



Article

# Depth Map Super-Resolution Based on Semi-Couple Deformable Convolution Networks

Botao Liu <sup>1</sup>, Kai Chen <sup>1,\*</sup>, Sheng-Lung Peng <sup>2</sup>  and Ming Zhao <sup>1</sup> 

<sup>1</sup> School of Computer Science, Yangtze University, Jingzhou 434023, China; liubotao920@163.com (B.L.); hitmzhao@gmail.com (M.Z.)

<sup>2</sup> Department of Creative Technologies and Product Design, National Taipei University of Business, Taipei 10051, Taiwan; slpeng@ntub.edu.tw

\* Correspondence: 201703455@yangtzeu.edu.cn

**Abstract:** Depth images obtained from lightweight, real-time depth estimation models and consumer-oriented sensors typically have low-resolution issues. Traditional interpolation methods for depth image up-sampling result in a significant information loss, especially in edges with discontinuous depth variations (depth discontinuities). To address this issue, this paper proposes a semi-coupled deformable convolution network (SCD-Net) based on the idea of guided depth map super-resolution (GDSR). The method employs a semi-coupled feature extraction scheme to learn unique and similar features between RGB images and depth images. We utilize a Coordinate Attention (CA) to suppress redundant information in RGB features. Finally, a deformable convolutional module is employed to restore the original resolution of the depth image. The model is tested on NYUv2, Middlebury, Lu, and a Real-Sense real-world dataset created using an Intel Real-sense D455 structured-light camera. The super-resolution accuracy of SCD-Net at multiple scales is much higher than that of traditional methods and superior to recent state-of-the-art (SOTA) models, which demonstrates the effectiveness and flexibility of our model on GDSR tasks. In particular, our method further solves the problem of an RGB texture being over-transferred in GDSR tasks.

**Keywords:** depth map super-resolution; guide image filter; deformable convolution; deep learning

**MSC:** 68T99



**Citation:** Liu, B.; Chen, K.; Peng, S.-L.; Zhao, M. Depth Map Super-Resolution Based on Semi-Couple Deformable Convolution Networks. *Mathematics* **2023**, *11*, 4556. <https://doi.org/10.3390/math11214556>

Academic Editors: Debo Cheng, Junbo Ma and Rongyao Hu

Received: 20 October 2023  
Revised: 2 November 2023  
Accepted: 3 November 2023  
Published: 5 November 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Depth information is widely applicable in augmented reality [1], pose estimation [2], autonomous driving [3], and robot navigation [4]. The industrial depth estimation sensors include time-of-fly (ToF), binocular cameras, and structured-light cameras to accurately obtain real-world depth information. Owing to the limitations of this technology, depth maps collected from consumer-oriented sensors have a low-resolution (LR) and sparse depth information, and they contain significant noise. Traditional interpolation methods, such as bilinear and bicubic interpolation, perform exceptionally poorly in handling sparse and noisy depth images. Meanwhile, high-resolution (HR) RGB images are more accessible to acquire compared to depth maps; however, both have remarkably similar low-frequency features. For example, depth maps of binocular cameras are obtained based on the principle of left and right visual differences in RGB images. He Kaiming proposed the concept of guided filtering in Guided Image Filtering [5] and discovered the local linear relationship between the guided and original images. Based on this theory, subsequent studies discovered a correlation between the discontinuous regions in depth maps and edges of RGB images [6]. Therefore, applying RGB images in guided depth map super-resolution (GDSR) has become a crucial research topic in cross-modal image processing. Recently, GDSR has been applied to reconstruct the collected LR depth maps (LRDMs) and assist with the visual depth estimation tasks [7].

The three main traditional GDSR methods are as follows: the filter method represented by joint bilateral filtering [8–11], optimizer method represented by Markov random fields [12–16], and learning method represented by sparse dictionary learning [17–19]. Filter methods use predefined filters to guide the depth maps using RGB images to preserve deep edges. However, owing to the rich textural information in RGB images, redundant edge features migrate to depth maps. Furthermore, artificially designed filters, which exhibit inferior generalization performance, cannot be applied to most scenarios. The optimizer method uses a predefined global energy equation to fit the depth map super-resolution problems; however, these tasks are difficult to express owing to their complexities. Learning-based methods, which have high computational complexities, are limited to low-dimensional signals.

Recently, numerous deep convolutional neural network (CNN)-based methods have been proposed [20–22], including the coarse-to-fine method progressive multi-branch aggregation network (PMBANet) [23] that surpasses the conventional accuracy and speed methods. These CNN-based methods have significantly improved the performance in GDSR tasks; however, they have limitations. During feature extraction, the convolution checks the entire input for the same operation and transfers the redundant features. The datasets used by these models are limited and augmented by random cropping or resizing of the NYUv2 dataset [24,25]. This dataset primarily comprises cluttered indoor data and a small collection of complicated objects in the GDSR tasks, which significantly affect the generalization ability of the model.

To this end, we propose a novel semi-coupled deformation convolutional network (SCD-Net) containing a semi-coupled feature extraction module and a learnable deformation convolution filter. Figure 1 shows the workflow of this model.

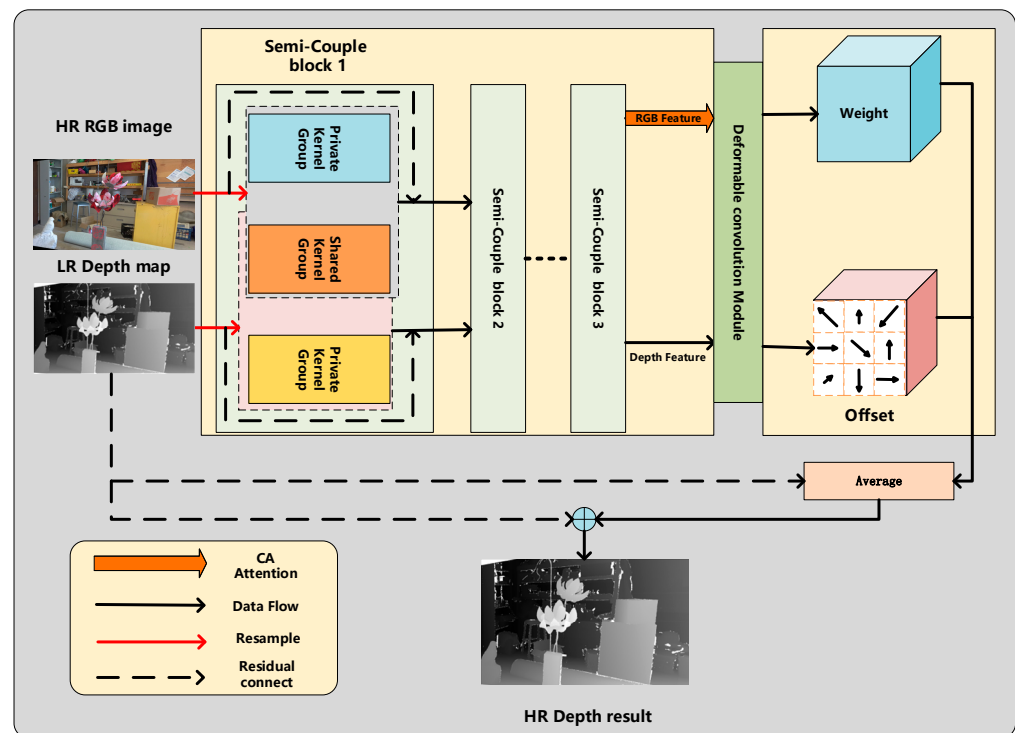


Figure 1. Overview of SCD-Net.

First, we describe the semi-coupled dictionary learning [26,27]. For feature extraction, a semi-coupled feature module was used to learn the connections between the edges in the RGB images and drastic changes in depth values at the interactions of the front and back scenes in the depth maps. Private kernels extract unique information from the depth and RGB images in each semi-couple block, whereas shared kernels extract common

information. The CA attention [28] was included after extracting the RGB features to highlight the valuable features. Therefore, this module can adaptively extract useful feature information for the GDSR.

Second, we propose a deformation convolution module to process the feature images obtained in the feature extraction section. Herein, the working mechanism of the deformation convolution [29] is transferred to the feature map processing, facilitating the element-wise learning of weights and offsets for paired RGB and depth features. To restore an HR depth map, we calculate the weighted average using the weight and original LRDM. The final depth result is obtained by concatenating it with the original LRDM.

