Guided Depth Map Super-Resolution: A Survey

ZHIWEI ZHONG, XIANMING LIU, JUNJUN JIANG, and DEBIN ZHAO, Harbin Institute of Technology
XIANGYANG JI, Tsinghua University

Guided depth map super-resolution (GDSR), which aims to reconstruct a high-resolution depth map from a low-resolution observation with the help of a paired high-resolution color image, is a longstanding and fundamental problem that has attracted considerable attention from computer vision and image processing communities. Myriad novel and effective approaches have been proposed recently, especially with powerful deep learning techniques. This survey is an effort to present a comprehensive survey of recent progress in GDSR. We start by summarizing the problem of GDSR and explaining why it is challenging. Next, we introduce some commonly used datasets and image quality assessment methods. In addition, we roughly classify existing GDSR methods into three categories: filtering-based methods, prior-based methods, and learning-based methods. In each category, we introduce the general description of the published algorithms and design principles, summarize the representative methods, and discuss their highlights and limitations. Moreover, depth-related applications are introduced. Furthermore, we conduct experiments to evaluate the performance of some representative methods based on unified experimental configurations, so as to offer a systematic and fair performance evaluation to readers. Finally, weconclude this survey with possible directions and open problems for further research. All related materials can be found at https://github.com/zhwzhong/Guided-Depth-Map-Super-resolution-A-Survey.

CCS Concepts: • **General and reference → Surveys and overviews;**

Additional Key Words and Phrases: Guided depth map super-resolution, survey, filtering, prior, learning

ACM Reference format:
Zhiwei Zhong, Xianming Liu, Junjun Jiang, Debin Zhao, and Xiangyang Ji. 2023. Guided Depth Map Super-Resolution: A Survey. *ACM Comput. Surv.* 55, 14s, Article 301 (July 2023), 36 pages.
https://doi.org/10.1145/3584860

1 INTRODUCTION

Depth information plays a fundamental role in a wide range of computer vision applications, such as 3D reconstruction [43, 87], semantic scene understanding [118, 124, 134], and autonomous driving [28, 122]. Currently, there are two main range sensing technologies to capture depth information, either passive or active. For passive sensing technologies, the most famous and commonly

This work was supported by the National Natural Science Foundation of China under grants 92270116 and 62071155.
Authors' addresses: Z. Zhong, X. Liu (corresponding author), J. Jiang, and D. Zhao, School of Computer Science and Technology, Harbin Institute of Technology, Xidazhi Street, Harbin 150001, China; emails: {zhwzhong, csxm, jiangjunjun, dbzhao}@hit.edu.cn; X. Ji, Department of Automation and BNRist, Tsinghua University, Qinghuayuan Street, Beijing 100084, China; email: xyji@tsinghua.edu.cn.

0360-0300/2023/07-ART301 $15.00
https://doi.org/10.1145/3584860

used approach is stereo reconstruction, which first obtains two images of the scene through two conventional monochrome or color cameras separated by a certain distance and then estimates the scene depth by using stereo matching algorithms. Over the past decades, the performance of stereo matching algorithms has improved markedly [99, 119]. However, as these methods are based on finding corresponds between the two views, the effect of depth missing and occlusion will occur for texture-less regions. In contrast to passive ones, active methods can capture depth maps more efficiently and robustly, especially for texture-less regions. The **Time-of-Flight (ToF)** camera [64] and structure light scanner [50] are two representative depth sensors for real-time active depth sensing. Despite the rapid development of active depth sensing, there is still a large gap between depth camera and color camera, especially in terms of spatial resolution. For instance, the spatial resolution of consumer-level ToF cameras is only about 200 kilopixels, which is much lower than off-the-shelf conventional color cameras (e.g., 40 megapixels for Huawei Mate 30).

The **Low-Resolution (LR)** depth maps acquired by these depth sensors would greatly affect the performance of algorithms in downstream applications. Therefore, it is highly desirable to develop effective and efficient depth super-resolution algorithms to facilitate the usage of depth information. In general, depth cameras are equipped with an additional RGB sensor that can capture color images with a resolution higher than that of depth maps. It is natural to improve the quality of the depth map by employing the **High-Resolution (HR)** color counterpart as guidance, which is known as **Guided Depth Map Super-Resolution (GDSR)**. Compared to the general image super-resolution, the guided depth image super-resolution has two unique characteristics: (1) the depth map contains smooth areas separated by sharp boundaries, which is known as **Piecewise Smooth (PWS)**, and (2) an additional HR color image can be used to guide the depth map reconstruction process. Generally, the depth map and its corresponding color image are geometric and photometric descriptions of the same scene, and they have strong structural similarity. Thus, most existing GDSR algorithms are elaborately designed to exploit this piecewise smoothness and take full advantage of the guidance image. Recently, GDSR has attracted more and more attention in both academic and industrial circles, which encourages us to offer this comprehensive survey of recent advances.

Fundamentally, GDSR is an ill-posed problem since there can be various HR depth maps with a slight difference in camera angle, illumination, material properties, and other variables for a certain LR depth map. During the past decades, a wide variety of methods have been developed to solve this inverse problem. To clearly describe the progress of the GDSR methods, we provide a brief chronology in Figure 1. Here, we classify these methods into the following three categories: filtering-based methods, optimization-based methods, and learning-based methods.

Filtering-based methods estimate the depth of pixels by performing a weighted average of local pixels, and the weights are obtained by the affinity calculated from RGB-D image pairs. The **Bilateral Filters (BFs)** [62, 88, 114, 139], **Non-Local Means (NL-Means)** filters [6, 53], and **Guided Filters (GFs)** [44, 72, 110, 130] are representative methods. Filtering-based methods enjoy simplicity in design and low computational complexity, but they are prone to generate annoying artifacts when the depth discontinuities are not consistent with those in their corresponding color image. Moreover, the filter kernel is typically designed for a specific task, which lacks flexibility. Contrary to local filtering-based methods, optimization-based methods formulate the problem of GDSR as a global optimization framework. The objective function usually contains two terms: the data fidelity term and regularization term. The data fidelity term is used to preserve depth consistency between the super-resolved result and the input depth map. The regularization term encourages the edges of the super-resolved result to consist with the guidance image as much as possible. The shortcomings of local filtering-based methods can be resolved to some extent by optimization-based methods. Nevertheless, since the optimization-based methods depend on

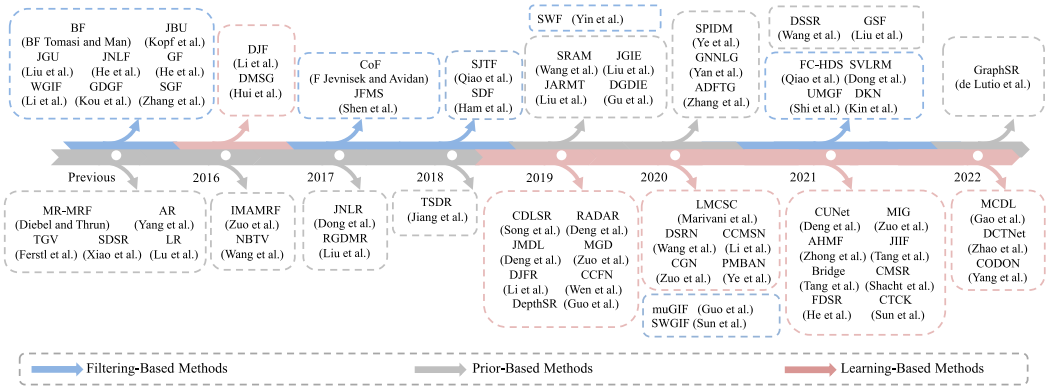


Fig. 1. Milestones of GDSR methods, including filter-based methods, prior-based methods, and learning-based methods.

handcrafted object functions, they may not reflect the real complex image priors. Furthermore, the iterative optimization process involved in these algorithms is usually time consuming, which impedes their applications in real-world scenarios. To obtain a better and more efficient GDSR model, learning-based methods are proposed and applied to large-scale datasets to solve the depth map super-resolution task. The learning-based methods can be further divided into two categories: sparse dictionary learning methods and deep learning methods. Dictionary learning, also known as sparse coding, aims at deriving a set of dictionary atoms in which only a small number of atoms can be linearly combined to approximate a given image. Due to its simplicity and flexibility for data representation, numerous methods have been introduced in the past decade and achieved impressive results [18, 19, 36]. Recent advances of deep learning algorithms have revolutionized the area of computer vision; tremendous progress has been achieved in a variety of domains, such as image classification [27, 46], object localization [131, 147], and semantic segmentation [108, 151]. GDSR is no exception: myriad neural networks have been developed and advanced the state of the art [17, 154, 156]. The learning-based method is suitable for a scene where there is a large amount of training data. When training data are insufficient, the traditional optimization-based or filter-based method may be a better choice.

Although significant progress has been made, there is no unified framework to understand and classify existing methods, particularly the recently popular deep learning based approaches. To the best of our knowledge, there are only two surveys [30, 137] available in the literature for the area of GDSR. The work of Eichhardt et al. [30] was published in 2017, with a main purpose to summarize the methods that are tailored for ToF depth maps. The work of Wang et al. [137] was published in 2019, which only introduced some representative methods published before 2019. Many advanced methods that achieve state-of-the-art performance have been proposed since then. Furthermore, neither of these two surveys introduces the benchmark datasets, evaluation metrics, or experimental comparisons of different methods. A comprehensive survey is desired that summarizes and compares these approaches with a unified perspective. In addition, inconsistent experimental protocols are utilized in different approaches. The task of GDSR lacks a benchmark for fairly evaluating state-of-the-art approaches. In view of these issues, we provide a comprehensive review of existing GDSR methods.

We summarize the key contributions of this survey as follows:

- We present a systematic investigation of recent GDSR techniques, including problem formulation, benchmark dataset, widely used evaluation metrics, and representative

methods. To introduce these methods clearly, we classify them into three categories: filtering based, optimization based, and learning based. For each category, we review representative methods and discuss their contributions, benefits, and weaknesses.
- We provide comprehensive and fair experimental comparison to evaluate state-of-the-art methods quantitatively and qualitatively on benchmark datasets based on the same experimental configurations. Moreover, we present some tricks that can be used to further improve the performance of depth map reconstruction.
- We discuss the open problems and challenges, and we envision prospects for future research.

The rest of this article is organized as follows. In Section 2, we introduce some background concepts and basic knowledge, including the formulation of the GDSR problem, widely used RGB-D datasets, and evaluation metrics. Then, detailed reviews of filtering-based (Section 3), optimization-based (Section 4), and learning-based (Section 5) methods are given. In Section 6, we introduce the depth-related applications. In Section 7, we systematically compare several representative methods using publicly available datasets with unified experimental configurations. Finally, we conclude the article and present some challenging issues as well as promising future research directions in Section 8.

