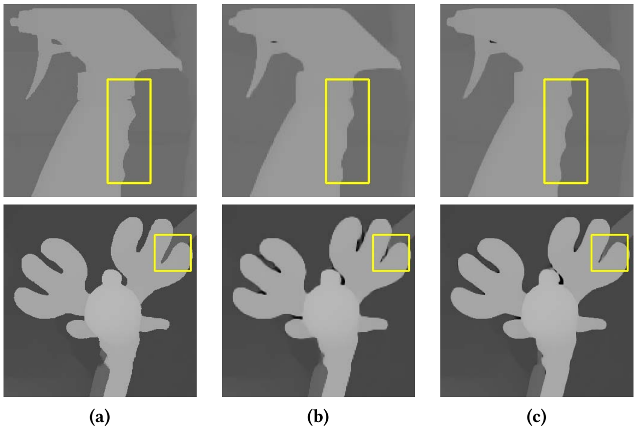

Figure 7: Visual comparisons of different components of our network (×8 case). (a) Ground truth. (b) DSRNet. (c) BridgeNet.

constraints. Moreover, when the two bridges work together (*i.e.*, full model), the performance reaches the best. We also provide some visual comparisons in Figure 7. From it, we can see that our model has clearer boundaries and more accurate depth values compared with DSRNet alone, as shown in the boxes of the figure.

To further demonstrate the effectiveness of HABdg, we replace it by feeding features of MDENet into DSRNet without any processing, as presented in Table 4. Compared with simply concatenating features in channel dimension and sending into DSRNet, our proposed HABdg can provide more effective high-frequency guidance for boosting the performance.

Table 4: Ablation studies (in MAD) of our HABdg on the Middlebury 2005 dataset (×8 case). 'w/o HABdg' refers to replacing the HABdg by directly propagating features from MDENet to DSRNet.

	Art	Books	Dolls	Laundry	Mobius	Reindeer	Avg.
w/ HABdg	0.58	0.24	0.34	0.34	0.26	0.31	0.343
w/o HABdg	0.65	0.25	0.37	0.38	0.28	0.33	0.376

5 CONCLUSION

In this paper, we explore a joint learning framework that combines the depth map super-resolution and monocular depth estimation to boost the depth SR performance without adding any other supervision labels. The center of this paper is how to design the interaction between the two subnetworks (*i.e.*, DSRNet and MDENet), thus we propose two bridges. On one hand, we let the MDENet provide high-frequency guidance information for the DSRNet through the HABdg in the feature encoder. On the other hand, we use the DSR branch to provide content guidance for depth estimation branch via the CGBdg in the feature decoder. Comprehensive experiments show that our method achieves competitive performance, especially for the cases of large scaling factors. Moreover, our network architecture is highly portable and can provide a paradigm for associating the DSR and MDE tasks.

ACKNOWLEDGMENTS

This work was supported by the Beijing Nova Program under Grant Z201100006820016, in part by the National Key Research and Development of China under Grant 2018AAA0102100, in part by the National Natural Science Foundation of China under Grant 62002014, Grant U1936212, in part by Elite Scientist Sponsorship Program by the China Association for Science and Technology under Grant 2020QNRC001, in part by General Research Fund-Research Grants Council (GRF-RGC) under Grant 9042816 (CityU 11209819), Grant 9042958 (CityU 11203820), in part by Hong Kong Scholars Program under Grant XJ2020040, in part by CAAI-Huawei MindSpore Open Fund, and in part by China Postdoctoral Science Foundation under Grant 2020T130050, Grant 2019M660438.

REFERENCES

[1] Badour Albahar and Jia-Bin Huang. 2019. Guided Image-to-Image Translation With Bi-Directional Feature Transformation. In *IEEE International Conference on Computer Vision*. 9015–9024.

[2] Oisin Mac Aodha, Neill D. F. Campbell, Arun Nair, and Gabriel J. Brostow. 2012. Patch Based Synthesis for Single Depth Image Super-Resolution. In *European Conference on Computer Vision*, Vol. 7574. 71–84.

[3] Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. 2011. A Database and Evaluation Methodology for Optical Flow. *International Journal of Computer Vision* 92, 1 (2011), 1–31.

[4] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. 2018. Estimating Depth From Monocular Images as Classification Using Deep Fully Convolutional Residual Networks. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 11 (2018), 3174–3182.

[5] Zuyao Chen, Runmin Cong, Qianqian Xu, and Qingming Huang. 2021. DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection. *IEEE Transactions on Image Processing* 30 (2021), 7012–7024.

[6] Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. 2019. Review of Visual Saliency Detection with Comprehensive Information. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 10 (2019), 2941–2959.

[7] Runmin Cong, Jianjun Lei, Huazhu Fu, Junhui Hou, Qingming Huang, and Sam Kwong. 2020. Going from RGB to RGBD Saliency: A Depth-Guided Transformation Model. *IEEE Transactions on Cybernetics* 50, 8 (2020), 3627–3639.