Finally, we created the Real-sense dataset to address data scarcity. This dataset was collected using an Intel Real-sense D455 structured-light camera to match 1100 RGB depth-image pair sets. The Real-sense dataset supplements the missing data about human body postures, gestures, and plants in the NYUv2 dataset.

The proposed model was evaluated using the following datasets: NYU v2 [25], Middlebury [30], and Lu [31], and the collected data were sampled using Intel real-sense cameras. Our findings indicate that SCD-Net can achieve excellent results in GDSR tasks while maintaining its speed.

We analyze the current GDSR methods and datasets in the “Section 2”, pointing out their advantages and disadvantages. Following this, in the “Section 3”, we introduce the dataset we have created to address the issues of scarcity and poor quality of datasets in current GDSR tasks. In the “Section 4”, we provide a detailed explanation of our approach. In the “Section 5”, we validate the effectiveness of our method and dataset through qualitative and quantitative analyses, as well as multiple sets of ablation experiments. The “Section 6”, summarizes the entire work, highlights the existing issues, and proposes ideas for future improvements.

## 2. Related Work

### 2.1. Benchmark Datasets

Currently, multiple RGB-D datasets are used to train and test models. These datasets can be divided into synthetic and real scenes. The depth errors of synthetic deep datasets, which are obtained using a computer graphics software, approach zero. Commonly used methods include SceneFlow [32], Sintel [33], and New Tsukuba [34]. Virtual scenes cannot satisfy the demands of real-world situations; therefore, deep datasets of natural scenes have been constructed. The Middlebury [30] dataset provides high-quality RGB images and corresponding depth map samples containing noise from 2001 to 2021. The New York University created the NYUv2 dataset [25] containing 407,024 pairs of color and raw depth images of various indoor scenes obtained using Kinectv2 depth cameras (Microsoft corporation, Redmond, WA, USA). The team annotated 1449 RGB images and provided the corresponding depth maps.

### 2.2. GDSR Algorithm

This section discusses the GDSR algorithm based on the traditional GDSR and deep learning-based methods. We summarize the comparison of four main current GDSR algorithms from three directions. The details of various methods are as follows (Table 1).

**Table 1.** Summary of the performance of the four methods.

Directions \ Methods	Filter-Based	Optimizer-Based	Learning-Based	Deep Learning
Computational speed	Fast	Slowest	Fast	Fastest
Accuracy	Lowest	Low	High	highest
Interpretability	Strong	Strong	Poor	Poorest

### 2.2.1. Traditional GDSR Method

**Filter-based methods:** Kopf et al. [35] learned from bilateral filtering [36] and proposed that an HR input image can provide contextual a priori information during up-sampling to yield better reconstruction results. In 2009, Kaiming et al. [5] proposed a guided filtering method that has a local linear transformation relationship between the filtering output and guided image. This overcomes the gradient flipping phenomenon of bilateral filtering and yields good results in applications such as image denoising and joint up-sampling. Lo et al. [37] proposed a GDSR-based novel joint trilateral filtering method that integrates the local gradient information of depth maps while reconstructing their HR outputs, thus overcoming the limitations of edge discontinuity. Li et al. [38] proposed a joint example-based super-resolution method for depth maps that learns mapping functions from a set of training samples and improves the resolution of the depth maps through sparse encoding. Lu et al. [10] proposed a method for generating detailed HR depth maps using subsampled depth maps to solve texture transfer problems. These filter-based methods are limited in recovering high-frequency details. Furthermore, artificially designed filters exhibit inferior generalization.

**Optimizer-based methods:** These methods model the interactions between pixels in RGB images and depth maps represented by Markov random fields [12]. Optimization methods, such as the derived conditional random fields [39], total variation regularization [40], limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) [41], approximate message passing [42], and non-negative matrix factorization [43], exhibit excellent performance in GDSR. However, these methods have the following limitations: high computational complexity, difficulty in adjusting parameters, and tendency to generate visual artifacts.

**Learning-based methods:** Tomic et al. [44] proposed a joint dictionary-learning method to identify shared information and correlations between cross-modal data to accomplish the GDSR tasks. Zhang et al. [45] introduced multiscale dictionary learning into the SR method, integrating local and nonlocal priors to suppress reconstruction artifacts and enrich the visual details. Wang et al. [46] proposed a semi-coupled dictionary-learning method that can enhance the shared information and correlation between data domains. Meanwhile, it preserves the independence and original features of each data domain and reduces the complexity of the model. These learning-based methods have the following limitations: high computational complexity and difficulty in hyperparameter selection.

### 2.2.2. Deep Learning GDSR Methods

Owing to the limitations of the conventional methods, GDSR tasks have been transferred to deep learning.

In 2016, Li et al. designed a directed filtering method, Deep Joint Filtering (DJF) [47], which encodes the target and guided images, concatenates their feature maps, and convolutes these maps to obtain HR depth maps. This method outperforms the guided filtering methods. Zeng et al. [26] modeled the GDSR problem as a neural implicit image-interpolation problem and proposed using joint implicit image functions (JIIF) to represent the learned interpolation weights and values. Tang et al. [48] proposed a joint learning network called BridgeNet, which designs a high-frequency attention bridge to guide GDSR tasks by learning high-frequency information from monocular depth-estimation tasks. Gu et al. [19] proposed a weighted analysis representation model for guided depth image enhancement. This model contains a guided weight function and task-driven learning formulas to optimize and guide the GDSR tasks. He et al. [49] proposed a fast and adaptive depth map super-resolution model that can adaptively decompose high-frequency components from RGB images to guide depth maps in accomplishing super-resolution tasks. Guo et al. [21] employed a multiscale-guided method using residual U-net to learn the residual information between bilateral interpolation up-sampling and ground truth values. Wu et al. [50] proposed a fast end-to-end trainable direct-filtering model, emphasizing the requirement for HR input-to-output mapping. Herein, the input is downsampled,

and the task is completed at LR. This method significantly improves the speed and use of computing resources.

Recently, Kim et al. [24] proposed a deformation convolution joint upsampling method, Deformable Kernel Network (DKN), that achieves adaptive filtering and improves the super-resolution accuracy. The image generated using this method exhibits a high overall accuracy. However, the technique is limited by simplistic feature extraction, and the receptive field is returned to the original image by each convolutional kernel. The depth image restored using this method contains numerous pixel points with fitting errors. Zhao et al. proposed a Discrete Cosine Transform Network (DCTNet) [51], which shifts the GDSR task from the spatial domain to a frequency domain and uses the discrete cosine transform method to solve an optimization problem channel-wise. This method outperforms the previous state-of-the-art methods in various datasets; however, it overwhelms the GPU.

### 2.3. Comparison with Existing Methods

The proposed SCD-Net learns from the conventional semi-coupled dictionary learning methods [46], compensates for the limitations of DKN [24], replaces the feature extraction section with a semi-coupled feature extraction module, and solves the problem of RGB texture error transfer. Furthermore, CA [28] attention modules were implemented with deformable convolution modules to enhance the ability of the model to select correct cross-modal features adaptively. Finally, bicubic interpolation was used to restore the sampled subpixels, further improving the accuracy of the results.

## 3. Real-Sense Dataset

In the NYU dataset, approximately six-pixel areas at the four boundaries of the image contained no information. While cropping them can prevent the impact on the model, it increases the workload. Additionally, the dataset is expected to be sufficiently standardized and user-friendly.

**Collection device:** The NYUv2 dataset used a Kinect v2 ToF camera, which is disadvantageous in that the errors for close-range objects can only be accurate to the order of centimeters. The resolution of the generated depth map was low, and the edge information feedback for the sampled objects was poor. The depth map generated by the structured-light camera has relatively straightforward edges. Furthermore, it can attain a resolution of  $1080 \times 720$  pixels with a minimum sampling error of 0.01 mm for nearby objects. Hence, an Intel RealSense D455 camera (Intel corporation, Santa Clara, CA, USA) was used as the acquisition device. The camera was recalibrated to ensure that the depth information coverage of the collected depth images was sufficiently high. Finally, the depth-map pair resolution was set according to that of NYUv2 at  $640 \times 480$  pixels.