2 FUNDAMENTALS

In this section, we introduce the problem formulation of GDSR and commonly used datasets and evaluation metrics.

2.1 Problem Formulation

Current range sensing systems are far from perfect due to a few technical limitations, thus a depth map acquired by these systems is a degraded version of the underlying ground truth depth map. Let $\boldsymbol{y}$ and $\boldsymbol{x}$ denote the acquired LR and underlying HR depth maps, respectively. The observation model in GDSR can be described as follows:

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{H}$ denotes the observation matrix and $\boldsymbol{n}$ is the introduced additive noise. For different sensing systems, the observation matrix can take different forms. For example, the depth maps captured by the Kinect camera usually experience both structure missing and random depth pixels missing, whereas the depth maps captured by the ToF camera might suffer both LR and noisy degradations. It is worth noting that in the field of color image processing, a similar formulation has been extensively studied. GDSR differs from them because of the existence of the guidance images. How to take full advantage of the guidance image to improve the performance of GDSR presents new opportunities and challenges to researchers.

To make the mathematical formulations clear, we summarize the notations used in this article and list them in Table 1. Here, we use normal case letters to represent the scalars, uppercase bold letters to represent the matrices, and lowercase bold letters to represent the vectors.

2.2 Datasets for GDSR

With the rapid development of GDSR algorithms, a great number of RGB-D datasets have been constructed in the past few decades. We describe 10 RGB-D datasets that are widely used in the literature in Table 2. The detailed descriptions of each dataset are as follows:

Middlebury dataset [51, 95, 97]: Middlebury is a dataset of indoor scenes, which was the most widely used dataset before the deep learning technique was introduced to the field of GDSR. It consists of three small-scale datasets: 2003 [97], 2005 [51, 95], and 2006 [51, 95]. As these

Table 1. Summary of Commonly Used Notations in This Article

Notation	Description	Notation	Description
$\boldsymbol{X} \in \mathbb{R}^{H \times W}$	The GT depth map	$\hat{\boldsymbol{Y}} \in \mathbb{R}^{H \times W}$	The interpolated LR depth map
$\boldsymbol{Y} \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s}}$	The LR depth map	$\boldsymbol{Y}^* \in \mathbb{R}^{H \times W}$	The super-resolved depth map
$\boldsymbol{G} \in \mathbb{R}^{H \times W}$	The guidance image	$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{g}$	The vectorized form of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{G}$, respectively

Table 2. Summary of Widely Used RGB-D Datasets for GDSR

Middlebury [51, 95, 97]	–	Indoor	-	Stereo Camera	32
ToFmark [35]	2013	Indoor	Time of Flight	PMD Nano	3
NYU v2 [104]	2012	Indoor	Structure Light	Kinect V1	1,449
MPI Sintel Depth [7]	2012	Indoor, Outdoor	Animated Film	Blender	1,064
Lu [84]	2014	Indoor	Structure Light	ASUS Xtion Pro	6
SUN RGB-D [106]	2015	Indoor	Time of Flight	Kinect v2	2,860
DIDOE [117]	2019	Indoor, Outdoor	Laser scanner	FARO Focus S350	27,858
DIML Indoor [12]	2021	Indoor	Time of Flight	Kinect v2	2,000
DIML Outdoor [12]	2021	Outdoor	Stereo Matching	ZED Stereo	2,000
RGB-D-D [47]	2021	Indoor, Outdoor	Time of Flight	Huawei P30 Pro, Helios	4,811

datasets do not provide depth maps, previous methods adopted disparity data as depth maps for evaluation.

ToFMark dataset [35]: This dataset consists of three RGB-D image pairs (i.e., *Books, Shark, Devil*) that are captured from real-world scenes. The LR depth maps are captured by a PMD Nano ToF camera, whereas the HR depth maps are first acquired by using a structured light scanner that is equipped with a high-speed projector. The final ground truth depth maps are then obtained by combining multiple acquisitions with slightly displaced projection angles.

NYU v2 dataset [104]: NYU v2 is an indoor dataset collected by a Microsoft Kniect v1 sensor. This dataset is a continuation of NYU v1 [103], which utilizes the same sensor and data structure but contains fewer scenes and image pairs.

MPI Sintel Depth dataset [7]: This dataset is generated from the open source 3D animated short film *Sintel*. MPI Sintel Depth contains more complex and realistic scenes, such as fog, large motions, and blur. It includes 1,064 image pairs used for the evaluation of the GDSR algorithm.

Lu dataset [84]: This dataset contains six representative RGB-D image pairs that are obtained by the ASUS Xtion Pro camera, and several GDSR methods use it as the test set.

SUN-RGB-D dataset [106]: This dataset is widely used for scene understanding tasks. It contains 10,355 RGB-D image pairs for training and 2,860 RGB-D image pairs for testing. The test RGB-D pairs are obtained by Microsoft Kinect v2.

DIDOE dataset [117]: DIDOE is the first public dataset that captures both indoor and outdoor RGB-D image pairs with the same camera. It contains HR RGB images with precise and far-range depth measurements.

DIML dataset [12, 59–61]: DIML is a newly collected large-scale RGB-D dataset that consists of more than 200 indoor/outdoor scenes. For indoor scenes, the authors capture the depth maps using the Kinect v2 sensor. For outdoor scenes, they first use an accurate stereo matching method to generate disparity maps and then convert them to depth maps using the calibration parameters.

RGB-D-D dataset [47]: This dataset consists of 4,811 image pairs captured from both indoor and outdoor scenes. The "D-D"represents the paired LR and HR depth maps acquired by the Huawei P30 Pro and the Lucid Helios camera.

2.3 Assessment Metrics for GDSR

For GDSR, quantitative comparison is an essential part of evaluating the performance of different methods. Two of the most frequently used metrics are **Mean Absolute Error (MAE)**, **Mean Square Error (MSE)**, and **Root Mean Square Error (RMSE)**, which are defined as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\boldsymbol{X}_i - \boldsymbol{Y}_i^*|, \quad \text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{X}_i - \boldsymbol{Y}_i^*)^2, \quad \text{RMSE} = \sqrt{\text{MSE}}, \tag{2}$$

where $\boldsymbol{X}_i$ means the i-th pixel value of the ground truth depth map, $\boldsymbol{Y}_i^*$ represents the i-th pixel value of the super-resolved depth map, and N denotes the total number of pixels. Some minority measurements, such as the **Structural Similarity Index Measure (SSIM)** and **Peak Signal-to-Noise Ratio (PSNR)** are also sometimes utilized. The PSNR can be calculated as follows:

$$\text{PSNR} = 10 \cdot \log_{10}(\text{MAX}^2/\text{MSE}), \tag{3}$$

where MAX is the maximum pixel value of the depth map. SSIM [126] is a traditional indicator of image quality assessment that employs a combination of structures, luminance, and contrast to calculate similarity:

$$\text{SSIM} = \frac{\left(2\mu_x\mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)}, \tag{4}$$

where μ_x and μ_y represent the means of the ground truth and the super-revolved depth maps, respectively; σ_x^2 and σ_y^2 denote the variance of the ground truth and the super-revolved depth maps, respectively; and c_1 and c_2 denote two variables that are used to stabilize the division.

3 FILTERING-BASED METHODS

In this section, we introduce the filtering-based methods. Based on the filter kernel construction process, we group existing methods into four categories: BF [114] and its variants, NL-Means filter [6] and its variants, GF [44] and its variants, and dynamic filters. The dynamic means that some factors in the filter are dynamic, such as dynamic guidance [42] and dynamic neighbors [145]. Note that for those methods that employ the optimization algorithm or deep convolution neural network for filter kernel generation, we classify them in this category, as they follow the basic framework of the image filters. A brief summary of the filtering-based methods can be found in Table 3.

3.1 Bilateral Filters

The BF, originally introduced by Tomasi and Manduchi [114], is a kind of image filter that aims to smooth the image as well as preserve the edge. The basic idea of the BF is to take into account both the spatial distance and the variation of intensities when constructing the filter kernels. Formally, for a pixel at the index p, the filtered output can be obtained by the following equation:

$$\boldsymbol{Y}_p^* = \frac{1}{\text{W}_p}\sum_{q\in\mathcal{N}_p} f(\|p-q\|_2^2)g(\|\boldsymbol{Y}_p - \boldsymbol{Y}_q\|_2^2)\boldsymbol{Y}_q, \tag{5}$$

where $\boldsymbol{Y}_p$ and $\boldsymbol{Y}_p^*$ are the input and output images at the pixel index p; $\mathcal{N}_p$ are the neighbors of the pixel p; $f(\cdot)$ and $g(\cdot)$ are two Gaussian kernels; and $\text{W}_p = \sum_{q\in\mathcal{N}_p} f(\|p-q\|_2^2)g(\|\boldsymbol{Y}_p - \boldsymbol{Y}_q\|_2^2)$ denotes the normalization factor. Following Tomasi and Manduchi [114], Eisemann and Durand [31] propose JBF, a joint bilateral filter that applies the range filter to the guidance image $\boldsymbol{G}$, and

Table 3. Brief Summary of the Filtering-Based Methods

Method	Published	Category	Basic Idea
Bilateral filter	BF [114]	ICCV-1998	Constructs the kernels by considering both the spatial distance and the variation of intensities.
	JBU [62]	TOG-2007	Employs the spatial filter on the LR depth map and the range filter on the HR intensity guidance image.
	JGU [75]	CVPR-2013	Uses the geodesic distance instead of the Euclidean distance to quantify the dissimilarities of the pixels.
	CoF [55]	CVPR-2017	Replaces the range filter in the BF with a co-occurrence matrix to distinguish the boundaries.
	JTF [82]	TCYB-2016	Constructs the kernels with the spatial distance, color difference, and local depth gradient.
Non-local means	JNLF [53]	CVPR-2010	Similar to the JBU [62] but calculates the affinity by using image patches instead of single pixels.
Guided filter	GF [44]	ECCV-2010	Assumes that the filtered result is a local linear transform of the guidance image, and the result is obtained by solving a least-squares problem.
	WGIF [72]	TIP-2014	Integrates an edge-aware weighting into the GF [44] to deal with the problem of halo artifacts.
	GDGF [63]	TIP-2015	Introduces explicit first-order edge-aware constrains into the GF [44] to better preserve the edges.
	SKWGIF [110]	TIP-2020	Employs the steering kernel to learn the boundary information and incorporates it into the filter process.
	SVLRM [24]	TPAMI-2021	Learns the linear representation coefficients (a_k, b_k) in the GF [44] by a deep convolution neural network.
	UMGF [102]	TIP-2021	A new formulation of GF inspired by unsharp masking, and it only needs to estimate one coefficient.
Dynamic filter	SGF [148]	CVPR-2015	Segment graph based filter that takes the tree distance calculated on the segment graph as guidance affinity.
	SWF [145]	CVPR-2019	A general filtering framework in which the neighbors of each pixel can be adjusted dynamically.
	JFMS [100]	IJCV-2017	Introduces the concept of mutual-structure, and the goal of the image filter can be achieved by correctly transferring the mutual-structures.
	SDF [42]	TPAMI-2018	Incorporates both static (guidance image) and dynamic (regularized target image) guidance into a unified framework to alleviate the effect of inconsistent structures.
	muGIF [41]	TPAMI-2020	Proposes a new measurement to manage the structure similarity of two inputs at the pixel level and a global optimization objective.
	DKN [58]	IJCV-2021	Joint learns the kernel weight and the neighbors by a neural network.

