Discrete Cosine Transform Network for Guided Depth Map Super-Resolution

Zixiang Zhao[1,2] Jiangshe Zhang[1*] Shuang Xu[1,3*] Zudi Lin[2] Hanspeter Pfister[2]
[1]Xi'an Jiaotong University, Xi'an, China
[2]Harvard University, Cambridge MA, USA
[3]Northwestern Polytechnical University, Xi'an, China
zixiangzhao@stu.xjtu.edu.cn, jszhang@mail.xjtu.edu.cn, xs@nwpu.edu.cn,
{linzudi,pfister}@g.harvard.edu

Abstract

Guided depth super-resolution (GDSR) is an essential topic in multi-modal image processing, which reconstructs high-resolution (HR) depth maps from low-resolution ones collected with suboptimal conditions with the help of HR RGB images of the same scene. To solve the challenges in interpreting the working mechanism, extracting cross-modal features and RGB texture over-transferred, we propose a novel Discrete Cosine Transform Network (DCTNet) to alleviate the problems from three aspects. First, the Discrete Cosine Transform (DCT) module reconstructs the multi-channel HR depth features by using DCT to solve the channel-wise optimization problem derived from the image domain. Second, we introduce a semi-coupled feature extraction module that uses shared convolutional kernels to extract common information and private kernels to extract modality-specific information. Third, we employ an edge attention mechanism to highlight the contours informative for guided upsampling. Extensive quantitative and qualitative evaluations demonstrate the effectiveness of our DCTNet, which outperforms previous state-of-the-art methods with a relatively small number of parameters. The code is available at https://github.com/Zhaozixiang1228/GDSR-DCTNet.

1. Introduction

With the popularity of consumer-oriented depth estimation sensors, *e.g.*, Time-of-Flight (ToF) and Kinect cameras, depth maps have promoted advancements in autonomous driving [24, 37], pose estimation [42, 56], virtual reality [20, 28], and scene understanding [10, 64]. Unfortunately, due to the technical limitations and suboptimal imaging conditions, depth images are often low-resolution (LR) and noisy. However, high-resolution (HR) RGB images (or intensity images) are relatively easy to obtain in the same scene when acquiring depth maps. Therefore, *guided depth map super-resolution* (GDSR) with RGB images has become an essential topic in multi-modal image processing and multi-modal super-resolution (SR). Our research is based on the assumption that there are statistical co-occurrences between the texture edges of RGB images and the discontinuities of depth maps [40]. In this way, information in RGB images can be utilized to restore HR depth maps when the LR depth maps are unsatisfactory for downstream applications.

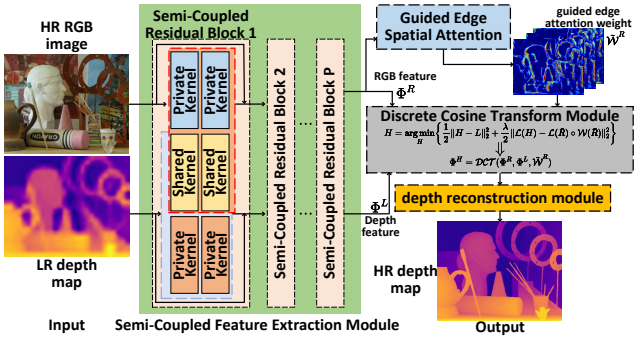


Figure 1. Overview of DCTNet. First, the SCFE module extracts shared and private features from the depth (LR) and RGB (HR) images. The GESA module employs the RGB feature to obtain edge attention weights useful for SR. The multi-modal features and attention weights are then processed by the DCT module, where DCT is utilized in each channel to get HR depth features. Finally, the reconstruction module outputs the SR depth map.

For image SR, deep neural networks have become the *de facto* methodology due to their ability in modeling the mapping from LR to HR images [5, 25, 62]. However, image SR mainly focuses on reconstructing fine details and textures, while depth SR models need to infer textureless and piecewise affine regions that have sharp depth discontinuities [40]. Besides, depth maps can be noisy and suffer from a lower tolerance for artifacts in real-world applications [54]. Therefore, we can hardly adopt the methods for image SR without appraising the unique characteristics of depth SR.

*Corresponding authors.

Conventional methods for GDSR can be divided into three categories, *i.e*., filter- [27, 29, 30, 33], optimization- [4,6,23,35,59] and learning-based [8,54,55] methods. Filter-based (or local) methods focus on preserving sharp depth edges under the guidance of the intensity image. However, for texture-rich RGB images, irrelevant edges may be transferred to depth images (known as texture *over-transferred*). In addition, the explicitly defined filters can only model a specific visual task and lack flexibility. Optimization-based (or global) methods design energy functions based on diverse data prior, with data-fidelity regularization terms constraint the solution space [38,63]. However, natural priors are often challenging to be explicitly represented and learned. The third category contains learning-based methods, which employ data-driven pipelines to learn the dependency between multi-modal inputs. Representative works in this category use sparse dictionary learning [15,19,51], which learn dictionaries in a group learning manner and set constraints on the sparse representations of different modalities [2,63].

Deep learning (DL) models are introduced to learn the mapping from LR to HR images [44,47,49,50,52,61], but they still often cooperate classic methods for depth upsampling. For example, learnable filter [16,53] (combination of DL and filter-based methods) and algorithm unrolling [3,58] (DL with optimization-based methods) have shown promising results. However, there are still challenges in conventional methods, including edge mismatch and texture over-transferred between the RGB/depth images, difficulty to learn of natural priors effectively, and limited interpretability for the internal mechanism of DL architectures.

To this end, we propose a *Discrete Cosine Transform Network* (DCTNet) for the GDSR task, inspired by coupled dictionary learning and physics-based modeling. It consists of four components: semi-coupled feature extraction (SCFE), guided edge spatial attention (GESA), discrete cosine transform (DCT) module, and a depth reconstruction (DR) module. The workflow is illustrated in Figure 1. Our contributions can be summarized as follows:

First, we propose the *semi-coupled* residual blocks to leverage the correlation between the intensity edge in RGB images and the depth discontinuities in depth images, but still preserve the unique properties like detailed texture and segment smoothness in two modalities. In each convolutional layer of this block, half of the kernels are responsible for extracting *shared* information in depth/RGB images, which is applied to both modalities. The rest half of the convolution kernels are designed to extract unique information in the depth and RGB images, respectively. Parameters in the *private* kernel are not shared. Thus the feature extractor with semi-coupled blocks can effectively extract informative features for GDSR from input image pairs.