**Data processing:** The obtained depth maps and RGB images cannot be directly used for training. Figure 2 shows that the collected depth maps often contain many holes, indicating missing information. Furthermore, the RGB and depth images are not aligned. To obtain aligned data, internal and external parameter matrices of the camera were used to process the image pairs. Then, the RGB pixels correspond to those of the depth map. After that, the following methods used in the NYUv2 dataset were employed:

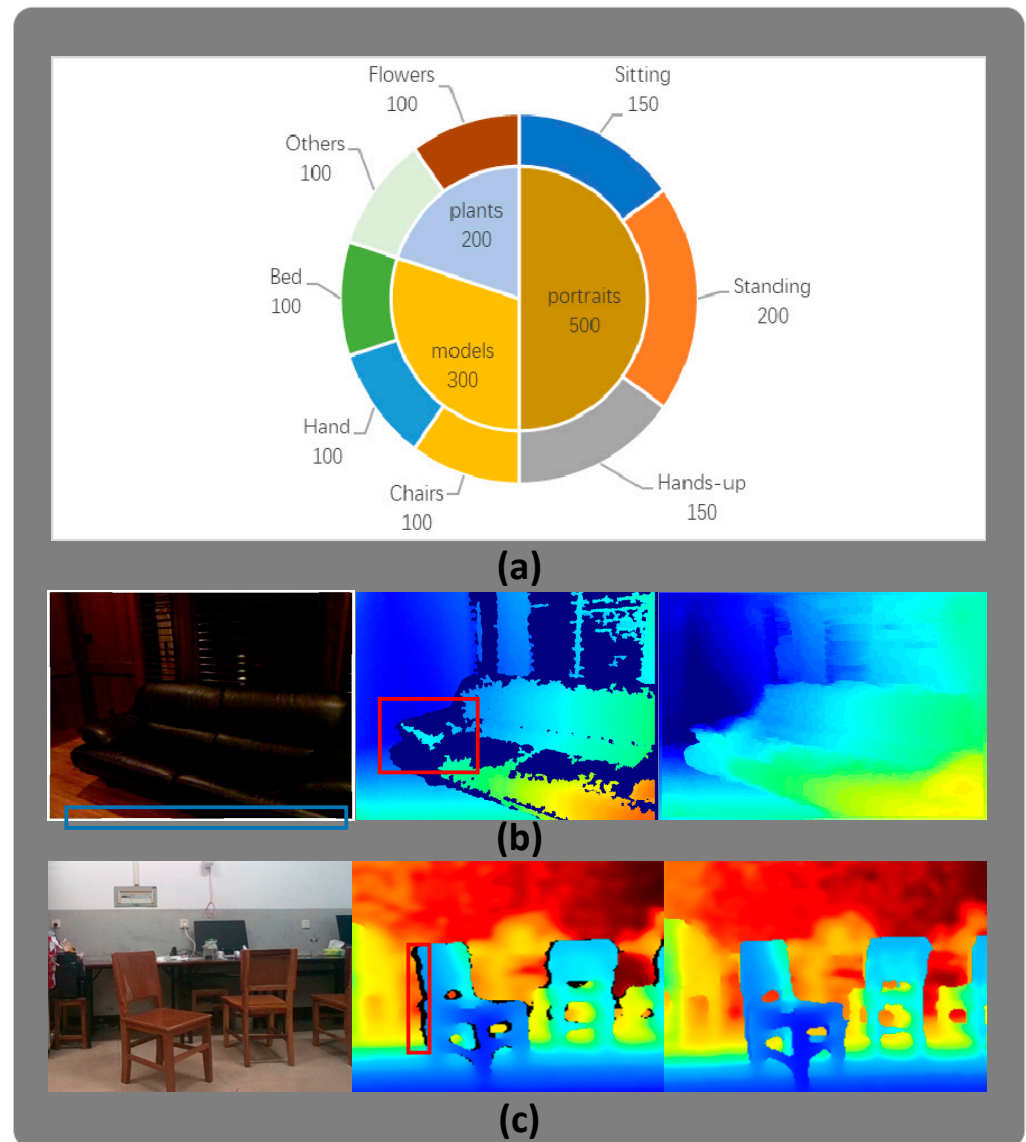
- (i) Implement the over-segmentation algorithm [27] to obtain rich color image boundary information. Then, use the coloring method [52] to fill the holes under the edge information supervision of the RGB image.
- (ii) A depth completion model [53] is employed to restore the collected images directly.

The final ground truth (GT) depth map was used as the optimal image selected between the two methods after human evaluation.

**Data composition:** Depth acquisition was performed for various scenarios. The deficiencies present in conventional depth datasets were supplemented, explicitly addressing the absence of specific targets:

- (i) Attitudes of people from different backgrounds.

- (ii) Various models, such as chairs and backpacks.
- (iii) Densely intertwined plants, such as flowers.



**Figure 2.** Statistics of the Real-sense dataset. (a) The scenes and the corresponding hierarchical content structures of the Real-sense dataset. (b) Example from the NYUv2 dataset (from left to right: RGB, raw depth, and GT). (c) Measure from the Real-sense dataset (from left to right: RGB, raw depth, and GT). Red and blue borders indicate the missing depth value and invalid boundary for NYU image data, respectively.

To ensure the accuracy of the data, we collected the data within the recommended range of 0.6–4 m for the camera. Figure 2 shows the data composition.

Figure 2b,c indicate that the depth coverages of most of the original HR depth maps in our dataset exceeded 85%. The prediction of depth values in depth-missing areas achieved higher accuracy during the depth completion process. Therefore, the accuracy of the GT image was greater than that of the NYUv2 dataset.

## 4. Methods

### 4.1. Problem Formula

Essentially, the GDSR tasks establish a relationship between LR depth maps (LRDMs) and HR-guided RGB images through the optimal objective function comparing the differ-

ence between them. In this relationship, RGB images of HR are used for guidance. The LR and GT depth maps are defined as  $D_L$  and  $D_H$ , respectively ( $D_L$  has  $H/K$  and  $W/K$  sizes). The HR-guided RGB image is defined as  $R_H$  (with image sizes  $H$  and  $W$ ), where  $K$  is set as the scaling factor. The mapping between  $D_L$  and  $R_H$  can be written as

$$D_H = G(D_L, R_H; \partial) \tag{1}$$

where  $G$  represents the established network model and  $\partial$  represents the parameters learned by the network. Using the energy equation to represent  $G$ , we have

$$G = \frac{1}{2} \|D_H - \hat{D}_L\|_2^2 + \frac{\gamma}{2} \|F(D_H) - F(\hat{R}_H) \cdot W(\hat{R}_H)\|_2^2, \tag{2}$$

where  $\hat{D}_L$  is the upsampled image of  $D_L$ ;  $\gamma$  controls the contribution of the second term;  $F()$  is a learnable filter;  $\hat{R}_H$  is the grayscale image of  $R_H$ ;  $\cdot$  is the matrix multiplication operator; and  $W()$  is the given threshold function that selects the useful edges for the GDSR problem. According to Equation (2), the proposed method overcomes the following limitations:

- (i) Learning a suitable threshold function to select the beneficial edges for the task;
- (ii) Modifying the model to adaptively select the value of  $\gamma$ . The following section presents a novel SCD-Net model to address these issues and prevent the excessive RGB texture migration in the cross-modal image processing.

#### 4.2. Overall Network Architecture

The proposed SCD-Net addresses the above three issues sequentially, as shown in Figure 3. It first utilizes a semi-coupled approach for feature extraction; then, it applies a CA attention for feature processing; it finally employs deformable convolution modules for feature fusion. Figure 1 shows a detailed diagram of SCD-Net with the following modules:

- (i) A set of HR RGB images and LRDMs is presented. The semi-coupled residual module extracts the shared and private features from the source image.
- (ii) For each element in the feature map, a set of matching weights and offsets are learned, enabling the filter to extract information beneficial to the task from different images.
- (iii) The obtained weights and offsets are multiplied and concatenated with the original LRDM to obtain an HR feature map.

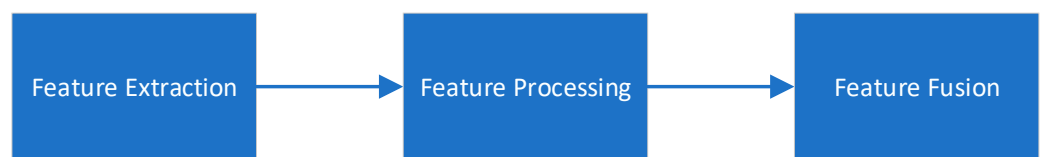


Figure 3. Method procedure.