it can be formulated as follows:

$$Y_p^* = \frac{1}{W_p} \sum_{q \in \mathcal{N}_p} f(\|p - q\|_2^2) g(\|G_p - G_q\|_2^2) Y_q. \tag{6}$$

Motivated by JBF [31], Kopf et al. [62] propose JBU, a a joint bilateral upsamplingto restore an HR depth map from an LR observation with an HR color image as guidance. JBU [62] uses the same kernel as JBF [31], and it can be expressed as follows:

$$Y_p^* = \frac{1}{W_p} \sum_{q \in \mathcal{N}_p} f(\|p_\downarrow - q_\downarrow\|_2^2) g(\|G_p - G_q\|_2^2) Y_{q_\downarrow}, \tag{7}$$

where p and q represent the coordinates in the HR image, and $p_\downarrow$ and $q_\downarrow$ mean the corresponding coordinates (possibly fractional) in the LR image. The BF cannot determine whether an edge is within a texture or not, as it is defined by a spatial Gaussian and a range Gaussian filter. Jevnisek and Avidan [55] propose the **Co-Occurrence Filter (CoF)** to address this problem. In particular, they propose a normalized co-occurrence matrix that can represent the co-occur frequency of pixel values to replace the range Gaussian filter used in the BF [114], and the CoF [55] is given by

$$Y_p^* = \frac{1}{\mathrm{W}_p} \sum_{q \in \mathcal{N}_p} f(p,q) \frac{C(a,b)}{h(a)h(b)} Y_{q_\downarrow}, \tag{8}$$

where $f(\cdot)$ is a Gaussian filter kernel, $h(a) = \sum_p [I_a^{\mathrm{g}} = a]$, and $[\cdot]$ equals 0 if the expression in the brackets is false and 1 otherwise, and C is a co-occurrence matrix, and the co-occurrence value of a and b can be obtained by the following equation:

$$C(a,b) = \sum_{p,q} f(\|p-q\|_2^2)[G_a = a][G_b = b]. \tag{9}$$

To simplify the computation, the authors calculate the co-occurrence matrix only in a local window.

Instead of employing the Euclidean distance as JBU [62], Liu et al. [75] propose JGU, a joint geodesic upsamplingthat uses the geodesic distance to measure the affinity between two coordinates. Existing methods are based on the assumption that the discontinuities in depth and the guidance are consistent. Actually, this assumption may not always be true, and this will lead to the problem of texture copying. To address this issue, Li et al. [71] propose a joint trilateral upsampling that generates filter kernels by incorporating edge information from the LR depth map:

$$f(\|p_\downarrow - q_\downarrow\|_2^2) g(\|G_p - G_q\|_2^2) h(\|Y_{p_\downarrow} - Y_{q_\downarrow}\|_2^2). \tag{10}$$

Lo et al. [82] propose a similar idea that advances the JBF [31] with the LR depth map for kernel construction, named **Joint Trilateral Filter (JTF)**. The kernel function of JTF [82] can be obtained by replacing the third term of Equation (10) with the gradient map of the interpolated LR depth map. JTF [82] can generate sharp edges for lower upsampling factors, but the performance will drop when the upsampling factor becomes large. Qiao et al. [89] propose to use the result of the segmentation of the guidance image as one of the weight terms for kernel generation, as they find that the segmented regions can adhere well to the depth boundary.

3.2 NL-Means Filters

The basic idea of the NL-Means filter [6] is similar to the BF [114] method that restores a pixel as a weighted average of its neighbor pixels. However, the definition of the kernel function $w(\cdot)$ is slightly different. For the NL-Means filter, the similarities are determined by the patches surrounding both pixels instead of the single-pixel values, which is formulated as

$$w(Y_p, Y_q) = \exp\left(-\frac{1}{h} \sum_{k \in N} f(\|k\|)(Y_{p+k} - Y_{q+k})^2\right), \tag{11}$$

where h means a smoothing parameter, N denotes the number of pixels in a patch, and k is the offset of the center pixel. Inspired by Kopf et al. [62] and the NL-Means filter [6], Huhle et al. [53] propose a **Joint Non-Local Means Filter (JNLF)**, which has the following form:

$$Y_p^* = \frac{1}{\mathrm{W}_p} \sum_{q \in \mathcal{N}_p} \xi_{pk} w(G_p, G_q) w(Y_{p_\downarrow}, Y_{q_\downarrow}) Y_{q_\downarrow}, \tag{12}$$

where $\xi_{pk} = \exp\left(-(Y_{p_\downarrow} - Y_{(p+k)_\downarrow})^2\right)/h$ represents a weighting factor to restrict affinity comparison within areas of similar depths.

3.3 Guided Filters

The widely used BF [114] and its variants may suffer from gradient reversal artifacts, especially for detail enhancement tasks [3, 29, 34], and are typically time consuming. To alleviate these problems, He et al. [44] propose GF, a novel translation-variant image filter based on the assumption that the filtered result $\boldsymbol{Y}^*$ can be obtained from the guidance image $\boldsymbol{G}$ by a local linear transform in a window $\mathcal{N}$, which can be expressed as follows:

$$Y_p^* = a_k G_p + b_k, \forall k \in \mathcal{N}_k, \tag{13}$$

where a_k and b_k are two linear coefficients that are assumed to be constant values in the window $\mathcal{N}_k$ centered on the pixel k. The values of these two linear coefficients can be obtained by minimizing an objective function that calculates the difference between the filtered result and the input image:

$$\arg\min_{a_k, b_k} \sum_{p \in \mathcal{N}_k} \left((a_k G_p + b_k)^2 - Y_p + \gamma a_k^2 \right), \tag{14}$$

where γ denotes a regularization parameter that is utilized to penalize large values for a_k. By solving a least-squares problem of Equation (14), we can obtain the optimal values of a_k and b_k, and they are

$$a_k = \frac{\frac{1}{|\mathcal{N}|} \sum_{p \in \mathcal{N}_k} G_p Y_p - \overline{G}_k \overline{Y}_k}{\sigma_k^2 + \gamma}, \quad b_k = \overline{Y}_k - a_k \overline{G}_k, \tag{15}$$

where $\overline{G}_k$ and $\overline{Y}_k$ are the mean of the guidance and target images in $\mathcal{N}_k$, respectively; σ_k^2 denotes the variance of the guidance image in $\mathcal{N}_k$; and $|\mathcal{N}|$ means the number of pixels in $\mathcal{N}$. A pixel p can be involved in different windows, which may produce different values of Y_p^* in Equation (13). The authors propose a naive strategy that averages all possible values of Y_p^* as the final filtered result:

$$Y_p^* = \overline{a}_p G_p + \overline{b}_p, \tag{16}$$

where $\overline{a}_p = \frac{1}{|\mathcal{N}|} \sum_{k \in \mathcal{N}_p} a_k$ and $\overline{b}_p = \frac{1}{|\mathcal{N}|} \sum_{k \in \mathcal{N}_p} b_k$.

As an edge-preserving smoothing filter, the GF [44] not only prevents gradient reversal artifacts but also enjoys the merit of high computational efficiency. However, it may suffer from halo artifacts near some edges because of the filter's local character, and the regularization parameter is set as a constant. To address these problems, Li et al. [72] propose WGIF, a guided image filter that incorporates an edge-aware weighting into the GF [44] to dynamically adjust the regularization parameter. The edge-aware weighting $\Gamma(k)$ is calculated by integrating local and global information from the guidance image and is defined as follows:

$$\Gamma(k) = \frac{1}{N} \sum_{i=1}^{N} \frac{\sigma_k^2 + \varepsilon}{\sigma_i^2 + \varepsilon}, \tag{17}$$

where σ_k^2 and σ_i^2 denote the variance of the guidance image in a 3×3 window centered on the pixel k and i, respectively; ε means a small constant, and the authors set it as $(0.001 \times L)^2$ where L represents the dynamic range of the input target image. After obtaining the weight, the authors incorporate it into the object function of the GF [44], which can be expressed as follows:

$$\arg\min_{a_k, b_k} \sum_{p \in \mathcal{N}_k} \left((a_k G_p + b_k)^2 - Y_p + \frac{\gamma}{\Gamma(k)} a_k^2 \right). \tag{18}$$

Compared to the GF [44], the WGIF [72] can produce sharper edges and alleviate halo artifacts. However, as mentioned in the work of Chen et al. [110], the GF [44] and the WGIF [72] tend to produce results with blurred boundaries, as they adopt the average strategy to calculate the final result. In this average strategy, the weights of all pixels in a local window are assigned with the same values, and the boundary directions will easily be neglected. To solve this problem and make better use of the boundary information, Sun et al. [110] propose to incorporate the structure priors into the WGIF [72], where the priors are learned from the guidance image by using the steering kernel [111].