Second, we propose a novel DCT module to improve the explainability of working mechanisms in the empirically-designed DL architectures. This component utilizes DCT to solve a well-designed optimization model for GDSR and inserts it in the DL model as a module to acquire HR depth map features guided by RGB features in the multi-channel feature domain. Therefore, besides learning the LR-to-HR mapping, our DCTNet focuses more on feature extraction and edge weight highlighting. Although recent works have used DCT for recognition [57] and image SR [32], we are the first to use it in restoring degraded depth maps to the best of our knowledge. We further make the tuning parameters in the DCT module learnable to improve model flexibility.

Third, to overcome the issue that texture details in RGB images are over-transferred, we employ the enhanced spatial attention (ESA) block from RFANet [26] in our GESA module to highlight the edges in RGB features useful for GDSR. In this way, part of the intensity edges is activated and associated with the depth discontinuities, achieving the adaptive transfer from the texture structure in guided images.

We conduct comprehensive evaluations on four popular RGBD datasets, including NYU v2 [43], Middlebury [13,41], Lu [31] and RGBDD [12]. The quantitative and qualitative results show that our DCTNet can achieve state-of-the-art performance in GDSR with a relatively small number of parameters.

2. Related Work

Super-resolution is a basic computer vision topic with many sub-fields and numerous approaches. Here we only discuss methods for GDSR.

2.1. Conventional GDSR methods

Filter-based methods. The filter-based (local) methods aim to use the RGB image to guide the joint filter to perceive the edge in depth map. Starting from joint bilateral upsampling [18] and its variants [1,60], RGB images guide the acquisition of bilateral weights. Liu *et al*. [27] replace the Euclidean distance with the geodesic distances to maintain the discontinuities of the depth image. Weighted mode filter [33], guided filtering [11] and its variants [30,48] are also widely used in the upsampling process. Lu *et al*. [29] use the smoothing method to process the image parts obtained by the depth map guided RGB image segmentation to solve the texture transferring issue.

Optimization-based methods. The optimization-based (global) methods model the interdependency between color images and depth maps by Markov Random Field [4], nonlocal means filtering [35], pixel-wise adaptive auto-regressive model [59], total generalized variation [6] and multi-pass optimization framework [23], respectively.

Learning-based methods. Earlier methods like bimodal co-sparse analysis [15] and joint dictionaries learning [51] capture the interdependency of the RGB and depth images.

A multi-scale dictionary learning strategy with RGB-D structural similarity measure and a robust coupled dictionary learning algorithm with local coordinate constraints are employed by Kwon *et al*. [19] and Xie *et al*. [55] to solve over-smoothing and over-fitting problems in information transfer, respectively. Gu *et al*. [8] establish a task-driven learning method to learn the dynamic guidance by a weighted analysis representation model. Xie *et al*. [54] learn an HR edge map inference method from external HR/LR image pair.

2.2. Deep Learning GDSR Methods

GDSR performance is further promoted with the powerful feature extraction capability of neural networks. Riegler *et al*. [39] adopt the first-order primal-dual algorithm and unroll the optimization processing to a network structure, establishing the relationship between the DL-based methods and the optimization-based methods. Li *et al*. [21, 22] use a two-stream end-to-end network with skip connection to learn the mapping of LR to HR depth maps. Hui *et al*. [14] propose multi-scale guidance for edge transfer. Similarly, Guo *et al*. [9] use the residual U-Net structure to learn the residual information between bicubic interpolation upsampling and ground truth under multi-scale guidance. CoIAST [2], which is based on the iterative shrinkage thresholding algorithm (ISTA) [7], regard the estimation of HR depth map as a linear combination of two LISTA branches. CU-Net [3] uses two modules to separate common/unique features by multi-modal convolutional sparse coding and elaborate the model interpretability. More recently, DKN [16] and FDSR [12] achieve adaptive filtering neighbors/weight calculation and high-frequency guided feature decomposition through spatially-variant kernels learning and octave convolution, respectively. They outperform previous state-of-the-art (SOTA) methods in synthetic and real scene datasets.

2.3. Comparison with existing approaches

Our proposed DCTNet is closely related to the optimization-based and DL-based coupled dictionary learning methods. (1) The DCT module in our model obtains the depth map features of HR by solving an optimization problem, and we are the first to use DCT to solve the problem to our knowledge. In addition, the DCT module is integrated into the DL framework to complete the multi-channel feature acquisition. The learnable parameters further enhance the flexibility of the optimization function in this module. (2) For the RGB texture over-transferred challenge, compared to local/global methods, we use the ESA module [26] to adaptively learn the edges attention weights in a data-driven manner. (3) Our feature extraction encoder is inspired by coupled dictionary learning, but we do not need to learn the dictionary explicitly. Instead, the private/shared feature extraction is accomplished by limiting whether the parameters are shared between the convolution kernels.

3. Method

In this section, we will elaborate on the details of our proposed DCTNet. We first show how to use discrete cosine transform (DCT) to solve an optimization problem for the GDSR task in the image domain. We then describe the architectural units and training objectives of DCTNet.

3.1. Problem formulation

We first define some important symbols for clarity. In the GDSR task, a model is expected to take the HR RGB image $R \in \mathbb{R}^{M\times N\times 3}$ and the LR depth image $\tilde{L} \in \mathbb{R}^{m\times n}$ as inputs, where $\{M, N\}$ and $\{m, n\}$ are the height and width of input RGB and depth images, respectively. We aim to obtain the HR depth image $H \in \mathbb{R}^{M\times N}$ under the guidance of R. We also perform some preprocessing to get $\tilde{R}$ and L, where $\tilde{R} \in \mathbb{R}^{M\times N}$ denotes the Y channel in YCrCb color space of R and $L \in \mathbb{R}^{M\times N}$ is the upsampled image of $\tilde{L}$. If R and $\tilde{L}$ in the same scene are given, H can be obtained by minimizing the following energy function:

$$\mathcal{F} = \frac{1}{2}\|H - L\|_2^2 + \frac{\lambda}{2}\|\mathcal{L}(H) - \mathcal{L}(\tilde{R}) \circ \mathcal{W}(\tilde{R})\|_2^2, \quad (1)$$

where $\mathcal{L}(\cdot)$ is the Laplacian filter, $\mathcal{W}(\cdot)$ can be regarded as a given threshold function to select the edges useful for GDSR. $\circ$ denotes element-wise multiplication, and λ is a parameter controlling the contribution of the second term. The optimal solution can be achieved when $\frac{\partial \mathcal{F}}{\partial H} = 0$, and we have

$$H + \lambda\mathcal{L}^2(H) = \lambda\mathcal{L}\left(\mathcal{L}(\tilde{R}) \circ \mathcal{W}(\tilde{R})\right) + L. \quad (2)$$