##### 4.2.1. Semi-Couple Feature Extractor

Based on the fundamentals of GDSR tasks, a set of corresponding cross-modal images comprising correlated and independent features is created. For instance, the edges in RGB images are height-correlated with the depth-discontinuities in depth maps, whereas detailed textures in RGB images are absent in depth maps. However, in current deep learning-based GDSR methods, it is impossible to effectively extract cross-modal features by cascading RGB and depth image features. Therefore, the semi-coupled feature extraction module emulates the coupled dictionary-learning method to achieve cross-modal extraction between the shared and private features. Figure 4 shows that the input RGB and depth images were resampled to obtain a feature pair ( $H/4, W/4, 16$ ).

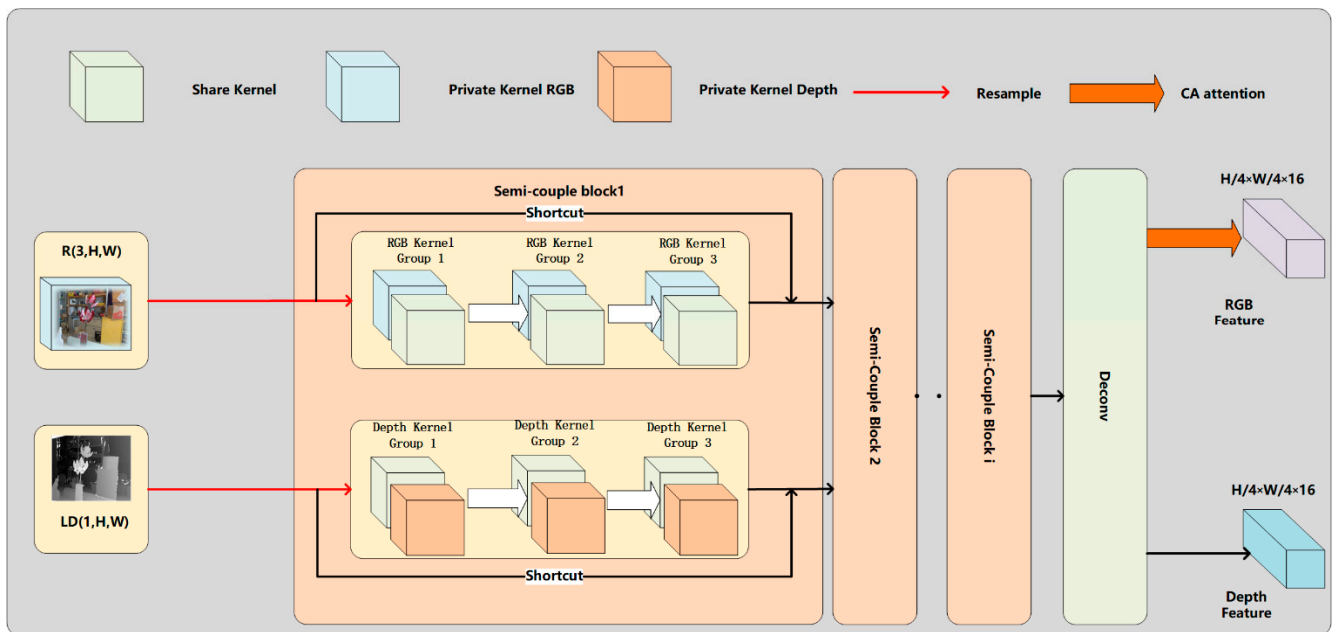


Figure 4. Semi-couple feature extractor.

Figure 5 illustrates the sampling method, wherein the samples were fed into the semi-coupled feature extraction module. In this module, the convolution kernels for RGB images and depth images comprise their respective private convolution kernels and a portion of shared convolution kernels.

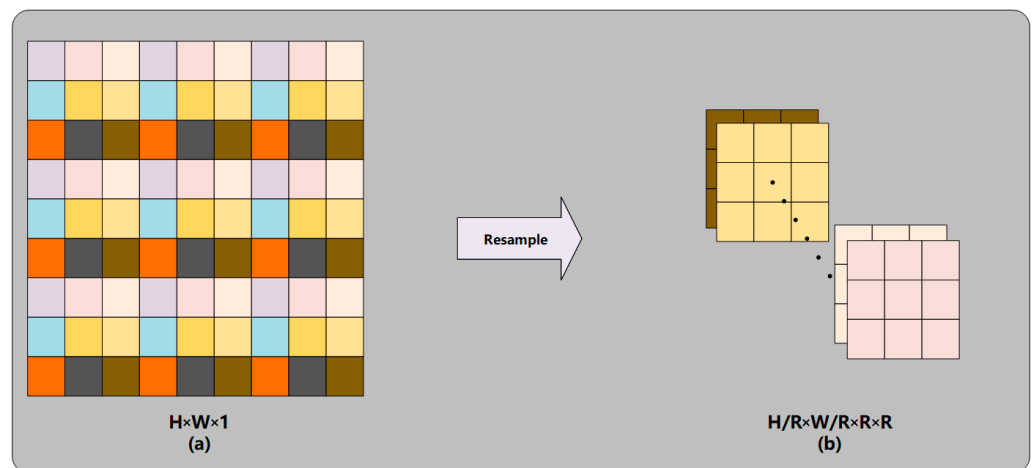


Figure 5. Resampling at  $R = 3$ . (The elements in (a) are uniformly distributed into the channels corresponding to the colors in (b).)

The outputs of the resample operation were set to  $S_0^R$  and  $S_0^{LD}$ . The corresponding inputs of each layer can be represented as  $S_{i,j}^R$  and  $S_{i,j}^{LD}$ , where  $i$  and  $j$  represent the  $i$ -th semi-coupled block and  $j$ -th kernel group in the  $i$ -th semi-coupled block, respectively. The corresponding features outputs of each group are denoted as  $\varphi_{i,j}^R$  and  $\varphi_{i,j}^{LD}$ . The feature size is  $(C, H, W)$ , where  $C$  represents the number of convolutional kernels within the kernel group. In Section 5, we determine the value of  $i$  and number of convolution kernels in each group during the ablation experiment. The correspondence between the output features and inputs is represented as

$$S_{i,j}^R(\varphi_{i,j-1}^R) = \varphi_{i,j-1}^R * \text{Concat}(K_{i,j}^{\text{shared}}, K_{i,j}^{\text{private}}), \tag{3}$$



$$S_{i,j}^{LD}(\varphi_{i,j-1}^{LD}) = \varphi_{i,j-1}^{LD} * \text{Concat}(K_{i,j}^{\text{shared}}, K_{i,j}^{\text{LD private}}), \tag{4}$$

where  $*$  represents the convolution operation;  $K_{i,j}^{\text{shared}}$ ,  $K_{i,j}^{\text{R private}}$ , and  $K_{i,j}^{\text{LD private}}$  represent the shared kernel corresponding to the  $j$ -th kernel group of the  $i$ -th semi-coupled block, private kernel for RGB feature extraction, and private kernel for depth map feature extraction, respectively.  $\text{Concat}$  represents the tensor-splicing operation in PyTorch. Based on Figure 4, the RGB feature output of each module can be represented as

$$\varphi_i^R = \text{ReLU}\{S_{i,3}^R(\text{ReLU}(S_{i,2}^R(\text{ReLU}(S_{i,1}^R(\varphi_{i-1}^R)))) + \varphi_{i-1}^R\}. \tag{5}$$

Equation (5) represents the same depth map output features. The CA attention module [28] was used to highlight the useful information in color features. Through this semi-coupled learning approach, the final output features include common and unique features in the cross-modal image pairs.

#### 4.2.2. Deformable Kernel for Guided Edge

In the previous section, RGB and depth features were obtained. Their corresponding unique features were determined using a semi-coupled feature extraction module. In order to suppress the redundant RGB features, a CA attention was used to highlight the information that improves up-sampling. Thereafter, a deformation convolution module was employed as a learnable filter. Consequently, deformable convolutional kernels can adapt to irregular image structures. Compared to traditional convolutional kernels with fixed shapes, they better capture textures and features in RGB images that help recover depth information. Additionally, using deformable convolutions in this context allows for the representation of a broader range of features with fewer parameters, thus reducing the model’s parameter count to some extent and enhancing computational efficiency. Finally, deformable convolutions excel in handling image boundaries, thereby minimizing the impact of boundary effects on recovering depth-discontinuities in GDSR tasks. Figure 6 shows the process described herein.