In the GF [44], the authors assume that the filtered result is a local linear transformation of the guidance image as written in Equation (13). According to this, we can see that the gradient of the target and the guidance images in a local window would satisfy the following equation:

$$\nabla Y_p^* = a_k \nabla G_p. \tag{19}$$

Equation (19) guarantees that the target image has the same structure as the guidance image since the coefficient a_k is constant for each local window. Therefore, inconsistent textures in the guidance image will be transferred to the target image, and this would lead to texture-copying artifacts. Furthermore, the GF [44] employs the mean filter to obtain the filtered output, which may blur the edges. To address this problem, Dong et al. [24] propose SVLRM, a spatially variant linear representation modelthat uses a deep convolution neural to learn the linear representation coefficients. The filtering process can be represented by the following equation:

$$Y_p^* = a(G, Y)G + b(G, Y), \tag{20}$$

where $\boldsymbol{I}^{\mathrm{g}}$ and $\boldsymbol{I}^{\mathrm{t}}$ denote the spatially variant linear representation coefficients that are learned by a deep neural network. The SVLRM [24] leverages the neural network to simultaneously estimate two coefficients of the GF; this is suboptimal and tends to generate inconsistent edges. The traditional unsharp masking [86] technique can enhance the edge by estimating only one coefficient. Inspired by this, Shi et al. [102] propose a novel and simplified formulation of the GF [44], namely the **Unsharp Mask Guided Filter (UMGF)**. The unsharp masking [86] technique can be used to enhance image sharpness, which is described as follows:

$$Y^* = \lambda \left(Y - \mathcal{F}_L(Y)\right) + Y, \tag{21}$$

where Y^* denotes the enhanced image and Y is the input image, $\mathcal{F}_L$ is a low-pass filter; $(Y - \mathcal{F}_L(Y))$ represents the unsharp mask, and λ is a coefficient to control the degree of enhancement. The filtered result of the GF [44] in Equation (16) can be rewritten as follows:

$$Y^* = \frac{1}{|\mathcal{N}|} \sum_{k \in \mathcal{N}_p} (a_k \overline{G}_p + b_k). \tag{22}$$

To eliminate b_k, we can put Equation (15) in Equation (22):

$$Y^* = \frac{1}{|\mathcal{N}|} \sum_{k \in \mathcal{N}_p} a_k G_p + \frac{1}{|\mathcal{N}|} \sum_{k \in \mathcal{N}_p} \left(\overline{Y}_p - a_k \overline{G}_k\right). \tag{23}$$

Then, the equation can be changed to the following equation:

$$Y^* = \frac{1}{|\mathcal{N}|} \sum_{k \in \mathcal{N}_p} a_k (G_p - \overline{G}_k) + \widetilde{Y}_p, \tag{24}$$

where $\widetilde{Y}_p = \frac{1}{|\mathcal{N}|} \sum_{k \in \mathcal{N}_p} Y_p$. As $\overline{Y}$ is the output of a mean filter, we can assume that $\overline{G}_p$ is close to its mean in a local window $\mathcal{N}_p$. Thus, Equation (24) can be rewritten as follows:

$$Y^* = \overline{a}_p (G_p - \widetilde{G}_p) + \overline{Y}_p, \tag{25}$$

where $\overline{a}_p = \frac{1}{|\mathcal{N}|}\sum_{k\in\mathcal{N}_p} a_k$ and $\widetilde{G}_p = \frac{1}{|\mathcal{N}|}\sum_{k\in\mathcal{N}_p} \overline{G}_k$. From Equation (25), we can see that for guided image filtering, the target image is first smoothed to remove redundant information, such as noise and textures. Next, the structure information $(G_p - \widetilde{G}_p)$ is transferred from the guidance image to the target image under the control of $\overline{a}_p$. The transferred structures are called the *unsharp mask*, which was mentioned previously. Equation (25) can be described in a more general way:

$$Y^* = f_a(Y_m, G_m) \odot G_m + \mathcal{F}_L(Y), \tag{26}$$

where $Y_m = Y - \mathcal{F}_L(Y)$ and $G_m = G - \mathcal{F}_L(G)$ are the unsharp masks calculated from the target and the guidance images, respectively and f_a is a function of controlling the degree of structure to be transferred. In this work, the authors propose to learn the function f_a using a deep neural network.

3.4 Dynamic Filters

Previous image filters are based on the assumption that the structures between the guidance image and depth image are consistent. However, inconsistent or even completely different structures usually exist in the two images. Directly adopting this guidance is prone to generate unsatisfactory results.

To solve this problem, Ham et al. [42] propose a **Static and Dynamic Filter (SDF)** to take advantage of static and dynamic guidance. The static guidance shares a similar idea with existing image filters, which modulates the target image with a weight function obtained from the guidance image. The dynamic guidance is derived from the regularized images—that is, iteratively using the already regularized images as the dynamic guidance to constrain the output with the aim of alleviating the effects of inconsistent structures. Shen et al. [100] propose the concept of mutual-structure to improve the performance of joint processing in the restoration of common structures. Shi et al. [100] define three types of structure: (1) *mutual-structure*: the common edges contained in the corresponding two patches; (2) *inconsistent structure*: the different structures contained in the two patches; and (3) *smooth regions*: the common low-variance smooth areas in the two patches. Obviously, the goal of the image filter is to transfer the mutual-structures while preventing the inconsistent ones. The mutual-structure of two patches can be measured by

$$\mathcal{S}(Y_p^*, Y_p) = \left(\sigma(Y_p^*)^2 + \sigma(G_p)^2\right)\left(1 - cov(Y_p^*, G_p)/\sqrt{\sigma(Y_p^*)\sigma(G_p)}\right)^2, \tag{27}$$

where $\sigma(Y_p^*)$ denotes the variance of a patch in a local window centered on the pixel p, and $cov(\cdot)$ means the covariance of the patch intensity. The patch-level filtering can be achieved by optimizing this equation. Compared to the SDF [42], this method can process two inputs simultaneously. However, the results of this method usually suffer from halo artifacts because of its local nature. To solve this problem, Guo et al. [41] propose muGIF, a mutually guided filtering that designs a global optimization method to improve filtering performance. In addition to the guidance, the neighbors can also be adjusted dynamically. Yin et al. [145] find that the edge pixels can be well recovered by a weighted average process if its neighboring pixels come from the same side of the edge. To achieve this goal, they propose a **Side Window Filtering (SWF)** technique where the side window of each pixel can be obtained by solving an optimization problem.

4 OPTIMIZATION-BASED METHODS

In this section, we introduce the optimization-based methods. These methods formulate GDSR as an optimization problem and introduce various priors to make the ill-posed problem traceable. Although the technique details are different, most of the methods in this category share a similar

Table 4. Brief Summary of the Prior-Based Methods

Category	Methods	Published	Basic Idea
Markov Random Field	MR-MRF [22]	NIPS-2005	The pioneering work that utilizes MRF for multi-modal image fusion.
	IMAMRF [163]	TCSVT-2016	Proposes a quantitatively metric to measure the inconsistency between the RGB-D pairs and embeds this metric into the MRF framework to alleviate texture-copying artifacts.
	MSFEE [164]	TIP-2018	Proposes to compute the guidance affinity by using the tree distance to maintain the depth boundaries.
Auto-Regressive Model	AR [138]	TIP-2014	Proposes an adaptive color-guided AR model for depth recover where the AR coefficients are derived from the local correction of the LR depth map and non-similarity of the guidance image.
	TSDR [56]	TIP-2018	A method that includes both transform and spatial domain regularizers. The transform domain regularizer is a patch-based AR model to characterize intra-patch correlations, and the spatial domain regularizer is a multi-directional TV to capture the geometrical structures.
	SRAM [121]	TMM-2019	A method that combines multi-direction sparse representation and AR to represent depth edges in both patch and pixel level.
Total Variation	TGV [35]	ICCV-2013	Employs the second-order TGV as the regularization for GDSR.
	FCN-PDN [90]	BMVC-2016	Models the TV regularizer in a non-local neighborhood and learns the weighting factor via a network.
	NBTV [123]	TIP-2016	Integrates the BF [114] into the TV to preserve the sharpness and consistency of edges.
Graph Laplacian	JGIE [80]	TIP-2019	Proposes a novel GDSR method by combining two graph domain regularizations: the internal smoothness prior and the external gradient consistency constrain.
	JARMT [79]	TIP-2019	Proposes a depth map restoration method that includes a local and a non-local manifold model to fully exploit the local smoothness prior and the non-local self-similarity structures.
	ADFTG [150]	TCSVT-2020	Proposes an adaptive data fidelity term to optimally generate each depth pixel and a unified graph-based regularization term to preserve the PWS and sharpen the edges.
	GNNLG [135]	TOMM-2020	Proposes a model that combines the nuclear norm and group-based graph prior to exploit the low-rank and non-local similarity properties of the depth map.
	DSSR [120]	TCSVT-2021	Proposes a dual normal-depth regularizer to exploit the geometric relationship between the depth and its corresponding normal and a reweighted graph Laplacian regularizer to obtain sharp edges.
	GraphSR [17]	CVPR-2022	Proposes to learn the affinity graph and then embeds the learned graph to a differentiable optimization layer to regularize the upsampling process.
Others	LR [84]	CVPR-2014	Proposes a low-rank matrix completion-based depth map enhancement method that is robust to the noise and has weak correlation between the depth map and its corresponding color image.
	JLNLR [25]	TMM-2017	Combines both local and non-local regularization for depth map recovery. The local regularization contains a weighted TV and a dual AR model, and the non-local regularization is a low-rank constrain.
	RGDMR [77]	TIP-2017	Proposes a robust penalty function as smoothness term to handle the issue of structure inconsistency.
	SPIDM [144]	PR-2020	Proposes a image decomposition model for depth map recovery, where the depth map is decomposed into a local smooth surface and a piecewise constant component. The former is modeled by a least-squares polynomial approximation, whereas the latter is modeled by a sparsity-promoting prior.
	DGDIE [37]	TPAMI-2019	Proposes a weighted analysis representation model for depth map reconstruction. The optimization process of this model is unfolded as a series of stage-wise operations that can be obtained by a task-driven training strategy.
	GSF [78]	TPAMI-2021	A generalized framework that combines the truncated Huber penalty function for image filtering. By varying the parameters of this function, the framework can obtain various smoothing natures.

framework, which can be formalized as follows:

$$\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}^*} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}^*\| + \lambda\phi(\boldsymbol{y}^*), \tag{28}$$

where $\boldsymbol{y}^* \in \mathbb{R}^N$ is the recovered HR depth map with N pixels, and y is the observed LR depth map. λ is a weight to balance the data term and the regularization term, and $\phi(\boldsymbol{y}^*)$ means the regularization term to smooth the result $\boldsymbol{y}^*$ and ensure that it has the same structure as the guidance image $\boldsymbol{g}$. The major distinction among the different approaches in this category is their regularization term. We present a brief summary of existing optimization-based methods in Table 4, and the technical details of each method can be found in the following section.