Eq. (2) can be treated as a 2D Poisson's equation (PE). Here we assume a "reflection padding" extension at the boundary of the image when performing convolution operations, which makes zero gradients on the image boundary. Thus, PE Eq. (2) is with the Neumann boundary condition (NBC). Technically, PE with the NBC can be solved via DCT [45]. Then we set $\lambda\mathcal{L}(\mathcal{L}(\tilde{R}) \circ \mathcal{W}(\tilde{R})) + L \triangleq E$, and implement DCT operation on both sides of the equation:

$$\mathcal{F}_c(H) + \lambda K^2 \circ \mathcal{F}_c(H) = \mathcal{F}_c(E), \quad (3)$$

where $\mathcal{F}_c(\cdot)$ is the DCT operation, $K_{ij} = \cos\left(\frac{i-1}{M}\pi\right) + \cos\left(\frac{j-1}{N}\pi\right)$ and $1 \le i \le M, 1 \le j \le N$. Finally, the HR depth images can be calculated by:

$$H = \mathcal{F}_c^{-1}\left\{\mathcal{F}_c(E) \oslash \left(I + \lambda K^2\right)\right\}, \quad (4)$$

where $\mathcal{F}_c^{-1}(\cdot)$ is the inverse DCT operation, $\oslash$ denotes the element-wise division, and I is the identity matrix. Due to space limitations, we refer readers to the supplementary material for the detailed derivation of the equation.

The above method has the following problems: (a) Although H can be solved by optimization, it requires additional edge perception methods to determine $\mathcal{W}(\cdot)$. (b) The

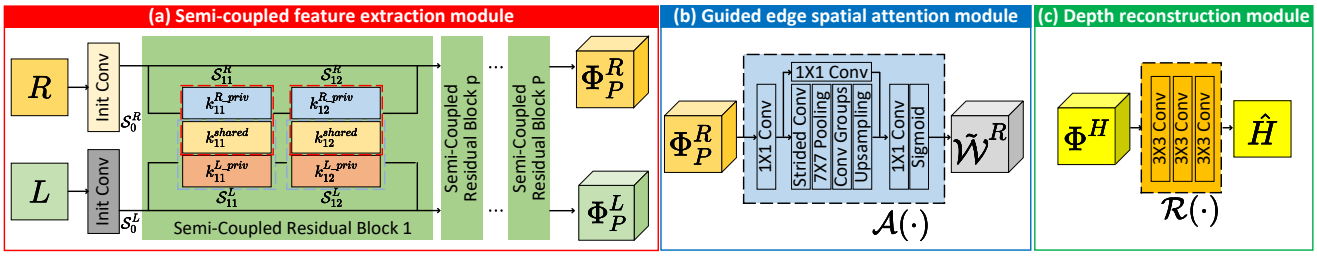


Figure 2. Detailed illustration of DCTNet workflow. Sub-figures (a)-(c) are the specific structures of SCFE, GESA, and DR modules in Fig. 1, which aims to extract cross-modality features, highlight RGB edge information, and reconstruct the HR depth map, respectively.

λ is manually given in Eq. (2), which restricts the model flexibility. (c) Optimizing a single channel in the image domain is difficult to effectively model the cross-modal internal feature correlation. Combining the challenges discussed in Sec. 1, *e.g.*, RGB texture over-transferred and the difficulty of natural prior learning, we propose a novel DCTNet in the following part to alleviate the above issues.

3.2. DCTNet

Our proposed DCTNet consists of four components including *semi-coupled feature extraction* (SCFE), *guided edge spatial attention* (GESA), *discrete cosine transform* (DCT) and *depth reconstruction* (DR) modules. The detailed illustrations are shown in Fig. 1 and 2.

We give an overview of the model. First, given a pair of L and R, the semi-coupled residual blocks extract shared and private features from the source images. The GESA module then processes the RGB feature to obtain the attention edge weights useful for SR. Subsequently, multi-channel RGB and depth features and the attention edge weights are input into the DCT module to acquire HR depth features. Finally, the depth reconstruction module outputs the SR depth map. The details of the modules are explained next.

3.2.1 Semi-coupled feature extraction

The RGB and depth maps in the same scene can have redundant information (*e.g.*, shape and edges) and complementary information (*e.g.*, RGB texture details and depth discontinuities). At the same time, based on the basic assumptions of the GDSR that some of the features in the cross-modal image should be interdependent, while other are modality-specific. Therefore, our SCFE module is designed to achieve cross-modal extraction of shared and private features.

As shown in Fig. 2(a), we can build the SCFE module as an encoder for feature extraction. The internal convolutions are two initial convolutions and P semi-coupled residual blocks. Here we denote the initial convolution layers corresponding to $\{L, R\}$ as $\{\mathcal{S}_0^L, \mathcal{S}_0^R\}$, and the qth convolution layer in the pth semi-coupled residual block corresponding to $\{L, R\}$ is denoted as $\{\mathcal{S}_{pq}^L, \mathcal{S}_{pq}^R\}$, where $p = 1, 2, \cdots, P$ and $q = 1, 2$. The output features of $\{\mathcal{S}_{pq}^L, \mathcal{S}_{pq}^R\}$ are denoted by $\{\Phi_{pq}^L, \Phi_{pq}^R\} \in \mathbb{R}^{M \times N \times C}$, where C is the number of kernels in $\{\mathcal{S}_{pq}^L, \mathcal{S}_{pq}^R\}$. P and C are determined in Sec 4.2. Note that when $q = 2$, $\{\Phi_{pq}^L, \Phi_{pq}^R\}$ can be simplified as $\{\Phi_p^L, \Phi_p^R\}$. The initialization layer generates $\Phi_0^R = \mathcal{S}_0^R(R)$, $\Phi_0^L = \mathcal{S}_0^L(L)$. Then taking the first convolution kernel in the pth semi-coupled residual block as an example, the semi-coupled convolution operation can be expressed as

$$\mathcal{S}_{p1}^R(\Phi_{p-1}^R) = \Phi_{p-1}^R * \mathcal{C}(k_{p1}^{shared}, k_{p1}^{R\_priv}), \quad (5)$$

$$\mathcal{S}_{p1}^L(\Phi_{p-1}^L) = \Phi_{p-1}^L * \mathcal{C}(k_{p1}^{shared}, k_{p1}^{L\_priv}), \quad (6)$$

where $*$ denotes convolution, $\{k_{p1}^{shared}, k_{p1}^{R\_priv}, k_{p1}^{L\_priv}\}$ denote the shared convolution kernels and the private ones corresponding to R and L, respectively. $\mathcal{C}(\cdot, \cdot)$ denotes the concatenation over the channel dimension. Then, the output feature Φ_p^R of R in the pth residual block becomes

$$\Phi_p^R = \texttt{ReLU}\left\{\mathcal{S}_{p2}^R(\texttt{ReLU}(\mathcal{S}_{p1}^R(\Phi_{p-1}^R))) + \Phi_{p-1}^R\right\}, \quad (7)$$

and that of Φ_p^L is similar to Eq. (7), only the superscript needs to be replaced from R to L. Finally, the outputs of SCFE module are Φ_P^L and Φ_P^R, which contain both the shared and the private features in the cross-modal image pair.