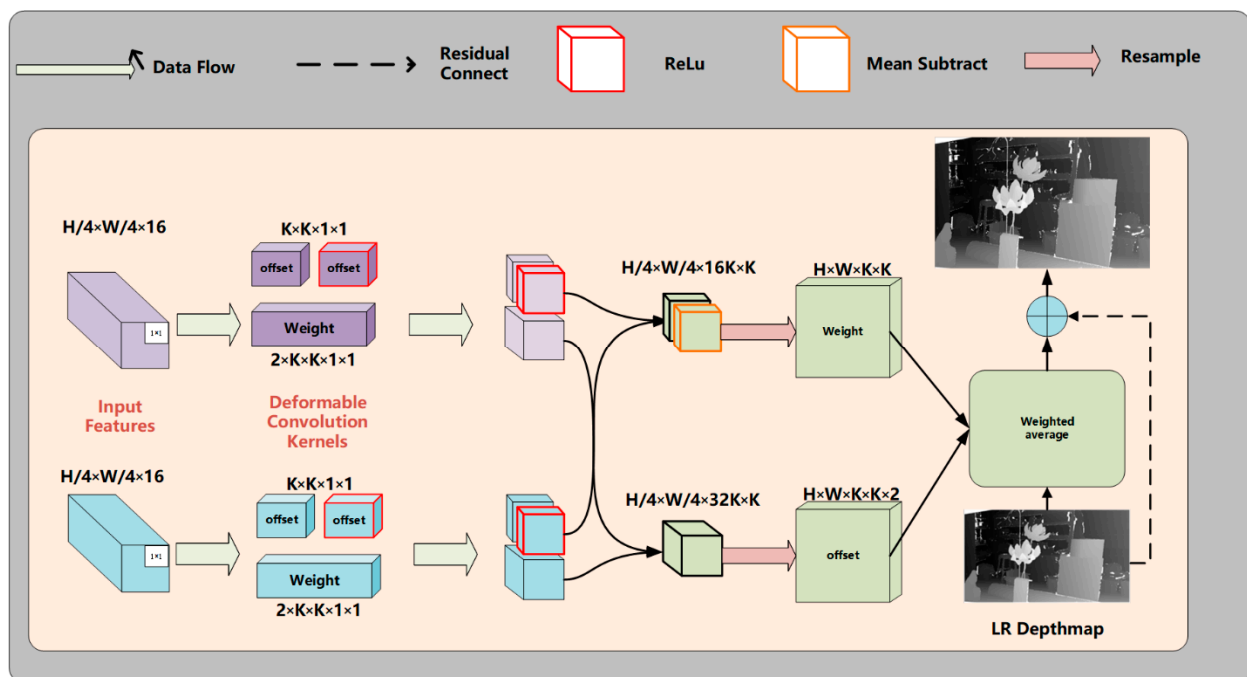


Figure 6. Deformable convolution module.

For general deformation convolution operations, the offsets should be learned along the  $x$ - and  $y$ -axis for each element of the feature map based on the following formula:

$$S(p_0) = \sum_{p_n \in \mathbb{R}} \omega(P_n) * \alpha(p_0 + P_n + (\Delta x + \Delta y)) \tag{6}$$

where  $p_0$  is the feature map point on which the convolution operation is performed;  $P_n$  is the standard offset of each point in the convolutional kernel relative to the center point;  $\omega(P_n)$  is the weight at the corresponding position of the convolutional kernel;  $\alpha(p_0 + P_n)$  is the element value at position  $p_0 + P_n$  on the input feature map;  $S(p_0)$  is the element value at position  $p_0$  on the output feature map;  $\Delta x$  and  $\Delta y$  denote the offsets of the additionally learned convolutional kernel elements along the  $x$ - and  $y$ -axis, respectively. This method outputs sets of neighbors and the corresponding weights adaptively for each pixel. The process is illustrated in Figure 7.

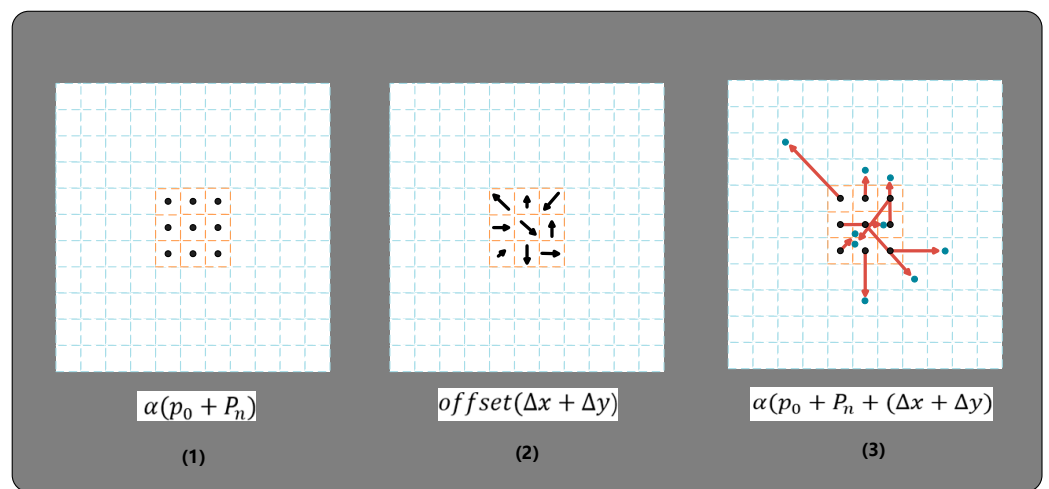


Figure 7. Deformable kernel.

After understanding the working principle of deformable convolution, we need to combine it with the GDSR task. To cover the entire feature map with the learned deformable convolutional kernel, a resample operation should be performed on the input feature map. To achieve this, we refer to the DKN approach [24].

$K \times K \times 1 \times 1$  convolutional kernels and  $2K \times K \times 1 \times 1$  convolutional kernels are used to process input features to learn the weights and offsets for each element of the assumed deformable convolutional kernel, respectively. Here,  $K$  represents the assumed deformable convolution kernel size. The two offset features are concatenated by multiplying the corresponding matrix points. The two weight features, after being activated using the ReLU function, are concatenated by multiplying the corresponding matrix points. The final weight is obtained by subtracting the mean values of RGB features and depth map features, respectively. Finally, the weight and offset features are combined to obtain the output, and the residual based on the LR depth images is calculated; then, the super-resolution depth map is obtained by adding it to the original low-resolution depth map.

Because most of the resulting offset elements were floating points, the sampled points were represented as subpixels. Therefore, bicubic interpolation was used to solve for the pixel values corresponding to the subpixel points. In Equation (6), the part representing the position is denoted as  $F_g$ , where  $j$  represents the sixteen points closest to subpixel point  $g$ , and  $f_d$  is the input target feature.  $B$  represents a bicubic interpolation function.

$$\alpha(p_0 + P_n + (\Delta x + \Delta y)) = F_g = \sum_{j \in R(g)} B(g, j) f_d \tag{7}$$

### 4.2.3. Training Loss

Our training model utilizes smooth Huber loss (L1-loss). In GDSR tasks, compared to L2-loss, L1-loss has been proven to be more robust in depth-discontinuities regions [24,49], and L1-loss can significantly suppress the influence of noise. The formula is as follows: the model output is set to  $D_H$ ; for a true HR depth map  $D_{gt}$ , we have

$$L(D_H, D_{gt}) = \sum_{i=1}^N |D_H - D_{gt}|_1. \quad (8)$$

## 5. Experiments

This section describes ablation experiments conducted with diverse datasets, showcasing the proposed model's effectiveness.

### 5.1. Setup

**Datasets:** The proposed model was trained based on recent studies [24,49]. We utilized an equal number of training samples to ensure fairness in the subsequent model comparison experiments. Of 1000 sheets selected from the NYUv2 dataset as the training set, 900 and 100 pairs were used for training and validation, respectively. The remaining 449 pairs were used as the test sets.

To verify its generalization ability, SCD-Net was tested using Middlebury and Lu datasets, the test set, and 100 pairs of RealSense data that did not participate in the training.

### 5.2. Experimental Details

The training samples were inputted using the original image ( $640 \times 480$ ). The depth images were normalized based on the maximum and minimum values in each depth image. For the color images, the standard deviation provided by ImageNet was used for normalization. The network was trained with 200 epochs, the mini-batch size was set to eight, and the Adam optimizer was employed at a learning rate of 0.0001. The proposed LRDM adhered to the DCT-Net [51] approach. HR images were down-sampled and restored to their original scale using bicubic interpolation. The processed image was used as the corresponding LRDM for the RGB image of the input model.