4.1 Markov Random Field

Markov Random Field (MRF) is an undirected graphical model that can be used to represent a joint probability distribution. It has been extensively applied in many computer vision and image processing tasks, such as semantic segmentation, image restoration, and depth map super-resolution. The MRF can be solved by optimizing the Gibbs energy function on a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ [13]:

$$E = \sum_{p \in \mathcal{V}} U(\boldsymbol{y}_p) + \lambda \sum_{(p,q) \in \mathcal{E}} V(\boldsymbol{y}_p, \boldsymbol{y}_q), \tag{29}$$

where $\mathcal{V}$ means the set of all pixels in an image and $\mathcal{E}$ means the set of edges that connect p and q. $U(\boldsymbol{y}_p)$ and $V(\boldsymbol{y}_p, \boldsymbol{y}_q)$ denote the data term and the regularization term, respectively. Based on this, Diebel and Thrun [22] propose a **Multi-Resolution Markov Random Field (MR-MRF)** framework for GDSR. In the MR-MRF, the depth map super-resolution is modeled as a pixel-labeling problem, and the MRF is utilized to estimate the true label for each pixel of the coarse depth map. The object function of MR-MRF [22] is as follows:

$$E = \sum_{p \in \mathcal{V}} (\boldsymbol{y}_p^* - \boldsymbol{y}_p)^2 + \lambda \sum_{(p,q) \in \mathcal{E}} (\boldsymbol{y}_p^* - \boldsymbol{y}_q^*)^2 \exp\left(-c(\boldsymbol{g}_p - \boldsymbol{g}_q)^2\right), \tag{30}$$

where $\boldsymbol{y}^*$ is the super-resolved depth map, $\boldsymbol{y}$ is the LR depth map, and $\boldsymbol{g}$ is the guidance image; c is a constant value that quantifies the degree of edge smoothing. Lu et al. [83] further extend this work by introducing a new data term that is more suitable for the characteristics of the depth map. The quadratic distance adopted in Equation (30) is prone to generate results with blurred-edge artifacts, and to solve this problem, Lo et al. [81] propose to use exponential-type functions to define the data term and the regularization term. Previous methods for GDSR are based on the assumption that the edges between the depth map and its corresponding guidance image are consistent. However, this assumption is not always true, and texture-copying artifacts will be introduced when the assumption is invalid. To overcome this problem, Zuo et al. [163] propose a quantitative score to measure the inconsistency between the depth map and its corresponding color image. This score can be integrated into the regularization term of the MRF-based framework to alleviate texture-copying artifacts and preserve depth boundaries. The cost function of Zuo et al. [163] is as follows:

$$E = \sum_{p \in \mathcal{V}} (\boldsymbol{y}_p^* - \boldsymbol{y}_p)^2 + \lambda \sum_{(p,q) \in \mathcal{E}} (\boldsymbol{y}_p^* - \boldsymbol{y}_q^*)^2 \exp\left(-\frac{|\nabla \boldsymbol{g}_{pq} \cdot (1 - \alpha_{pq})| + |\nabla \hat{\boldsymbol{y}}_{pq} \cdot \alpha_{pq}|}{2\sigma^2}\right), \tag{31}$$

where $\nabla \boldsymbol{g}_{pq}$ and $\nabla \boldsymbol{y}_{pq}^*$ denote the difference between the pixel p and q for the guidance image and the coarsely interpolated depth map $\hat{\boldsymbol{y}}_{pq}$, respectively; σ is a constant value; and α is a confidence map that is used to describe the inconsistency between the depth map and its corresponding color image, and the detailed calculation process can refer to the original paper [163]. Existing MRF-based GDSR methods that calculate the similarity in the regularization term are only based on the differences between the pixel and its neighbors. This strategy neglects the local structure on the depth map and tends to generate the depth map with blurred edges. To better maintain the depth edges, Zuo et al. [164] propose to calculate the similarity in the minimum spanning trees space where the paths in this space can represent the local structure of the depth map.

4.2 Auto-Regressive Model

The center theme of the **Auto-Regressive (AR)** model is that a signal can be represented as a linear combination of itself with carefully modified AR coefficients. Let $\boldsymbol{y} \in \mathbb{R}^N$ be the input signal, and the output of the AR model at position p is given by $\boldsymbol{y}_p^* = \sum_{q \in \mathcal{N}_p} \alpha_{p,q} \boldsymbol{y}_q$, where $\mathcal{N}_p$ means the neighborhood pixel of the pixel p, and α are the AR coefficients. Yang et al. [138] find that the

depth signals can be well represented by the AR model if we can adjust the coefficients based on the unique characteristics of the depth signals. To this end, the authors propose an adaptive AR model for GDSR that adjusts the coefficients by considering both the local correlation in the LR depth map and the non-local similarity in the HR guidance image. This model is given by

$$\arg\min_{\boldsymbol{y}^*} \sum_{p\in O} (\boldsymbol{y}_p^* - \hat{\boldsymbol{y}}_p)^2 + \sum_p \left(\boldsymbol{y}_p^* - \frac{1}{C_p} \sum_{q\in\mathcal{N}_p} \exp\left(-\frac{(\boldsymbol{y}_p^* - \boldsymbol{y}_q^*)^2}{2\sigma_1^2} \right) \exp\left(-\frac{\sum_{i\in C} \left\| \mathbf{B}_p \circ \left(\mathcal{P}_p^i - \mathcal{P}_q^i \right) \right\|_2^2}{2\times 3\times \sigma_2^2} \right) \boldsymbol{y}_q^* \right), \quad (32)$$

where O denotes the set of pixels of the coarsely interpolated depth map $\hat{\boldsymbol{y}}$, and $\mathcal{N}_p$ means the neighborhood pixel of the pixel p; σ_1 and σ_2 are constant values to control the decay rate of the exponential function; $\mathcal{P}_p^i$ represents the operator used to extract a patch centered at p in the color channel i; $\circ$ is pixel-wise multiplication; and $\mathbf{B}$ denotes the bilateral kernel function that is used to capture shape information of local image structures. Jiang et al. [56] find that the pixel-based AR model tends to produce undesired artifacts when there is a large difference between the center pixel and its neighbors. To remedy this limitation, the authors propose a patch-based AR model where the relationship is modeled by non-local similar patches. Wang et al. [121] propose a compound method by combining dictionary learning and the AR model. The dictionary learning is used to represent the inter-patch (patch-level) correlation of depth edges, whereas the AR model is used to describe the intra-patch (pixel-level) correlation of depth edges. Wang et al. [123] extend the AR model by selecting the largest variance channel as input, as they find that the average operator adopted in the original AR model tends to impair structural awareness.

4.3 Total Variation

Total Variation (TV), or the ROF model, is a classical regularization functional proposed by Rudin et al. [92] in 1992 and has been widely used in the area of image restoration. The TV [92] distinguishes the edge and the noise using the image gradients, and diffusion is performed only along the edge direction; thus, the edge information can be well preserved. Unfortunately, the TV can only approximate the piecewise constant function and tends to generate the "staircasing artifacts" for the smooth region. To solve this problem, Bredies et al. [4] propose **Total Generalized Variation (TGV)**, a generalized mathematical model of TV that can approximate the piecewise multinomial function of any order. Motivated by this, several methods have been proposed that incorporate TV and its variants as regularizers and achieved promising performance for the GDSR.

Ferstl et al. [35] propose to use a second-order TGV weighted by an anisotropic diffusion tensor as the regularization function to upsample the depth map, which can avoid surface flattening. The energy function is formulated as follows:

$$\min_{\boldsymbol{y}^*, \boldsymbol{z}} \|\boldsymbol{y}^* - \hat{\boldsymbol{y}}\|_2^2 + \alpha_1 \boldsymbol{T} |\nabla \boldsymbol{y}^* + \boldsymbol{z}| + \alpha_0 |\nabla \boldsymbol{z}|, \quad (33)$$

where the last two terms of this equation denote the second-order TGV regularizer; T is the anisotropic diffusion calculated from the HR guidance image. Riegler et al. [91] propose to incorporate the second-order TGV model into a deep convolution network, which combines the merits of deep learning and variational methods. In this model, all the optimization procedures of the variational model are unrolled so that the whole model can be trained end-to-end. Riegler et al. [90] propose a **Non-Local Total Variation (NLTV)** regularization with a robust norm for GDSR. The objective function of NLTV is formulated as follows:

$$E = \sum_p \sum_{q\in\mathcal{N}_p} \exp\left(-\frac{\|p-q\|_2}{\sigma_d} - \frac{\|\hat{\boldsymbol{g}}_p - \hat{\boldsymbol{g}}_q\|_2}{\sigma_v} \right) |\boldsymbol{y}_p^* - \boldsymbol{y}_q^*|_\epsilon, \quad (34)$$

where $|\cdot|_\epsilon$ denotes the Huber norm [52], which not only can smooth the depth reconstruction but also can preserve sharp boundaries. ϵ is the parameter for the Huber norm; $\hat{\boldsymbol{g}}$ is the guidance image learned by a neural network. Wang et al. [123] propose **Normalized Bilateral Total Variation (NBTV)**, which extends the TV by involving the BF [114] and the normalized sparse measure. Compared to the TV, the NBTV [123] can produce results with sharper and more precise edges. The edges of the depth map are typically oriented at arbitrary directions. Nevertheless, most of the existing TV-based models only deal with the horizontal and vertical directions, which may oversmooth the fine details. Jiang et al. [56] propose WNTV, a weighted multi-direction TV that represents depth edges with spatially varying orientations. The objective function of WMTV is as follows:

$$E = \sum_i \|\curlyvee \boldsymbol{W}_i \nabla_{\theta_i} \boldsymbol{y}^*\|_1, \tag{35}$$

where ∇_{θ_i} means the gradient along the direction θ_i and $\curlyvee \boldsymbol{W}_i$ denotes a weighting matrix that is used to assign different values to different directions for each pixel.

4.4 Graph Laplacian

One of the most significant image priors for depth map restoration is that depth maps contain smooth areas separated by sharp boundaries, which is also known as PWS. Recent studies have shown the powerful capability of the graph Laplacian to deal with the PWS signal. Inspired by this, several works have been proposed to use it to exploit the smoothness of the depth map. In the following, we first present the general formulation of the graph Laplacian and then introduce some representative methods that employ the graph Laplacian as the regularizer.