Compared with fully-shared or independent settings, the semi-coupled convolution kernels in the SCFE module can learn the shared/private parts of their respective input features, which extract features more effectively. The effectiveness of the SCFE module is demonstrated in Sec. 4.4.

3.2.2 Guided edge spatial attention

To prevent the problem that irrelevant textures are transferred to the SR depth map H when the guide RGB image contains rich textures, we adopt the ESA block from RFANet [26] that achieves excellent results in single image SR into our GESA module, as shown in Fig. 2(b). The ESA block can highlight the attention weight in a lightweight and efficient manner, facilitating the learning of discriminative features. This motivation meets our requirements for the GESA module. We use $\mathcal{A}(\cdot)$ to represent the operation in this module,

and the guided edge attention weight can be obtained by

$$\tilde{\mathcal{W}}^R = \mathcal{A}(\Phi_P^R) \in \mathbb{R}^{M\times N\times C}. \tag{8}$$

This module replaces the operation of obtaining $\mathcal{W}(\tilde{R})$ by manually giving $\mathcal{W}(\cdot)$ in Eq. (1). Thus, part of the edges in intensity features can be highlighted. Compared with conventional methods that manually design criteria to extract edge weights useful for upsampling, data-driven strategies can achieve adaptive extraction of attention weights.

3.2.3 Discrete cosine transform

In the above subsections, we have acquired the multi-channel features Φ^R and Φ^L corresponding to R, L[1], and guided edge attention weight $\tilde{\mathcal{W}}^R$. In this subsection, we will use them to accomplish the depth feature upsampling. In Eq. (4), we illustrate that given a pair of L, R and a threshold functions $\mathcal{W}(\cdot)$, the HR depth image can be reconstructed through DCT operation. Thus we consider the DCT algorithm as a module, which can be integrated into our DCTNet framework. Furthermore, it can be expanded to obtain the multi-channel HR depth map features by completing the DCT operation on each feature channel. Mathematically, the calculation of the DCT module, denoted as $\mathcal{DCT}(\cdot,\cdot,\cdot)$, is

$$\Phi^H = \mathcal{DCT}(\Phi^R, \Phi^L, \tilde{\mathcal{W}}^R), \tag{9}$$

where $\Phi^H \in \mathbb{R}^{M\times N\times C}$ is the guided upsampling feature of depth map L. More specifically, $\mathcal{DCT}(\cdot,\cdot,\cdot)$ calculates

$$\Phi^E[c] \triangleq \tilde{\lambda}_c \mathcal{L}\left(\mathcal{L}(\Phi^R[c]) \circ \tilde{\mathcal{W}}^R[c]\right) + \Phi^L[c], \tag{10}$$

$$\Phi^H[c] = \mathcal{F}_c^{-1}\left\{\mathcal{F}_c(\Phi^E[c]) \oslash \left(I + \tilde{\lambda}_c K^2\right)\right\}, \tag{11}$$

where $\Phi^H[c] \in \mathbb{R}^{M\times N}$ is the cth channel feature map of Φ^H. We want to emphasize that, compared with the manually given λ in Eq. (1) and Eq. (4), the $\tilde{\lambda} \in \mathbb{R}^C$ in Eq. (10) is set to be learnable. The channel-wise parameters are updated with the training progress, improving model flexibility.

To summarize, there are two main advantages of using the DCT module. First, besides $\tilde{\lambda} \in \mathbb{R}^C$, the acquisition of the feature map Φ^H is learning-free, which can reduce the network size with less learnable weights. Second, using the DCT operation to directly calculate the output features makes this component more interpretable than a neural network that usually works like a black box.

3.2.4 Depth reconstruction

Finally, the depth reconstruction module aims to predict the HR depth map from its feature map Φ^H, which is the output of the DCT module. The detailed structure is shown in Fig. 2(c). Specifically, the function $\mathcal{R}(\cdot)$ of this module can be expressed as $\hat{H} = \mathcal{R}(\Phi^H)$, where $\hat{H} \in \mathbb{R}^{M\times N}$ is the predicted HR depth map of DCTNet.

[1]We denote $\{\Phi_P^L, \Phi_P^R\}$ as $\{\Phi^L, \Phi^R\}$ for simplicity.

Impact of network depth P ($C = 64$)					
Setting	2	3	4	5	6
×4	2.378	1.989	1.544	**1.521**	1.527
×8	4.644	3.963	**3.152**	3.174	3.166
×16	8.245	6.904	**5.764**	5.787	5.776
Impact of network width C ($P = 4$)					
Setting	8	16	32	64	128
×4	2.798	2.300	1.992	1.544	**1.529**
×8	5.694	4.476	3.808	**3.152**	3.171
×16	9.531	7.695	6.976	5.764	**5.734**

Table 1. The impacts of depth P and width C on the DCTNet using the validation set. **Bold** indicates the best RMSE result.

3.2.5 Training loss

Consistent with recent works [3, 21, 22], we choose ℓ_2-loss as the training objective. That is, $\mathcal{D}(\hat{H}_i, H_i) = \sum_{i=1}^{N} \|\hat{H}_i - H_i\|_2^2$, where H_i is the ground truth HR depth map.

4. Experiment

In this section, we conduct comprehensive quantitative and qualitative experiments on several datasets to demonstrate the effectiveness of our proposed DCTNet.

4.1. Setup

Datasets. We use popular GDSR benchmarks following the protocol in recent works [16, 21, 22, 46]. Specifically, we select the first 1000 pairs of NYU v2 dataset [43] as the training set (900 pairs for training the network and 100 pairs for validation), and the last 449 pairs as a test set. We also utilize Middlebury [13, 41] (30 pairs) and Lu [31] (6 pairs) provided by Lu *et al*. [31] as the test sets. In addition, 405 pairs of images in the RGBDD dataset [12] are incorporated for evaluation. We train our DCTNet on the NYU v2 dataset [43] and test on the four datasets mentioned above.

In our experiments, all the LR depth images are synthesized by applying bicubic down-sampling of the HR depth maps. Finally, to verify the generalization ability of our models in natural scenes, we test them on the *real-world branch* of the RGBDD dataset[2] [12]. Please refer to the supplementary material for more descriptions of all the datasets.

Metric and implementation details. The training samples are resized to 256×256 in the pre-processing stage. The network is trained for 1000 epochs with a mini-batch size of

[2]This branch dataset contains 2215/405 pairs of RGBD images as training/test set. The LR depth maps and target HR depth maps are all acquired in real scenes, with the sizes of 192×144 and 512×384, respectively.