In the testing section, the conventional super-resolution tasks were emulated, and the root mean square error (RMSE) was used to evaluate the error between the GT and output images. The RMSE value was inversely proportional to the image quality. PyTorch was used to perform all the experiments, and an Nvidia GeForce RTX 3090 GPU was used for training and testing.

### 5.3. Comparison with Other Methods

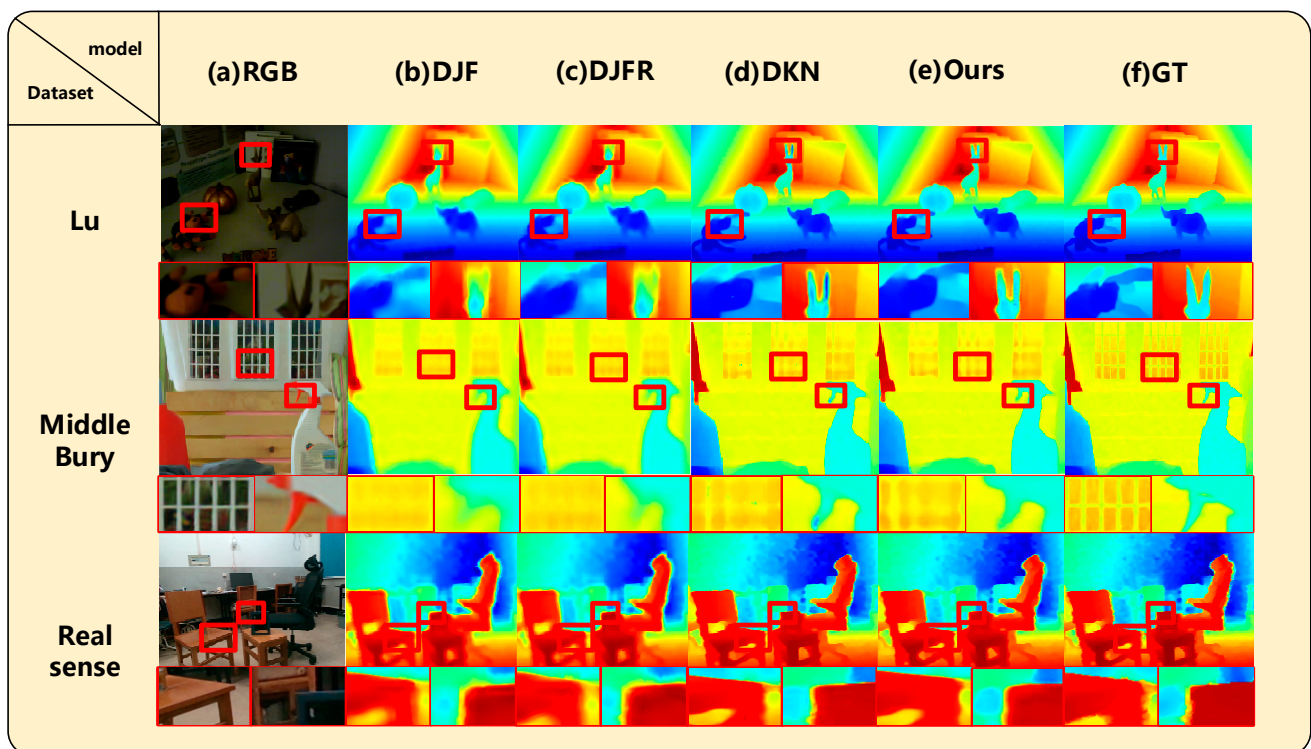
SCD-Net was evaluated on the NYUv2, Middlebury, Lu, and Real-sense datasets and compared to the conventional bicubic interpolation methods and current depth map super-resolution models, such as DJF [47], DJFR [54], pixel-adaptive convolutional neural network (PAC) [55], common and unique information splitting network (CUNet) [56], and DKN [24].

To verify the superiority of the proposed model, we used  $\times 4$ ,  $\times 8$ ,  $\times 16$ , and three amplification coefficients to display the output images of various models trained using the same dataset, i.e., NYUv2, as shown in Figure 6. Table 2 shows the RMSE comparison of these models after applying the super-resolution models on different test sets. Our method outperforms traditional models significantly at the most demanding  $\times 8$  and  $\times 16$  scales, and it also surpasses recent state-of-the-art models. Compared to the parameter-heavy CUNet and PAC methods, our model achieves optimal or near-optimal results across the three datasets, demonstrating the robustness of our approach.

**Table 2.** Quantitative comparison between the proposed SCD-Net and existing state-of-the-art approaches on the three benchmark datasets (Bold indicates the best RMSE result).

Methods	Middlebury			NYU v2			Lu		
	×4	×8	×16	×4	×8	×16	×4	×8	×16
Bicubic	4.44	7.58	11.87	8.16	14.22	22.32	5.07	9.22	14.27
DJF [47]	1.68	3.24	5.62	2.80	5.33	9.46	1.65	3.96	6.75
DJFR [54]	1.32	3.19	5.57	2.38	4.94	9.18	1.15	3.57	6.77
PAC [55]	1.32	2.62	4.58	1.89	<b>3.33</b>	<b>6.78</b>	1.20	2.33	5.19
CUNet [56]	1.10	2.17	4.33	1.92	3.70	<b>6.78</b>	0.91	2.23	4.99
DKN [24]	<b>1.08</b>	2.17	4.50	1.86	3.58	6.96	<b>0.82</b>	2.10	5.05
ours	1.13	<b>2.13</b>	<b>4.39</b>	<b>1.68</b>	3.45	6.88	0.86	<b>1.92</b>	<b>4.88</b>

In order to verify the generalization ability of the model for various datasets, five ×8 scale restoration models trained on the NYUv2 dataset were selected, as shown in Figure 8. The super-resolution was performed on the Lu, Middlebury, and Real-sense images obtained using bicubic interpolation. The complicated parts of the GDSR task are marked and enlarged in red boxes. Compared to DJF and DJFR, the proposed model can restore deep discontinuous edges in depth maps. Compared to DKN, which can also recover depth discontinuities, the proposed model suppresses the RGB texture error migration (the image restored using the DKN model contains erroneous depth points). It is evident from this that the semi-coupled feature extraction module we have proposed, in combination with the CA attention-based feature processing approach, contributes to improving the model’s robustness.



**Figure 8.** Visual comparison of ×8 depth map SR results without NYUv2 dataset. The depth map processing was performed using JET color bar. (a) RGB, (b) DJF [47], (c) DJFR [54], (d) DKN [24], (e) the proposed SCD-Net, and (f) GT images. (The area marked with red squares is enlarged for display).

#### 5.4. Ablation Study

In order to determine the structure of the proposed semi-coupled feature extraction method, ablation experiments were performed on the number of convolutional kernels in each layer and that of residual layers. The proposed model was uniformly trained using the NYUv2 dataset. We defined  $i$  and  $j$  as the number of convolutional kernels within a layer and that of semi-coupled layers, respectively. The results are summarized in Table 3.

**Table 3.** Results of ablation experiments on the NYUv2 test set. Bold figures indicate the best score in terms of RMSE.