A depth map can be modeled as a signal on a weighted undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{W})$, where $\mathcal{V}$ denotes the set of vertices (or nodes); $\mathcal{E}$ means the set of edges, and each edge connects a pair of vertices in $\mathcal{V}$; and $\boldsymbol{W}$ denotes the weighted adjacent matrix that is used to encode the similarity between two connected vertices. Note that we only consider $\mathcal{G}$ as an undirected graph with non-negative edge weights, so we have $\boldsymbol{W}_{ij} = \boldsymbol{W}_{ji}$ and $\boldsymbol{W}_{ij} > 0$. The graph Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where $\boldsymbol{D}_{ii} = \sum_j \boldsymbol{W}_{ij}$ means the diagonal degree matrix. Given the graph Laplacian matrix $\boldsymbol{L}$, the smoothness of a signal $\boldsymbol{y}$ on the graph can be formulated as a quadratic form of $\boldsymbol{L}$:

$$\boldsymbol{y}^\top \boldsymbol{L}\boldsymbol{y} = \frac{1}{2}\sum_{i,j} \boldsymbol{W}_{ij}(\boldsymbol{y}_i - \boldsymbol{y}_j)^2, \tag{36}$$

and by minimizing Equation (36), the signal x is promoted to be smooth with respect to the graph $\mathcal{G}$.

Generally, there are two main graph-based strategies used in the literature for the inverse problem: graph-based regularization and graph-based transform. For the graph-based regularization, the signal is generally assumed to be smooth with respect to the graph $\mathcal{G}$, and the image restoration task can be solved by the following equation:

$$\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}^*} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}^*\|_2^2 + \gamma \boldsymbol{y}^\top \boldsymbol{L}\boldsymbol{y}^*, \tag{37}$$

and the regularization term is small if each pair of nodes (i, j) that are connected by an edge has similar values or the edge weight $\boldsymbol{W}_{ij}$ is small. Since the Laplacian matrix $\boldsymbol{L}$ is a real, symmetric, positive semi-definite matrix, we can decompose it into a set of orthogonal eigenvectors, denoted by $\{\boldsymbol{u}_l\}_{l=1,\ldots,n}$, with real non-negative eigenvalues $0 = \eta_1 \le \eta_2 \le \cdots \le \eta_n$:

$$\boldsymbol{L} = \boldsymbol{U}\Lambda\boldsymbol{U}^\top, \tag{38}$$

where $\boldsymbol{U}$ is the eigen-matrix with $\boldsymbol{u}_i$'s as columns and Λ is the diagonal matrix with η_i's on its diagonal. The graph-based transform (GFT) can be formulated as $\alpha = \boldsymbol{U}^\top \boldsymbol{y}$.

The graph-based regularization can achieve better performance for the image super-resolution task, whereas the graph-based transform works well for the image denoising task [150]. To this end, Zhang et al. [150] propose a unified regular function named *transferred graph regularization* for image restoration tasks, which combines the advantages of the preceding two graph-based strategies. Liu et al. [80] propose a novel framework for GDSR that contains an internal smooth prior and an external gradient consistency constraint. For the internal smooth prior, the authors propose a graph Laplacian regularizer where the weight matrix is tailored to fully utilize information between the depth map and its corresponding color image. The external gradient consistency constraint is proposed to address the problem of structure discrepancy, and this constraint is motivated by the observation that the depth map gradient is small apart from the edge-separating region. Hence, the depth map gradient can be obtained by discarding the finite parts of the guidance image gradient:

$$\nabla_{\mathcal{G}} \boldsymbol{y} = \mathcal{T}_{\tau}(\nabla_{\mathcal{G}} \mathbf{g}), \tag{39}$$

where $\mathcal{T}_{\tau}$ denotes the hard thresholding operator with parameter τ, $\nabla_{\mathcal{G}} \boldsymbol{y}$ and $\nabla_{\mathcal{G}} \mathbf{g}$ mean the gradients of the depth map and the guidance image, and the gradient operator is defined as follows:

$$\nabla_{\mathcal{G}} \boldsymbol{y}_i = \sum_{j, j \sim i} \boldsymbol{W}(i, j)(y(i) - y(j)), \tag{40}$$

where $\boldsymbol{W}_{ij}$ means the edge weight between nodes i and j, and $j \sim i$ denotes that node j is connected to node i in graph $\mathcal{G}$. The final object function of this work is given by

$$\arg\min_{\boldsymbol{y}^*} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}^*\|_2^2 + \alpha \boldsymbol{y}^{*\top} \hat{\boldsymbol{L}} \boldsymbol{y}^* + \beta \|\nabla_{\mathcal{G}} \boldsymbol{y}^* - q\|_1, \tag{41}$$

where $\hat{\boldsymbol{L}} = \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}$; q is a constant value that means the threshold of the guidance image gradient. The second term in Equation (41) is the internal smoothness prior, whereas the third term in Equation (41) means the external consistency constraint. Liu et al. [79] propose a depth map restoration method that includes local and non-local manifold models. The local manifold is used to maintain the smoothness of the local manifold structure, and the non-local manifold model is designed to make full use of the self-similarity structures. Instead of constructing the graph with the pixels, Yan et al. [135] propose to construct the graph with a group of similar patches, as the group-based image restoration model can efficiently gather the similarity of different patches. Wang et al. [120] find that the edge weight distribution of an area with sharp edges is a bimodal distribution. Inspired by this, they propose a reweighted graph Laplacian regularizer to preserve sharp edges and promote the bimodal distribution of edge weights. To fully exploit the contextual information of the guide image and ensure strict fidelity of the super-resolved depth map to the LR one, de Lutio et al. [17] propose to learn the graph regularization. Specifically, they first use a deep neural network to learn the affinity graph; then, the learned affinity graph is sent to a differentiable optimization layer to regularize the upsampling process.

4.5 Others

In this section, we introduce some methods that employ optimization algorithms to address the ill-posed GDSR problem but do not belong to any of the preceding categories.

In the literature, the GDSR methods are mainly based on two assumptions: (1) the guidance image should be of high quality, and (2) the depth map and its corresponding guidance image should be highly correlated. However, these two assumptions may be invalidated in some cases. For example, the guidance image is usually noisy when the depth camera is used in low light conditions. The noisy image is prone to mislead the depth restoration algorithm and pollute the depth edges. To overcome these problems, Lu et al. [84] propose a low-rank matrix completion-based method

that takes into account the noisy guidance image. Their method is based on the observation that similar RGB-D patches lie in a very low-dimensional subspace. By assembling similar patches into a matrix and restricting low-rank constraint, the latent structure structure in the RGB-D patches can be essentially captured. Dong et al. [25] propose to combine both local and non-local regularization strategy for the depth map reconstruction. The local regularization contains two complementary local constraints: the gradient domain weighted TV and the spatial domain dual AR model. The non-local regularization is a low-rank constraint that aims to exploit the global characterization of color-depth dependency. Instead of designing an effective guidance weight, Liu et al. [77] propose a robust penalty function as a smoothness term to alleviate texture-copying artifacts. Ye et al. [144] find that the depth map mainly consists of smooth regions separated by sharp edges and propose an image decomposition model for depth map recovery. Concretely, the authors first decompose the depth map into a local smooth surface and an approximately piecewise constant component. Then, they propose using a polynomial smoothing filter to approximate the local smooth surface and a sparsity-promoting prior to model the piecewise constant component. Gu et al. [37, 38] propose WSAR, a weighted analysis representation model for guided depth map reconstruction, in which they combine the analysis representation regularization term and the guidance weighting term to model the relationship of the RGB-D pairs. The optimization of WASR is unfolded; thus, the proposed model can make full use of previous expertise as well as training data.

5 LEARNING-BASED METHODS

5.1 Dictionary Learning Methods

In this section, we introduce the learning-based methods, which include the dictionary learning methods and the deep learning methods. The dictionary learning methods aim at finding a set of basis atoms (dictionary) such that the input signal can be represented as a sparse combination of these atoms. Contrary to dictionary learning methods, the deep learning methods directly learn the non-linear mapping from the two inputs (i.e., the guidance image and the LR depth map) to the ground truth depth map by using the deep neural network. A brief summary of the learning-based methods can be found in Table 5.

In this section, we introduce the dictionary learning methods. We first present the general formulation of dictionary learning, then introduce some representative methods that employ dictionary learning to solve the ill-posed GDSR problem. Note that the convolution dictionary learning methods are also included in this section.

Generally, sparse dictionary learning consists of two steps: sparse coding and dictionary learning. Let $\boldsymbol{y} \in \mathbb{R}^N$ be an input signal and $\boldsymbol{D} \in \mathbb{R}^{N \times M} (N \ll M)$ be an over-complete dictionary; the sparse coding process aims to find a representation of $\boldsymbol{y}^*$ in the form of a linear combination of a small number of atoms in $\boldsymbol{D}$—that is, $\boldsymbol{y}^* \simeq \boldsymbol{D}\boldsymbol{\alpha}$ where $\boldsymbol{\alpha} \in \mathbb{R}^M$ should be sparse. The process of finding the sparse representation $\boldsymbol{\alpha}$ can be formulated as the following optimization problem:

$$\arg\min_{\alpha} \|\boldsymbol{y}^* - \boldsymbol{D}\boldsymbol{\alpha}\|_2 + \|\boldsymbol{\alpha}\|_0, \tag{42}$$

where $\|\cdot\|_0$ means the l_0 pseudo-norm counting the number of non-zero elements in the vector $\boldsymbol{\alpha}$. The preceding minimization problem is not convex due to the l_0 norm, and solving this problem is NP-hard. There are two typically ways to solve the preceding problem: one is to employ greedy methods such as orthogonal matching pursuit (OMP) [116], and the other is to relax the l_0 norm to the l_1 norm and solve it by using basis pursuit [8] or LASSO [93]. The dictionary learning step aims to find an over-complete dictionary such that the fidelity term in Equation (42) can be minimized when the $\boldsymbol{\alpha}$ is fixed. In the literature, many approaches have been proposed to learn the dictionary, such as MOD [33], K-SVD [1], and the Lagrange dual method [66].