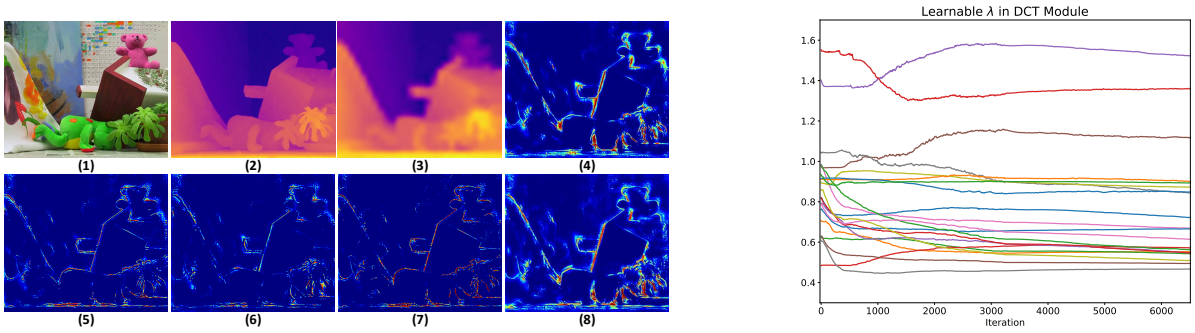


Figure 3. Visual results of (left) highlighted edge attention weights and (right) the changing curves of learnable λ. Left: (1)-(3): Input R, ground truth H and Input L, respectively. (4)-(8): Representative highlighted edge weights produced by the GESA module. Right: the values of the learnable parameters $\tilde{\lambda}$ against iterations during training. Different colored lines denote $\tilde{\lambda}_c$ corresponding to different channels.

64. We use the Adam [17] optimizer with a learning rate of 10^{-3}. In the test phase, we follow common practice and use the root-mean-square error (RMSE) to measure the depth SR performance against the ground-truth maps. A smaller RMSE implies a better quality of predicted depth images. The scripts are mainly implemented with Pytorch [36]. The training and testing are carried out on a PC with two NVIDIA GeForce RTX 3090 GPUs. We randomly initialize the learnable parameter $\tilde{\lambda}$ to e^{θ}, where $\theta \sim \mathcal{N}(0.1, 0.3)$. The number of semi-coupled residual blocks P and the kernel numbers C of semi-coupled filters in each convolution layer are set to 4 and 64, respectively. The choices of P and C are verified using the validation set in Sec. 4.2.

4.2. Validation experiments

Impact of network depth and width. For our proposed DCTNet, the network depth P and the width C play an important role in the effectiveness of super-resolution. We show the results among different combinations of $\{P, C\}$ on the validation set. We first fix $C = 64$, and calculate the prediction quality when $P = 2, 3, 4, 5, 6$ on the validation set. Then we verify the SR results for $C = 8, 16, 32, 64, 128$ when fixing $P = 4$. The results are demonstrated in Table 1. When $P < 4$, the model capability is restricted. When $P > 4$, increasing the depth does not achieve obvious performance gain but makes the model heavier. Similarly, when C exceeds 64, there is no significant performance improvement but increases the training cost. To have a good balance of model performance and computational cost, we set $\{P = 4, C = 64\}$ for the following experiments.

Highlighting edge attention weights. We visualize the first three and last two channels of the guided edge attention weights $\tilde{\mathcal{W}}^R$ in Eq. (8) from a representative sample pair (Fig. 3). We can clearly see that after the weight attention operation in the GESA module, the contour of the object is effectively highlighted, and the texture information inside the object is smoothed, which can alleviate the issue for texture over-transferred and benefit the GDSR task.

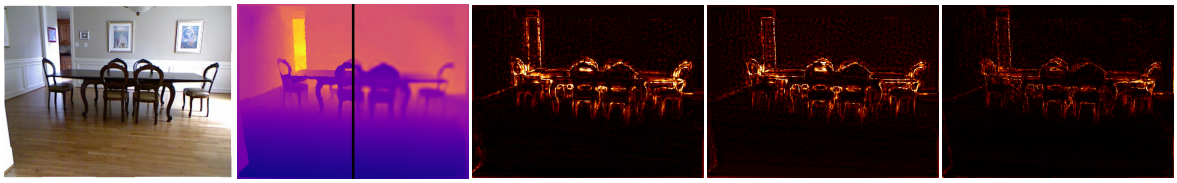
Evolution of the learnable parameters in DCT. One of our contributions is to make the tuning parameter $\tilde{\lambda}$ in Eq. (10) a list of channel-wise learnable parameters to improve the flexibility of DCTNet. Here we show the changing curve of $\tilde{\lambda}$ in each channel against the iteration number during training (Fig. 3). The plot shows that under the data-driven setting, $\tilde{\lambda}$ can adaptively adjust the importance between the fidelity term and the regular term. Compared with the manually given λ in Eq. (4), our design is more capable of leveraging the characteristics of different data domains.

4.3. Comparison with the state-of-the-arts

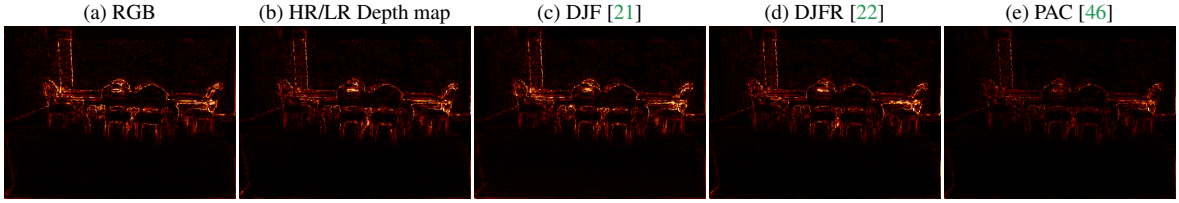
In this section, we test our DCTNet on the NYU v2, Middlebury, Lu and RGBDD benchmarks, and compare the results with state-of-the-art methods including DJF [21], DJFR [22], PAC [46], CUNet [3], DKN [16], FDKN [16] and FDSR [12] to demonstrate its performance.

Qualitative Comparison. We show the comparison of error maps for the SR depth maps in Fig. 4 and 5. Qualitatively, the depth predictions of DCTNet have lower prediction errors and are closer to the ground truth images. More visual comparisons are shown in the supplementary material.