Number of Kernel-Group $i$ ( $j = 3$ )					
Setting	8	16	32	64	128
$\times 4$	1.9146	1.8432	1.8247	1.6816	<b>1.6729</b>
$\times 8$	4.1318	3.8965	3.8424	3.4579	<b>3.4413</b>
$\times 16$	8.3753	7.8421	7.6025	<b>6.8872</b>	6.9037
Number of Kernel-Group $j$ ( $i = 64$ )					
Setting	2	3	4	5	6
$\times 4$	2.0121	1.6816	1.6732	<b>1.6553</b>	1.6599
$\times 8$	3.8937	<b>3.4579</b>	3.4599	3.4603	3.4623
$\times 16$	7.5524	6.8872	6.8551	6.8324	<b>6.8121</b>

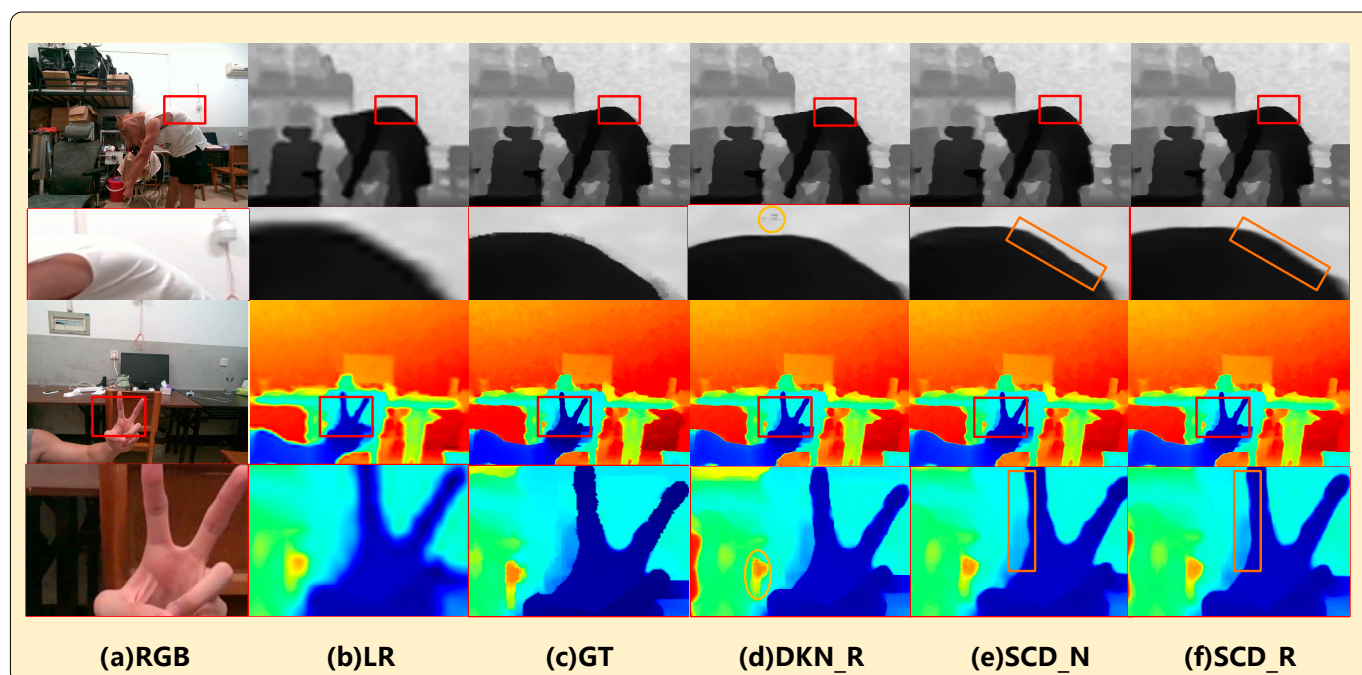
Based on Table 3, the accuracy of the generated depth map exhibits an upward trend at higher convolutional kernels and a greater number of semi-coupling layers. However, a small accuracy improvement significantly increases the parameter. (For example, when  $j = 4$ , increasing the size of  $i$  from 64 to 128 causes the model's parameter to change from 714,608 to 2,789,992. However, the improvement in performance is marginal.) The speed and accuracy of the model were weighed, and a  $64 \times 4$  semi-coupled feature extraction structure was selected.

In order to verify the effectiveness of the proposed dataset, the proposed model was retrained using the Real-sense dataset. Before training, the dataset format was preprocessed to ensure consistency with NYUv2. The training parameter settings were consistent with those described in Section 4.2. The results are presented in Table 4 and Figure 8. The performance of the model trained using the proposed dataset improved significantly, and the overall RMSE and accuracy of the object edge depth values increased.

**Table 4.** Quantitative depth map SR results on the Real-sense dataset. SCD\_R is SCD-Net trained on the Real-sense dataset. (Bold indicates the best RMSE result).

RMSE	DJF	DJFR	DKN	SCD	SCD_R
$\times 4$	2.66	2.23	1.75	1.58	<b>1.47</b>
$\times 8$	5.22	4.61	3.35	3.22	<b>3.01</b>
$\times 16$	9.11	8.89	6.56	6.34	<b>6.03</b>

As shown in Figure 9, we selected a pair of human-pose depth data that did not participate in the training to test the generalization ability of all models for natural scenes. Among them,  $\_N$  and  $\_R$  represent the models trained using the NYU and Real-sense datasets, respectively. Figure 9e indicates a slight texture error migration, and the target edge is not sufficiently flat in the area marked by the orange box. After retraining the proposed model using the Real-sense dataset, super-resolution was performed on the LR image, which returned typical depth values in this area.



**Figure 9.** Visual comparison of  $\times 8$  depth map SR results for Real-sense. (a) RGB image, (b) lower-resolution depth map, (c) ground truth, (d) DKN [23] training on Real-sense dataset, (e) SCD-net training on NYUv2 dataset, (f) SCD-net training on Real-sense dataset. (The area marked with red squares is enlarged for display).

## 6. Conclusions

This study proposes SCD-Net with semi-coupled feature extraction and deformable convolution to overcome the limitation of degradation in depth discontinuities when using traditional interpolation methods for depth map super-resolution and RGB texture over-transfer in GDSR tasks. Furthermore, we addressed the issue of limited datasets and poor quality of the current depth-map super-resolution tasks. However, the deformable convolution module needs a resampled input, which makes it difficult to migrate it to other models. At the same time, fully convolutional neural networks find it difficult to interpret the working mechanism. For GDSR datasets, existing methods are unable to establish a genuine correspondence between HR depth maps and LRDMs. This is crucial for real-world GDSR tasks. In future studies, we will improve the deformable convolution module structure to allow it to participate flexibly in other models. We will further shift the GDSR task from the spatial domain to the frequency domain to replace the fully convolutional network and enhance the interpretability of the model. To further enhance the quality of our dataset, we will delve into the principles of structured-light camera imaging. We plan to construct a collection system that combines low-resolution and high-resolution structured-light cameras in order to achieve a genuine correspondence between HR depth maps and LRDMs.

**Author Contributions:** Investigation, K.C.; resources, S.-L.P.; supervision, M.Z. and B.L.; funding acquisition, M.Z.; writing—original draft preparation, K.C.; writing—review and editing, M.Z., B.L., S.-L.P. and K.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the China University Industry, University and Research Innovation Fund of the Ministry of Education (grant number 2022 IT 036).