Table 5. Brief Summary of the Learning-Based Methods

Category	Methods	Published	Basic Idea
Dictionary Learning	JIDSR [115]	TIP-2014	Proposes a model to learn joint sparse representations of the depth and the guidance images. In the model, the dictionaries and the coefficients for each image are different but share a common sparse support. Moreover, it proposes a Joint Basis Pursuit algorithm to infer the sparse coefficients.
	CDLSR [105]	TCI-2019	Proposes a coupled dictionary learning method to explicitly capture the multi-modal dependency in the sparse feature domain. To better characterize the similarities and disparities between different modalities, the authors propose to use common and unique sparse representations to represent the multi-modal images.
	RADAR [21]	TCSVT-2019	Proposes a method that combines the strengths finite rate of innovation theory with multi-modal dictionary learning for robust depth map super-resolution. Finite rate of innovation theory is used to do initial upsampling, and multi-modal dictionary learning is used to capture meaningful information from the guidance image.
	JMDL [20]	TIP-2019	Proposes a joint multi-modal dictionary learning algorithm that simultaneously learns three dictionaries and two transform matrices. The learned dictionaries are used to represent the LR depth map, the guidance image, and the HR depth map, whereas the learned transform matrices are employed to model the relationship among the spare representations of each modality.
	LMCSC [85]	TIP-2020	Proposes a multi-modal sparse coding model where different modalities are described by different dictionaries, and the same sparse representations for each modalities are assumed to be similar.
	CUNet [19]	TPAMI-2021	Proposes to represent each modality with two convolutional dictionaries, where one dictionary is utilized to represent the common features and the other is utilized to represent unique features.
	MCDL [36]	TIP-2022	Proposes to solve the multi-modal convolution dictionary learning problem by traditional optimization, which is demonstrated to be less sensitive than the deep unfolding strategy with limited training data.
Deep Learning	DMSG [54]	ECCV-2016	Proposes a multi-scale guided convolution network with multi-scale fusion strategy and high-frequency domain training method to boost the performance of depth map reconstruction.
	MGD [160]	TCSVT-2019	Proposes a method to improve the effectiveness of the guidance features. The method extracts the multi-scale guidance features in the HR domain and revisits the extracted features via dense connection.
	DJFR [70]	TPAMI-2019	Proposes a two-branch network with skip connection for joint image filtering.
	CCFN [128]	TIP-2019	Proposes a coarse-to-fine convolution neural network for depth map super-resolution. In the coarse stage, the authors use large filter kernels to obtain crude HR depth, and in the fine stage, the authors use small filter kernels to produce more accurate results.
	DepthSR [40]	TIP-2019	Proposes a residual U-Net architecture with hierarchical guidance branch and input pyramid branch.
	DSRN [127]	PR-2020	Proposes a novel framework that is based on deep edge-aware learning. This framework first learns the edge information, then recovers the HR depth map by using the proposed fast depth filling module and cascaded network.
	CCMSN [68]	PR-2020	Proposes a multi-scale symmetric network for depth map super-resolution, and its core is a correlation-controlled color guidance block which aims to improve the color guidance accuracy.
	CGN [159]	TMM-2020	Proposes a coarse to fine framework that contains a depth-guided intensity features filtering module and an intensity-guided depth features refinement module to mitigate the artifacts caused by edge misalignment.
	PMBAN [143]	TIP-2020	Proposes a progressive multi-branch aggregation network (PMBAN), which contains stacked MBA blocks to progressively reconstruct the depth map. The MBA block has multiple branches to extract multi-modal features, and the extracted feature are fused by the proposed fusion block.
	MIG [161]	TMM-2021	Proposes a multi-stage network to enhance the LR depth map. In each stage, the depth features in the image and the gradient domains are iteratively refined by the guidance features through two parallel streams.
	AHMF [156]	TIP-2021	Proposes a novel model that contains a multi-modal attention-based fusion module and a bi-directional hierarchical feature collaboration module to fuse multi-modal and multi-level complementary features.
	JIIF [112]	MM-2021	Models the GDSR as a neural implicit interpolation process, and the interpolation weights and values are learned by a proposed novel joint implicit image function.
	Bridge [113]	MM-2021	Proposes an MTL framework that combines the depth map super-resolution and monocular DE to improve the quality of reconstructed depth maps. The two tasks are interacted with a high-frequency bridge and a context guidance bridge.
	CMSR [98]	CVPR-2021	Proposes a training strategy for GDSR, which only need one weakly aligned RGB-D pair as the training data.
	FDSR [47]	CVPR-2021	Proposes a high-frequency guidance network that employs the octave convolution [10] to capture the high-frequency components of the RGB image to guide the depth map super-resolution process.
	CTCK [109]	CVPR-2021	Proposes to utilize the guidance information by knowledge distillation. In this method, the guidance image is only required at the training stage.
	DCTNet [153]	CVPR-2022	Proposes a method that extends the discrete cosine transform to a deep learning module to make the network explainable, and combines this module with a shared/private feature extraction and guided edge attention mechanism to address issues of cross-modal feature extraction difficulty and RGB texture over-transferred.

Sparse dictionary learning has long been employed to represent images and signals. To represent multi-modal images, Tosic and Drewes [115] propose a sparse generative model to learn the joint representation of the depth map $\boldsymbol{y}^*$ and the guidance image $\boldsymbol{g}$, where they represent different modalities with different dictionaries and different coefficients:

$$\boldsymbol{g} = \boldsymbol{D}^g \boldsymbol{a} + \boldsymbol{\eta}^g, \quad \boldsymbol{y}^* = \boldsymbol{D}^y \boldsymbol{b} + \boldsymbol{\eta}^y, \tag{43}$$

where $\boldsymbol{\eta}$ is the noise vector, and $\boldsymbol{a} = [a_1, \ldots, a_K]$ and $\boldsymbol{b} = [b_1, \ldots, b_K]$ are the sparse coefficients for the guidance and depth images, respectively. As the depth and guidance images are captured from the same scenes, the authors couple these two sets of sparse coefficients via latent variables ϕ_i, which can be defined as follows:

$$a_i = m_i^g \phi_i, \quad b_i = m_i^x \phi_i, \tag{44}$$

where m_i^g and m_i^x denote the magnitudes of the sparse coefficients, and ϕ_i represents the activity of the i-th sparse coefficient. By restricting the vector $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_K]$ to be sparse, we can obtain the sparse representation of both the depth and the guidance images. To better reconstruct the depth edges, Wang et al. [121] propose a multi-direction dictionary learning algorithm, where they first classify the patches according to the geometrical directions and then train the dictionaries for each class, respectively. Song et al. [105] propose CDLSR, a coupled dictionary learning based multi-modal image super-resolution method to exploit complex dependencies between different modalities in a learned sparse feature domain. The CDLSR [105] is made up of a coupled dictionary learning phase and a coupled super-resolution phase. The learning phase aims at learning a set of dictionaries to couple the multi-modal images together in the sparse feature domain and can be formulated as follows:

$$\underset{\substack{\{\boldsymbol{\Psi}_c^l, \boldsymbol{\Psi}^l, \boldsymbol{\Psi}_c^h, \boldsymbol{\Psi}^h, \boldsymbol{\Phi}_c, \boldsymbol{\Phi}\} \\ \{\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{V}\}}}{\arg\min} \left\| \begin{bmatrix} \boldsymbol{Y}^l \\ \boldsymbol{Y}^* \\ \boldsymbol{G} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Psi}_c^l & \boldsymbol{\Psi}^l & \boldsymbol{0} \\ \boldsymbol{\Psi}_c^h & \boldsymbol{\Psi}^h & \boldsymbol{0} \\ \boldsymbol{\Phi}_c & \boldsymbol{0} & \boldsymbol{\Phi} \end{bmatrix} \begin{bmatrix} \boldsymbol{Z} \\ \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \right\|_F^2 \text{s.t. } \|\boldsymbol{z}_i\|_0 + \|\boldsymbol{u}_i\|_0 + \|\boldsymbol{v}_i\|_0 \leq s, \forall i, \tag{45}$$

where $\boldsymbol{Y}^l, \boldsymbol{Y}^*$ are the LR and the HR depth maps, respectively; $\boldsymbol{G}$ is the guidance image; $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ represent the learned dictionaries and $\boldsymbol{Z}, \boldsymbol{U}$, and $\boldsymbol{V}$ represent the sparse representations; and $\|\cdot\|_F$ is the Frobenius norm. To better characterize the similarities and disparities between different modalities, the authors make the sparse representations $\boldsymbol{Z}$ and $\boldsymbol{U}$ shared for the same modality, and the sparse representations $\boldsymbol{U}$ and $\boldsymbol{V}$ different for the different modalities. To further improve the performance of this method, the authors propose to incorporate the multistage strategy and the neighborhood regression technique into CDLSR [105]. Deng et al. [21] propose a robust algorithm for GDSR, named *RADAR*, which combines the strengths of finite rate innovation theory and multi-modal dictionary learning. Deng and Dragotti [20] propose JMDL, a joint multi-modal dictionary learning algorithm for GDSR. In JMDL, the authors use three dictionaries to represent the LR depth map, the guidance image, and the HR depth map, respectively. Considering that the three images are captured from the same scene, the authors assume that the sparse representations of them are related.

To guarantee the sparse representation, the dictionary $\boldsymbol{D}$ should be sufficiently over-complete. This makes the sparse coding model only represent the low-dimensional signals and may suffer from curse of dimensionality for the high-dimensional signals, such as image data. To address this problem, the researchers propose to split the image into overlapping patches and perform sparse coding on each patch. However, this mechanism neglects the dependencies between patches, which tends to produce unsatisfactory results. To overcome these shortcomings, the convolutional sparse coding model [5, 129] has been proposed, which employs convolution filters to replace

traditional dictionary atoms. The convolutional sparse coding model is spatially invariant and avoids splitting the image into patches; thus, the consistency prior can be naturally exploited. Given an input image $\boldsymbol{y}$ and dictionary filters $\boldsymbol{D} = \{\boldsymbol{d}_i | i = 1, \ldots, N\}$, the goal of the convolution sparse coding model is to calculate a set of convolution sparse maps $\boldsymbol{R} = \{\boldsymbol{\alpha}_i | i = 1, \ldots, N\}$, which is expressed as follows:

$$\arg\min_{\{\boldsymbol{\alpha}_i\}} \frac{1}{2} \left\| \boldsymbol{y} - \sum_{i=1}^{N} \boldsymbol{d}_i \right\|_2 + \lambda \sum_i \|\boldsymbol{\alpha}_i\|_1, \tag{46}$$

where $\{\boldsymbol{\alpha}_i\}$ means the set of sparse maps, and the size of $\boldsymbol{\alpha}_i$ is the same as that of the image $\boldsymbol{y}$. To solve this optimization problem, many efficient algorithms [5, 125, 129] have been proposed. The convolutional sparse coding model has been successfully adopted in many image processing tasks, such as image super-resolution [39], image denoising [155], and multi-modal image restoration [133].

Marivani et al. [85] propose a multi-modal convolutional sparse representation model where each image modality is represented with its own dictionary filter, and different modalities are restricted to have similar sparse codes. Unlike in the work of Marivani et al. [85], where each image modality is represented by a single dictionary, Deng and Dragotti [19] propose to represent each image modality with two convolution dictionaries, where one dictionary is used to represent common features and the other is used to represent unique features. The work of Deng and Dragotti [19] restricts the number of common and unique features to be the same; however, this may not always be true. To overcome this limitation, Xu et al. [132] propose a model named *CUNet++*, which assigns different filters for different modalities. Moreover, they employ a recurrent architecture to extract the common and unique features to fully mimic the sparse model. Gao et al. [36] propose MCDL, a novel multi-modal convolutional dictionary learning model that models the correlation of different modalities at the image level. Similar to Deng and Dragotti [19], they employ the common and unique dictionaries to represent each image modality. The difference is that Gao et al. [36] propose to solve this multi-modal convolutional dictionary learning problem through traditional optimization, as they believe that the traditional optimization is more robust than the unfolding strategy when training data are limited.