[8] Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Chunping Hou. 2018. Co-Saliency Detection for RGBD Images Based on Multi-Constraint Feature Matching and Cross Label Propagation. *IEEE Transactions on Image Processing* 27, 2 (2018), 568–579.

[9] Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Nam Ling. 2019. HSCS: Hierarchical Sparsity Based Co-Saliency Detection for RGBD Images. *IEEE Transactions on Multimedia* 21, 7 (2019), 1660–1671.

[10] Runmin Cong, Jianjun Lei, Huazhu Fu, Weisi Lin, Qingming Huang, Xiaochun Cao, and Chunping Hou. 2019. An Iterative Co-Saliency Framework for RGBD Images. *IEEE Transactions on Cybernetics* 49, 1 (2019), 233–246.

[11] Runmin Cong, Jianjun Lei, Changqing Zhang, Qingming Huang, Xiaochun Cao, and Chunping Hou. 2016. Saliency Detection for Stereoscopic Images Based on Depth Confidence Analysis and Multiple Cues Fusion. *IEEE Signal Processing Letters* 23, 6 (2016), 819–823.

[12] Riccardo de Lutio, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. 2019. Guided Super-Resolution As Pixel-to-Pixel Transformation. In *IEEE International Conference on Computer Vision*. 8828–8836.

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2014. Learning a Deep Convolutional Network for Image Super-Resolution. In *European Conference on Computer Vision*, Vol. 8692. 184–199.

[14] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *Advances in Neural Information Processing Systems*. 2366–2374.

[15] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. 2020. Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks. *IEEE Transactions on Neural Networks and Learning Systems* 32, 5 (2020), 2075–2089.

[16] David Ferstl, Christian Reinbacher, René Ranftl, Matthias Rüther, and Horst Bischof. 2013. Image Guided Depth Upsampling Using Anisotropic Total Generalized Variation. In *IEEE International Conference on Computer Vision*. 993–1000.

[17] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 2019. Real-Time 3D Hand Pose Estimation with 3D Convolutional Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 4 (2019), 956–970.

[18] Shuhang Gu, Wangmeng Zuo, Shi Guo, Yunjin Chen, Chongyu Chen, and Lei Zhang. 2017. Learning Dynamic Guidance for Depth Image Enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition*. 712–721.

[19] Chunle Guo, Chongyi Li, Jichang Guo, Runmin Cong, Huazhu Fu, and Ping Han. 2019. Hierarchical Features Driven Residual Learning for Depth Map Super-Resolution. *IEEE Transactions on Image Processing* 28, 5 (2019), 2545–2557.

[20] Bumsub Ham, Minsu Cho, and Jean Ponce. 2018. Robust Guided Image Filtering Using Nonconvex Potentials. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 1 (2018), 192–207.

[21] Lei He, Jiwen Lu, Guanghui Wang, Shiyu Song, and Jie Zhou. 2021. SOSD-Net: Joint Semantic Object Segmentation and Depth Estimation from Monocular images. *Neurocomputing* 440 (2021), 251–263.

[22] Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. 2021. Towards Fast and Accurate Real-World Depth Super-Resolution: Benchmark Dataset and Baseline. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9229–9238.

[23] Heiko Hirschmüller and Daniel Scharstein. 2007. Evaluation of Cost Functions for Stereo Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[24] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. 2016. Depth Map Super-Resolution by Deep Multi-Scale Guidance. In *European Conference on Computer Vision*, Vol. 9907. 353–369.

[25] Sunghoon Im, Hyowon Ha, Gyeongmin Choe, Hae-Gon Jeon, Kyungdon Joo, and In So Kweon. 2019. Accurate 3D Reconstruction from Small Motion Clip for Rolling Shutter Cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 4 (2019), 775–787.

[26] Christian Kerl, Jürgen Sturm, and Daniel Cremers. 2013. Dense Visual SLAM for RGB-D Cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2100–2106.

[27] Beomjun Kim, Jean Ponce, and Bumsub Ham. 2019. Deformable Kernel Networks for Guided Depth Map Upsampling. *ArXiv Preprint ArXiv:1903.11286* (2019).

[28] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. 2016. Deeper Depth Prediction with Fully Convolutional Residual Networks. In *International Conference on 3D Vision*. 239–248.

[29] Chongyi Li, Runmin Cong, Sam Kwong, Junhui Hou, Huazhu Fu, Guopu Zhu, Dingwen Zhang, and Qingming Huang. 2021. ASIF-Net: Attention Steered Interweave Fusion Network for RGB-D Salient Object Detection. *IEEE Transactions on Cybernetics* 50, 1 (2021), 88–100.

[30] Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy. 2020. RGB-D Salient Object Detection with Cross-Modality Modulation and Selection. In *European Conference on Computer Vision*. 225–241.

[31] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. 2016. Deep Joint Image Filtering. In *European Conference on Computer Vision*, Vol. 9908. 154–169.