Quantitative Comparison. The quantitative results on four test sets with scaling factors $\times 4$, $\times 8$, and $\times 16$ are shown in Tab. 2. Compared with existing approaches that only perform well on a certain dataset or super-resolution factor, our DCTNet achieves the best or second-best performance for multiple datasets and different super-resolution scales. This shows the advantage of our model upon previous state-of-the-arts. Moreover, following [12], for the *real-world branch* of RGBDD dataset, we use the $\times 4$ models trained in Tab. 2 to verify their generalization ability in real-world scenes. All the models are tested directly without additional finetuning. The quantitative results are shown in Table 3. Our proposed DCTNet achieves lower RMSE than previous

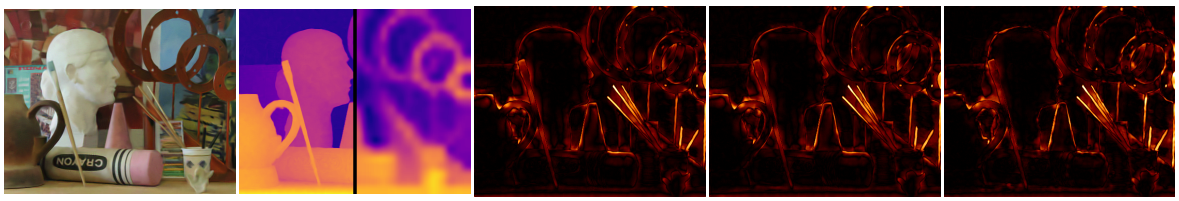


(a) RGB (b) HR/LR Depth map (c) DJF [21] (d) DJFR [22] (e) PAC [46]

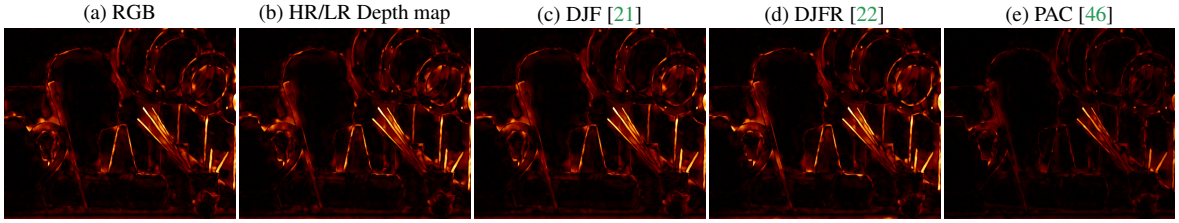


(f) CUNet [3] (g) DKN [16] (h) FDKN [16] (i) FDSR [12] (j) DCTNet (**Ours**)

Figure 4. Visual comparison of error maps for "Image_1365" in the NYU v2 dataset for 8× super-resolution.



(a) RGB (b) HR/LR Depth map (c) DJF [21] (d) DJFR [22] (e) PAC [46]



(f) CUNet [3] (g) DKN [16] (h) FDKN [16] (i) FDSR [12] (j) DCTNet (**Ours**)

Figure 5. Visual comparison of error maps for "05-Art" in the Middlebury dataset for 16× super-resolution.

methods, demonstrating its generalization ability.

Parameter Comparison. We discussed in Sec. 3 that the semi-coupled feature extraction (SCFE) module and the DCT module can reduce the number of learnable parameters while improving the interpretability of the model. Therefore, we show the number of model parameters *vs*. RMSE on the NYU v2 dataset in Fig. 6. Our model compares favorably against existing approaches with a relatively small number of parameters, demonstrating promising future directions in building lightweight network architectures.

4.4. Ablation Studies

We further validate the design choices of our DCTNet through ablation experiments (Tab. 4). Due to space limitations, we refer readers to the supplementary material for details about the network structures in Exp. III and V.

Semi-coupled filters. Besides the default semi-coupled filters in the SCFE module, we also tested independent (Exp. I) or fully-coupled (Exp. II) cases, where the parameters in each residual block are not shared or fully-shared, respectively. Exp. I result shows that the ability of independent kernels in extracting features is weaker than that of the semi-coupled filters, which demonstrates the necessity of using shared kernels to extract common features. On the other hand, Exp. II shows fully-shared filters lead to worse performance than the semi-coupled ones, which indicates the importance of considering the disparity between two modalities.

The DCT module. In Exp. III, we remove the DCT module and use a three-layer CNN to learn the mapping in Eq. (9). Removing the DCT module not only increases the number of learnable parameters but also reduces the prediction quality, which proves the effectiveness of the DCT module that follows the optimization-based methodology.

Learnable parameters $\tilde{\lambda}$. Instead of using learnable ones, we fix $\tilde{\lambda}$ to $e^{0.1}$ in Exp. IV (the mean of their initialization values). The result shows that a fixed tuning coefficient can reduce the flexibility of the model and restrict the SR ability.

Residual skip connection. In Exp. V, we remove the residual connection in the SCFE module and only use a stack of convolution kernels. The results show that residual connections play an important role in the feature extraction stage, as removing them leads to significant performance degradation.

Methods	Middlebury			NYU V2			Lu			RGBDD		
	×4	×8	×16	×4	×8	×16	×4	×8	×16	×4	×8	×16
DJF [21]	1.68	3.24	5.62	2.80	5.33	9.46	1.65	3.96	6.75	3.41	5.57	8.15
DJFR [22]	1.32	3.19	5.57	2.38	4.94	9.18	1.15	3.57	6.77	3.35	5.57	7.99
PAC [46]	1.32	2.62	4.58	1.89	3.33	6.78	1.20	2.33	5.19	1.25	1.98	3.49
CUNet [3]	1.10	2.17	4.33	1.92	3.70	6.78	0.91	2.23	4.99	1.18	1.95	3.45
DKN [16]	1.23	2.12	4.24	1.62	3.26	6.51	0.96	2.16	5.11	1.30	1.96	3.42
FDKN [16]	**1.08**	2.17	4.50	1.86	3.58	6.96	**0.82**	2.10	5.05	1.18	1.91	3.41
FDSR [12]	1.13	2.08	4.39	1.61	3.18	5.86	1.29	2.19	5.00	1.16	1.82	3.06
DCTNet (Ours)	1.10	**2.05**	**4.19**	**1.59**	**3.16**	**5.84**	0.88	**1.85**	**4.39**	**1.08**	**1.74**	**3.05**

Table 2. Quantitative comparison between our DCTNet and previous state-of-the-art approaches on four benchmark datasets. We use the RMSE metric (lower is better). The best and the second-best values are highlighted by **bold** and underline, respectively.