**Data Availability Statement:** Data sharing not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Chatzopoulos, D.; Bermejo, C.; Huang, Z.; Hui, P. Mobile augmented reality survey: From where we are to where we go. *IEEE Access* **2017**, *5*, 6917–6950. [[CrossRef](#)]
2. Shotton, J.; Girshick, R.; Fitzgibbon, A.; Sharp, T.; Cook, M.; Finocchio, M.; Moore, R.; Kohli, P.; Criminisi, A.; Kipman, A.; et al. Efficient human pose estimation from single depth images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 2821–2840. [[CrossRef](#)] [[PubMed](#)]
3. Rasouli, A.; Tsotsos, J.K. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 900–918. [[CrossRef](#)]
4. DeSouza, G.N.; Kak, A.C. Vision for mobile robot navigation: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 237–267. [[CrossRef](#)]
5. He, K.; Sun, J.; Tang, X. Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1397–1409. [[CrossRef](#)]
6. Riegler, G.; R  ther, M.; Bischof, H. Atgv-net: Accurate depth super-resolution. In Proceedings of the Computer Vision—ECCV 2016 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part III 14. Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 268–284.
7. Liu, B.; Chen, K.; Peng, S.L.; Zhao, M. Adaptive Aggregate Stereo Matching Network with Depth Map Super-Resolution. *Sensors* **2022**, *22*, 4548. [[CrossRef](#)]
8. Liu, M.Y.; Tuzel, O.; Taguchi, Y. Joint geodesic upsampling of depth images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 169–176.
9. Min, D.; Lu, J.; Do, M.N. Depth video enhancement based on weighted mode filtering. *IEEE Trans. Image Process.* **2011**, *21*, 1176–1190. [[PubMed](#)]
10. Lu, J.; Forsyth, D. Sparse depth super resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2245–2253.
11. Lu, J.; Shi, K.; Min, D.; Lin, L.; Do, M.N. Cross-based local multipoint filtering. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: New York, NY, USA, 2012; pp. 430–437.
12. Diebel, J.; Thrun, S. An application of markov random fields to range sensing. In *Advances in Neural Information Processing Systems, Proceedings of the 18th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 5–8 December 2005*; MIT Press: Cambridge, MA, USA, 2005; pp. 291–298.
13. Li, Y.; Min, D.; Do, M.N.; Lu, J. Fast guided global interpolation for depth and motion. In Proceedings of the Computer Vision—ECCV 2016 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part III 14. Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 717–733.
14. Ferstl, D.; Reinbacher, C.; Ranftl, R.; R  ther, M.; Bischof, H. Image guided depth upsampling using anisotropic total generalized variation. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 993–1000.
15. Park, J.; Kim, H.; Tai, Y.W.; Brown, M.S.; Kweon, I. High quality depth map upsampling for 3D-TOF cameras. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; IEEE: New York, NY, USA, 2011; pp. 1623–1630.
16. Yang, J.; Ye, X.; Li, K.; Hou, C.; Wang, Y. Color-guided depth recovery from RGB-D data using an adaptive autoregressive model. *IEEE Trans. Image Process.* **2014**, *23*, 3443–3458. [[CrossRef](#)]
17. Xie, J.; Feris, R.S.; Yu, S.S.; Sun, M.T. Joint super resolution and denoising from a single depth image. *IEEE Trans. Multimed.* **2015**, *17*, 1525–1537. [[CrossRef](#)]
18. Xie, J.; Feris, R.S.; Sun, M.T. Edge-guided single depth image super resolution. *IEEE Trans. Image Process.* **2015**, *25*, 428–438. [[CrossRef](#)] [[PubMed](#)]
19. Gu, S.; Zuo, W.; Guo, S.; Chen, Y.; Chen, C.; Zhang, L. Learning dynamic guidance for depth image enhancement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3769–3778.
20. Kiechle, M.; Hawe, S.; Kleinsteuber, M. A joint intensity and depth co-sparse analysis model for depth map super-resolution. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1545–1552.
21. Guo, C.; Li, C.; Guo, J.; Cong, R.; Fu, H.; Han, P. Hierarchical features driven residual learning for depth map super-resolution. *IEEE Trans. Image Process.* **2018**, *28*, 2545–2557. [[CrossRef](#)] [[PubMed](#)]
22. Hui, T.W.; Loy, C.C.; Tang, X. Depth map super-resolution by deep multi-scale guidance. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 353–369.
23. Ye, X.; Sun, B.; Wang, Z.; Yang, J.; Xu, R.; Li, H.; Li, B. PMBANet: Progressive multi-branch aggregation network for scene depth super-resolution. *IEEE Trans. Image Process.* **2020**, *29*, 7427–7442. [[CrossRef](#)]
24. Kim, B.; Ponce, J.; Ham, B. Deformable kernel networks for joint image filtering. *Int. J. Comput. Vis.* **2021**, *129*, 579–600. [[CrossRef](#)]
25. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from RGBD images. In Proceedings of the Computer Vision—ECCV 2012 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Proceedings, Part V 12. Springer Berlin Heidelberg: Berlin/Heidelberg, Germany, 2012; pp. 746–760.
26. Tang, J.; Chen, X.; Zeng, G. Joint implicit image function for guided depth super-resolution. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 4390–4399.

27. Nguyen, H.T.; Worring, M.; Van Den Boomgaard, R. Watersnakes: Energy-driven watershed segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 330–342. [\[CrossRef\]](#)
28. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
29. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
30. Hirschmuller, H.; Scharstein, D. Evaluation of cost functions for stereo matching. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; IEEE: New York, NY, USA, 2007; pp. 1–8.
31. Lu, S.; Ren, X.; Liu, F. Depth enhancement via low-rank matrix completion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3390–3397.
32. Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4040–4048.
33. Butler, D.J.; Wulff, J.; Stanley, G.B.; Black, M.J. A naturalistic open source movie for optical flow evaluation. In Proceedings of the Computer Vision–ECCV 2012 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Proceedings, Part VI 12. Springer Berlin Heidelberg: Berlin/Heidelberg, Germany, 2012; pp. 611–625.
34. Peris, M.; Martull, S.; Maki, A.; Ohkawa, Y.; Fukui, K. Towards a simulation driven stereo vision system. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; IEEE: New York, NY, USA, 2012; pp. 1038–1042.
35. Kopf, J.; Cohen, M.F.; Lischinski, D.; Uyttendaele, M. Joint bilateral upsampling. *ACM Trans. Graph. (ToG)* **2007**, *26*, 96-es. [\[CrossRef\]](#)
36. Tomasi, C.; Manduchi, R. Bilateral filtering for gray and color images. In Proceedings of the Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), Bombay, India, 7 January 1998; IEEE: New York, NY, USA, 1998; pp. 839–846.
37. Lo, K.H.; Wang, Y.C.; Hua, K.L. Joint trilateral filtering for depth map super-resolution. In Proceedings of the 2013 Visual Communications and Image Processing (VCIP), Kuching, Malaysia, 17–20 November 2013; IEEE: New York, NY, USA, 2013; pp. 1–6.
38. Li, Y.; Xue, T.; Sun, L.; Liu, J. Joint example-based depth map super-resolution. In Proceedings of the 2012 IEEE International Conference on Multimedia and Expo, Melbourne, VIC Australia, 9–13 July 2012; IEEE: New York, NY, USA, 2012; pp. 152–157.
39. Kasetkasem, T.; Arora, M.K.; Varshney, P.K. Super-resolution land cover mapping using a Markov random field based approach. *Remote Sens. Environ.* **2005**, *96*, 302–314. [\[CrossRef\]](#)
40. Strong, D.; Chan, T. Edge-preserving and scale-dependent properties of total variation regularization. *Inverse Probl.* **2003**, *19*, S165. [\[CrossRef\]](#)
41. Saputro, D.R.S.; Widyaningsih, P. Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method for the parameter estimation on geographically weighted ordinal logistic regression model (GWOLR). In Proceedings of the AIP Conference Proceedings, Yogyakarta, Indonesia, 15–16 May 2017; AIP Publishing: Melville, NY, USA, 2017; p. 1868.
42. Bi, H.; Zhang, B.; Zhu, X.X.; Hong, W.; Sun, J.; Wu, Y.  $L_1$ -regularization-based SAR imaging and CFAR detection via complex approximated message passing. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3426–3440. [\[CrossRef\]](#)
43. Lee, D.; Seung, H.S. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems, Proceedings of the 13th International Conference on Neural Information Processing Systems, Denver, CO, USA, 1 January 2000*; MIT Press: Cambridge, MA, USA, 2000; pp. 556–562.
44. Tosić, I.; Olshausen, B.A.; Culpepper, B.J. Learning sparse representations of depth. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 941–952. [\[CrossRef\]](#)
45. Zhang, K.; Gao, X.; Tao, D.; Li, X. Multi-scale dictionary for single image super-resolution. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: New York, NY, USA, 2012; pp. 1114–1121.
46. Wang, S.; Zhang, L.; Liang, Y.; Pan, Q. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: New York, NY, USA, 2012; pp. 2216–2223.
47. Li, Y.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep joint image filtering. In Proceedings of the Computer Vision–ECCV 2016 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part IV 14. Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 154–169.
48. Tang, Q.; Cong, R.; Sheng, R.; He, L.; Zhang, D.; Zhao, Y.; Kwong, S. Bridgenet: A joint learning network of depth map super-resolution and monocular depth estimation. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 2148–2157.
49. He, L.; Zhu, H.; Li, F.; Bai, H.; Cong, R.; Zhang, C.; Lin, C.; Liu, M.; Zhao, Y. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9229–9238.
50. Wu, H.; Zheng, S.; Zhang, J.; Huang, K. Fast end-to-end trainable guided filter. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1838–1847.



51. Zhao, Z.; Zhang, J.; Xu, S.; Lin, Z.; Pfister, H. Discrete cosine transform network for guided depth map super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5697–5707.
52. Levin, A.; Lischinski, D.; Weiss, Y. Colorization using optimization. In *ACM SIGGRAPH 2004 Papers*; Association for Computing Machinery: New York, NY, USA, 2004; pp. 689–694.
53. Jeon, J.; Lee, S. Reconstruction-based pairwise depth dataset for depth image enhancement using CNN. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 422–438.
54. Li, Y.; Huang, J.B.; Ahuja, N.; Yang, M.H. Joint image filtering with deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1909–1923. [[CrossRef](#)]
55. Su, H.; Jampani, V.; Sun, D.; Gallo, O.; Learned-Miller, E.; Kautz, J. Pixel-adaptive convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11166–11175.
56. Deng, X.; Dragotti, P.L. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3333–3348. [[CrossRef](#)] [[PubMed](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.