[32] Ming-Yu Liu, Oncel Tuzel, and Yuichi Taguchi. 2013. Joint Geodesic Upsampling of Depth Images. In *IEEE Conference on Computer Vision and Pattern Recognition*. 169–176.

[33] Jiangbo Lu, Keyang Shi, Dongbo Min, Liang Lin, and Minh N. Do. 2012. Cross-Based Local Multipoint Filtering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 430–437.

[34] Jinshan Pan, Jiangxin Dong, Jimmy S. J. Ren, Liang Lin, Jinhui Tang, and Ming-Hsuan Yang. 2019. Spatially Variant Linear Representation Models for Joint Filtering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1702–1711.

[35] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S. Brown, and In-So Kweon. 2011. High Quality Depth Map Upsampling for 3D-TOF Cameras. In *IEEE International Conference on Computer Vision*. 1623–1630.

[36] Gernot Riegler, Matthias Rüther, and Horst Bischof. 2016. ATGV-Net: Accurate Depth Super-Resolution. In *European Conference on Computer Vision*, Vol. 9907. 268–284.

[37] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling. 2014. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. In *German Conference on Pattern Recognition*, Vol. 8753. 31–42.

[38] Daniel Scharstein and Chris Pal. 2007. Learning Conditional Random Fields for Stereo. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1–8.

[39] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *European Conference on Computer Vision*, Vol. 7576. 746–760.

[40] Vishwanath A. Sindagi and Vishal M. Patel. 2020. HA-CCN: Hierarchical Attention-Based Crowd Counting Network. *IEEE Transactions on Image Processing* 29 (2020), 323–335.

[41] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik G. Learned-Miller, and Jan Kautz. 2019. Pixel-Adaptive Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 11166–11175.

[42] Baoli Sun, Xinchen Ye, Baopu Li, Haojie Li, Zhihui Wang, and Rui Xu. 2021. Learning Scene Structure Guidance via Cross-Task Knowledge Transfer for Single Depth Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7792–7801.

[43] Jin Wang, Wei Xu, Jian-Feng Cai, Qing Zhu, Yunhui Shi, and Baocai Yin. 2020. Multi-Direction Dictionary Learning Based Depth Map Super-Resolution With Autoregressive Modeling. *IEEE Transactions on Multimedia* 22, 6 (2020), 1470–1484.

[44] Lijun Wang, Jianming Zhang, Yifan Wang, Huchuan Lu, and Xiang Ruan. 2020. CLIFFNet for Monocular Depth Estimation with Hierarchical Embedding Loss. In *European Conference on Computer Vision*, Vol. 12350. 316–331.

[45] Yang Wen, Bin Sheng, Ping Li, Weiyao Lin, and David Dagan Feng. 2019. Deep Color Guided Coarse-to-Fine Convolutional Network Cascade for Depth Image Super-Resolution. *IEEE Transactions on Image Processing* 28, 2 (2019), 994–1006.

[46] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional Block Attention Module. In *European Conference on Computer Vision*, Vol. 11211. 3–19.

[47] Jun Xie, Cheng-Chuan Chou, Rogério Schmidt Feris, and Ming-Ting Sun. 2014. Single Depth Image Super Resolution and Denoising via Coupled Dictionary Learning with Local Constraints and Shock Filtering. In *IEEE International Conference on Multimedia and Expo*. 1–6.

[48] Jun Xie, Rogério Schmidt Feris, and Ming-Ting Sun. 2016. Edge-Guided Single Depth Image Super Resolution. *IEEE Transactions on Image Processing* 25, 1 (2016), 428–438.

[49] Xinchen Ye, Xiangyue Duan, and Haojie Li. 2018. Depth Super-Resolution with Deep Edge-Inference Network and Edge-Guided Depth Filling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 1398–1402.

[50] Xinchen Ye, Baoli Sun, Zhihui Wang, Jingyu Yang, Rui Xu, Haojie Li, and Baopu Li. 2020. PMBANet: Progressive Multi-Branch Aggregation Network for Scene Depth Super-Resolution. *IEEE Transactions on Image Processing* 29 (2020), 7427–7442.

[51] Yu Yin, Joseph P. Robinson, Yulun Zhang, and Yun Fu. 2020. Joint Super-Resolution and Alignment of Tiny Faces. In *AAAI Conference on Artificial Intelligence*. 12693–12700.

[52] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. 2018. Joint Task-Recursive Learning for Semantic Segmentation and Depth Estimation. In *European Conference on Computer Vision*, Vol. 11214. 238–255.

[53] Lijun Zhao, Huihui Bai, Jie Liang, Bing Zeng, Anhong Wang, and Yao Zhao. 2019. Simultaneously Color-Depth Super-Resolution with Conditional Generative Adversarial Network. *Pattern Recognition* 88 (2019), 356–369.

[54] Yi-Fan Zuo, Yuming Fang, Yong Yang, Xiwu Shang, and Bin Wang. 2019. Residual Dense Network For Intensity-Guided Depth Map Enhancement. *Information Sciences* 495 (2019), 52–64.