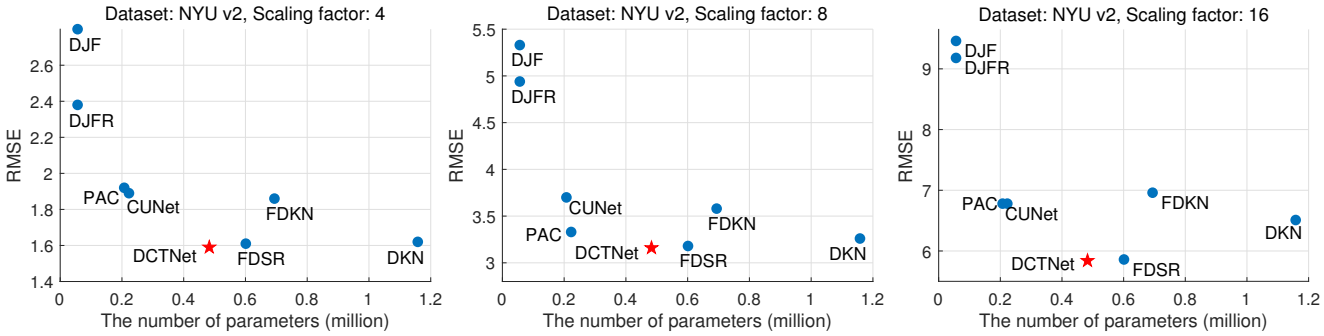


Figure 6. The number of model parameters *vs*. RMSE on the NYU v2 dataset for ×4, ×8, and ×16 SR scales. Our DCTNet (red star) achieves better or comparable performance with a relatively small number of parameters than existing models (blue dots).

Methods	RMSE	Methods	RMSE
SVLRM [34]	8.05	DKN [16]	7.38
DJF [21]	7.90	FDSR [12]	7.50
DJFR [22]	8.01	DCTNet	**7.37**
FDKN [16]	7.50		
FDSR* [12]	5.49	DCTNet*	**5.43**

Table 3. Quantitative results on the *real-world branch* of the RGBDD dataset. The best and second best values are highlighted by **bold** and underline, respectively. FDSR* and DCTNet* represent the results after finetuning on real-world branch data.

	Configurations	×4	×8	×16
I	w/ Independent Filters	1.74	3.34	6.06
II	w/ Fully-shared Filters	1.80	3.48	6.46
III	w/o DCT Module	1.78	3.46	6.61
IV	w/o Learnable Parameters	1.77	3.55	6.63
V	w/o Residual Connection	1.91	3.84	7.06
	Ours	**1.59**	**3.16**	**5.84**

Table 4. Results of ablation experiments on the NYU v2 test set. **Bold** indicates the best score in terms of RMSE.

4.5. Limitation

Although our proposed DCTNet is more interpretable and compares favorably against existing approaches with less learnable parameters, one limitation is that the components in the model make the formulation more complex than approaches using an end-to-end deep neural network to regress the HR depth map. In our future work, we will explore different ways to simplify the network design while keeping the merit of interpretability and the good tradeoff between network parameters and SR performance. We will also investigate challenges ubiquitous for most guided depth SR approaches, *e.g*., low illumination and blurry boundary in the paired RGB images.

5. Conclusion

In this paper, we propose a novel guided depth super-resolution (GDSR) model, DCTNet, based on discrete cosine transform, semi-coupled convolutional feature extraction, and adaptive edge attention. Our DCTNet incorporates intuitive motivations into the design choices to alleviate the challenges of RGB texture over-transferred, ineffective cross-modal feature extraction, and unclear working mechanism of network components in existing methods. In the future, we hope that more multi-modal image processing tasks can benefit from all or some components in DCTNet.

Acknowledgement

This work has been supported by the National Natural Science Foundation of China under Grant 61976174.

References

[1] Massimo Camplani, Tomás Mantecón, and Luis Salgado. Depth-color fusion strategy for 3-d scene modeling with kinect. *IEEE Trans. Cybern.*, 43(6):1560–1571, 2013. 2

[2] Xin Deng and Pier Luigi Dragotti. Deep coupled ISTA network for multi-modal image super-resolution. *IEEE Trans. Image Process.*, 29:1683–1698, 2020. 2, 3

[3] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3333–3348, 2021. 2, 3, 5, 6, 7, 8

[4] James Diebel and Sebastian Thrun. An application of markov random fields to range sensing. In *NIPS*, pages 291–298, 2005. 2

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014. 1

[6] David Ferstl, Christian Reinbacher, René Ranftl, Matthias Rüther, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *ICCV*, pages 993–1000. IEEE, 2013. 2

[7] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *ICML*, pages 399–406. Omnipress, 2010. 3

[8] Shuhang Gu, Wangmeng Zuo, Shi Guo, Yunjin Chen, Chongyu Chen, and Lei Zhang. Learning dynamic guidance for depth image enhancement. In *CVPR*, pages 712–721. IEEE Computer Society, 2017. 2, 3

[9] Chunle Guo, Chongyi Li, Jichang Guo, Runmin Cong, Huazhu Fu, and Ping Han. Hierarchical features driven residual learning for depth map super-resolution. *IEEE Trans. Image Process.*, 28(5):2545–2557, 2019. 3

[10] Saurabh Gupta, Ross B. Girshick, Pablo Andrés Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *ECCV*, pages 345–360. Springer, 2014. 1

[11] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(6):1397–1409, 2013. 2

[12] Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *CVPR*, pages 9229–9238. IEEE Computer Society, 2021. 2, 3, 5, 6, 7, 8

[13] Heiko Hirschmüller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*. IEEE Computer Society, 2007. 2, 5

[14] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *ECCV*, pages 353–369. Springer, 2016. 3

[15] Martin Kiechle, Simon Hawe, and Martin Kleinsteuber. A joint intensity and depth co-sparse analysis model for depth map super-resolution. In *ICCV*, pages 1545–1552. IEEE, 2013. 2

[16] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *Int. J. Comput. Vis.*, 129(2):579–600, 2021. 2, 3, 5, 6, 7, 8

[17] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6

[18] Johannes Kopf, Michael F. Cohen, Dani Lischinski, and Matthew Uyttendaele. Joint bilateral upsampling. *ACM Trans. Graph.*, 26(3):96, 2007. 2

[19] HyeokHyen Kwon, Yu-Wing Tai, and Stephen Lin. Data-driven depth map refinement via multi-scale sparse representation. In *CVPR*, pages 159–167. IEEE Computer Society, 2015. 2, 3

[20] Ling Li, Xiaojian Li, Shanlin Yang, Shuai Ding, Alireza Jolfaei, and Xi Zheng. Unsupervised-learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery. *IEEE Trans. Ind. Informatics*, 17(6):3920–3928, 2021. 1

[21] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *ECCV*, pages 154–169. Springer, 2016. 3, 5, 6, 7, 8

[22] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Joint image filtering with deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1909–1923, 2019. 3, 5, 6, 7, 8

[23] Yu Li, Dongbo Min, Minh N. Do, and Jiangbo Lu. Fast guided global interpolation for depth and motion. In *ECCV*, pages 717–733. Springer, 2016. 2

[24] Miao Liao, Feixiang Lu, Dingfu Zhou, Sibo Zhang, Wei Li, and Ruigang Yang. DVI: depth guided video inpainting for autonomous driving. In *ECCV*, pages 1–17. Springer, 2020. 1

[25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 1

[26] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, pages 2356–2365. IEEE Computer Society, 2020. 2, 3, 4

[27] Ming-Yu Liu, Oncel Tuzel, and Yuichi Taguchi. Joint geodesic upsampling of depth images. In *CVPR*, pages 169–176. IEEE Computer Society, 2013. 2

[28] Xianming Liu, Deming Zhai, Rong Chen, Xiangyang Ji, Debin Zhao, and Wen Gao. Depth super-resolution via joint color-guided internal and external regularizations. *IEEE Trans. Image Process.*, 28(4):1636–1645, 2019. 1

[29] Jiajun Lu and David A. Forsyth. Sparse depth super resolution. In *CVPR*, pages 2245–2253. IEEE Computer Society, 2015. 2

[30] Jiangbo Lu, Keyang Shi, Dongbo Min, Liang Lin, and Minh N. Do. Cross-based local multipoint filtering. In *CVPR*, pages 430–437. IEEE Computer Society, 2012. 2

[31] Si Lu, Xiaofeng Ren, and Feng Liu. Depth enhancement via low-rank matrix completion. In *CVPR*, pages 3390–3397. IEEE Computer Society, 2014. 2, 5

[32] Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *ICCV*, pages 4268–4277. IEEE, 2021. 2

[33] Dongbo Min, Jiangbo Lu, and Minh N. Do. Depth video enhancement based on weighted mode filtering. *IEEE Trans. Image Process.*, 21(3):1176–1190, 2012. 2

[34] Jinshan Pan, Jiangxin Dong, Jimmy S. J. Ren, Liang Lin, Jinhui Tang, and Ming-Hsuan Yang. Spatially variant linear representation models for joint filtering. In *CVPR*, pages 1702–1711. IEEE Computer Society, 2019. 8

[35] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S. Brown, and In-So Kweon. High quality depth map upsampling for 3d-tof cameras. In *ICCV*, pages 1623–1630. IEEE, 2011. 2

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[37] Wanli Peng, Hao Pan, He Liu, and Yi Sun. IDA-3D: instance-depth-aware 3d object detection from stereo vision for autonomous driving. In *CVPR*, pages 13012–13021. IEEE Computer Society, 2020. 1

[38] Haotong Qin, Yifu Ding, Xiangguo Zhang, Aoyu Li, Jiakai Wang, Xianglong Liu, and Jiwen Lu. Diverse sample generation: Pushing the limit of data-free quantization. *CoRR*, abs/2109.00212, 2021. 2

[39] Gernot Riegler, David Ferstl, Matthias Rüther, and Horst Bischof. A deep primal-dual network for guided depth super-resolution. In *BMVC*. BMVA Press, 2016. 3

[40] Gernot Riegler, Matthias Rüther, and Horst Bischof. Atgv-net: Accurate depth super-resolution. In *ECCV*, pages 268–284. Springer, 2016. 1

[41] Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *CVPR*. IEEE Computer Society, 2007. 2, 5

[42] Jamie Shotton, Ross B. Girshick, Andrew W. Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, and Andrew Blake. Efficient human pose estimation from single depth images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2821–2840, 2013. 1

[43] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, pages 746–760. Springer, 2012. 2, 5

[44] 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*, pages 5630–5639. IEEE Computer Society, 2020. 2

[45] Gilbert Strang. The discrete cosine transform. *SIAM review*, 41(1):135–147, 1999. 3

[46] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik G. Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *CVPR*, pages 11166–11175. IEEE Computer Society, 2019. 5, 6, 7, 8

[47] Baoli Sun, Xinchen Ye, Baopu Li, Haojie Li, Zhihui Wang, and Rui Xu. Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution. In *CVPR*, pages 7792–7801. IEEE Computer Society, 2021. 2

[48] Xiao Tan, Changming Sun, and Tuan D. Pham. Multipoint filtering with local polynomial approximation and range guidance. In *CVPR*, pages 2941–2948. IEEE Computer Society, 2014. 2

[49] Jiaxiang Tang, Xiaokang Chen, and Gang Zeng. Joint implicit image function for guided depth super-resolution. In *ACM Multimedia*, pages 4390–4399. ACM, 2021. 2

[50] Qi Tang, Runmin Cong, Ronghui Sheng, Lingzhi He, Dan Zhang, Yao Zhao, and Sam Kwong. Bridgenet: A joint learning network of depth map super-resolution and monocular depth estimation. In *ACM Multimedia*, pages 2148–2157. ACM, 2021. 2

[51] Ivana Tosic and Sarah Drewes. Learning joint intensity-depth sparse representations. *IEEE Trans. Image Process.*, 23(5):2122–2132, 2014. 2

[52] Yang Wen, Bin Sheng, Ping Li, Weiyao Lin, and David Dagan Feng. Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution. *IEEE Trans. Image Process.*, 28(2):994–1006, 2019. 2

[53] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Fast end-to-end trainable guided filter. In *CVPR*, pages 1838–1847. IEEE Computer Society, 2018. 2

[54] Jun Xie, Rogério Schmidt Feris, and Ming-Ting Sun. Edge-guided single depth image super resolution. *IEEE Trans. Image Process.*, 25(1):428–438, 2016. 1, 2, 3

[55] Jun Xie, Rogério Schmidt Feris, Shiaw-Shian Yu, and Ming-Ting Sun. Joint super resolution and denoising from a single depth image. *IEEE Trans. Multim.*, 17(9):1525–1537, 2015. 2, 3

[56] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2J: anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *ICCV*, pages 793–802. IEEE, 2019. 1

[57] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *CVPR*, pages 1740–1749. IEEE Computer Society, 2020. 2

[58] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *CVPR*, pages 1366–1375. Computer Vision Foundation / IEEE, 2021. 2

[59] Jingyu Yang, Xinchen Ye, Kun Li, Chunping Hou, and Yao Wang. Color-guided depth recovery from RGB-D data using an adaptive autoregressive model. *IEEE Trans. Image Process.*, 23(8):3443–3458, 2014. 2

[60] Qingxiong Yang, Ruigang Yang, James Davis, and David Nistér. Spatial-depth super resolution for range images. In *CVPR*. IEEE Computer Society, 2007. 2

[61] Xinchen Ye, Baoli Sun, Zhihui Wang, Jingyu Yang, Rui Xu, Haojie Li, and Baopu Li. Pmbanet: Progressive multi-branch aggregation network for scene depth super-resolution. *IEEE Trans. Image Process.*, 29:7427–7442, 2020. 2

[62] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *CVPR*, pages 3214–3223. IEEE Computer Society, 2020. 1

[63] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *CVPR*, pages 2808–2817. IEEE Computer Society, 2017. 2

[64] Yinda Zhang, Mingru Bai, Pushmeet Kohli, Shahram Izadi, and Jianxiong Xiao. Deepcontext: Context-encoding neural pathways for 3d holistic scene understanding. In *ICCV*, pages 1201–1210. IEEE, 2017. 1