5.2 Deep Learning Methods

In this section, we introduce the deep learning based methods. For this category, a straightforward way is to employ a two-stream network to extract features from the depth and color images separately, then the extracted features are fused to recover the degraded depth maps. However, this simple framework may not have the ability to produce the desired HR depth maps. To this end, researchers have developed various additional modules into this simple two-stream architecture to improve the depth map reconstruction performance. To better understand these methods, we introduce them from different perspectives as follows: (1) multi-modal information fusion strategy, (2) **Multi-Task Learning (MTL)**, (3) prior knowledge guided network design, (4) novel network architecture, and (5) loss function.

Multi-Modal Information Fusion Strategy. The depth map and its corresponding color image are geometric and photometric descriptions of the same scene, and they have strong structural similarity. Therefore, designing effective fusion algorithms to fuse these two images plays a significant role in GDSR. There are two widely used fusion tactics: single-level fusion and multi-level fusion. Dde Lutio et al. [17] fuse the multi-modal information at the date level, in which the original guidance and the pre-upscaled depth images are concatenated and sent to a convolution neural network for feature extraction. Kim et al. [58] employ two different neural networks to learn the kernel weight and offset from the guidance image and the depth map, respectively. Then, the

learned kernel weight and offset are fused to generate the final weight and offset via element-wise multiplication. Instead of fusing multi-modal information at the image level, Li et al. [69, 70] propose to fuse multi-modal information at the feature level. Specifically, they first send the depth and color images to a two-stream network for feature extraction, then the extracted deep features are concatenated and sent to another network to generate the final result.

However, such a single-level fusion may not fully exploit the latent correlations between the depth and the guidance images. To address this problem, Hui et al. [54] propose a multi-scale fusion strategy, where the multi-scale guided features are obtained by a VGG-like neural network. Motivated by this, various works have been developed and have achieved promising performance. For example, Guo et al. [40] propose to use a UNet-based framework. In this framework, the encoder sub-network is used to extract features from the depth maps, whereas the decoder sub-network is used to fuse the guidance features at the corresponding scale. Zuo et al. [161] propose a multi-scale alternatively guided network, where the depth map is iteratively upsampled by 2×. In each upsampling, the depth features are alternatively guided by its corresponding gradient features and guidance features. In another work, Zuo et al. [160] propose a novel framework to improve the effectiveness of guidance information for the deeper network. First, they extract the multi-scale guidance features in HR to enhance the quality of guidance features. Moreover, they utilize dense connection to maintain multi-scale guidance features, thus the depth features at each scale can be refined by guidance features from the current and all coarse scales.

The methods mentioned previously mainly fuse the multi-modal features by concatenation or multiplication, which tends to produce texture-copying artifacts as they treat the features from different modalities equally. To solve this problem, Zhong et al. [156] propose MMAF, a multi-modal attention based fusion module to select and fuse the multi-modal features in a learning manner. MMAF consists of a feature enhancement block and a feature recalibration block. The former is used to adaptively select informative feature from the guidance image, whereas the latter is designed to rescale the multi-modal features so that the heterogeneous features can be directly fused by naive operations. Song et al. [107] propose an attention-based iterative residual learning framework to deal with the real-world depth map super-resolution. This framework contains an iterative residual learning module to progressively learn the high-frequency component for the depth map and an channel attention strategy to let the network pay more attention to the informative features. Liu et al. [76] introduce an adaptive full-feature fusion module to enhance multi-scale feature representation ability. To fully exploit the complementary information contained in the depth and color image (i.e., enhance the edge information contained in the depth image and suppress the texture information in the color image), Yang et al. [140] propose CAC, a cross-domain attention conciliation module where the core is a parallel channel-spatial attention mechanism. Li et al. [68] propose CCGB, a correlation-controlled color guidance block to improve the accuracy of the guidance image and avoid texture-transfer and depth-bleeding artifacts. CCGB is made up of three major parts: correlation computing, weight learning, and guidance channel reweighting.

Multi-Task Learning. MTL, also known as learning with auxiliary task, is a sub-field of machine learning that aims to improve the performance of a specific task by simultaneously training multiple tasks. Motivated by this, some researchers propose to use the MTL framework to boost the performance of depth map super-resolution. For example, Tang et al. [113] propose a joint learning framework in which the tasks of depth map super-resolution and **Depth Estimation (DE)** are solved at the same time. To make these two tasks cooperate with each other, the authors propose two different guidance strategies. One is the high-frequency attention bridge, which aims to learn the high-frequency information of the DE task to guide the depth super-resolution process. The other is the content guidance bridge, which is designed to learn the content guidance from the depth super-resolution task to guide the DE task. In their framework, these two tasks are trained

separately with their corresponding loss functions. In a practical scenario, the paired RGB-D data may not always exist, especially for the aligned image pairs. In addition, processing HR guidance image will cause additional resource consumption. To solve these problems, Sun et al. [109] propose a novel cross-task distillation framework in which the guidance image is only required in the training stage. Similarly to Tang et al. [113], they also employ DE as an auxiliary task. Based on this, they further propose a cross-task interaction module to encourage the DE and depth map super-resolution tasks to learn from each other. Unlike Tang et al. [113], the authors propose to use knowledge distillation to realize bilateral knowledge transfer between these two tasks. In this framework, the roles of teacher and student are exchanged between these tasks based on their current performance for depth map recovery so that these two tasks can be tightly coupled and cooperated. When the model is trained, we can restore the LR depth map with the depth map super-resolution branch.

Besides DE, Yan et al. [136] propose to use depth completion as an auxiliary task to boost the performance of DSR. To make full use of the complementary information flows between these two tasks, they further design JCL, a novel joint correlation learning module, and IC, an iterative-cross module. The JCL is designed to capture long-range contextual dependencies, and it is made up of a lightweight pixel-wise correlation and a channel-wise correlation. The IC is an iterative module that aims to reuse complementary information flows between the output layers.

Prior Knowledge Guided Network Design. Over the past few years, deep learning based depth map super-resolution methods have achieved unprecedented success due to their powerful representation ability. Despite promising performance, the black box nature makes it difficult to interpret, and performance improvements usually come from stacking new modules at the expense of increasing the model complexity. To alleviate this problem, some works propose to integrate the deep models into traditional optimization algorithms. For example, Riegler et al. [90] combine the advantages of the deep convolution neural network and the variational method and propose a deep primal-dual network. In their framework, they first use a stack of convolution layers to generate a guidance and a roughly super-resolved depth map, then the obtained guidance and depth images are sent to a non-local variational model to reconstruct the final depth map. Zhou et al. [158] propose a maximal posterior estimation model for GDSR with a local implicit prior and a global implicit prior. The local implicit prior is designed for modeling the complex relationship between the reconstructed depth image and the HR guidance image from a local perspective, whereas the global implicit prior models the relationship between these two images from a global perspective. Moreover, they propose a persistent memory mechanism to improve the information representation ability in both image and feature spaces.

Marivani et al. [85] propose to use convolution sparse representation to represent different image modalities at the image level. In their model, different modalities have different convolution dictionaries but with the same sparse representation. Deng and Dragotti [19] propose to model each image modality with two convolution dictionaries, where the one is used to represent the common features and the other is used to model the unique features. De Lutio et al. [17] propose to use a deep neural network to learn the affinity graph, and the learned graph is implemented as a differentiable optimization layer that can be used as a regularizer for depth map upsampling. This method can inherit the merits of both optimization and deep learning based methods: the optimization layer ensures the fidelity, and the deep neural network makes the learned graph have a large context. Zhao et al. [153] propose to use a DCT to solve a optimization model for GDSR, then extend the solution of this optimization model to a deep neural network to solve the unexplainable problem of an existing empirical-designed neural network. In addition, they propose a shared/private feature extraction module to extract cross-modal features and a guided edge-attention module to alleviate texture over-transfer problem.

Novel Network Architecture. Designing an effective and efficient neural network architecture is a hot topic in deep learning, and in GDSR, researchers have tried numerous novel architectures to better exploit the intrinsic characteristics and correlations of multi-modal features. For example, Zhao et al. [152] propose CDcGAN, a color-depth conditional generative network to simultaneously super-resolve the LR depth and color images. To reduce the artifacts caused by the differences of distributions between the depth and its corresponding color image, Zuo et al. [159] propose a depth-guided affine transformation network in which depth-guided color feature filtering and color-guided depth feature refinement are performed iteratively to progressively enhance the network representation ability. Moreover, all the refined depth features are concatenated to make full use of the iterations. Based on deep edge-aware learning, Wang et al. [127] propose a novel GDSR framework to preserve the sharpness of the depth edges. Their model first learns the edge information of the depth map and then uses the learned edges to guide the depth reconstruction process. Ye et al. [143] propose a progressive multi-branch aggregation network that consists of a reconstruction branch, a multi-scale branch, and a guidance branch. The reconstruction branch is an attention-based error feed-forward/-back module, which is used to iteratively highlight the informative features. The multi-scale branch aims to learn multi-scale representation. The outputs of these three branches are sent to a fusion block, which is actually a channel attention mechanism to adaptively select and fuse the valuable features.

Zhong et al. [156] propose to use a bi-directional RNN to encourage the cooperation multi-level features. Tang et al. [113] propose a high-frequency attention bridge module and a content guidance bridge module to associate the tasks of depth super-resolution and DE. In another work, Tang et al. [112] introduce the implicit image function into GDSR to improve the model capability. Liu et al. [76] propose to use the deformable convolution [16] to align the depth features according to its corresponding guidance features. He et al. [47] propose a fast depth super-resolution baseline, where they use the octave convolution to fuse multi-modal features and reduce the computation complexity. Chen et al. [10] propose to decompose the high-frequency component from the guidance features to guide the depth map reconstruction. Li et al. [68] propose a correlation-controlled color guidance block, which aims to improve the accuracy of the guidance information. Yang et al. [140] propose CODON, a cross-domain orchestration network to effectively exploit the complementary information between depth and guidance features. Their model has two essential modules: RMC, a recursive multi-scale convolutional residual module, and CAC, a cross-domain attention conciliation module. RMC is used to address the scale variance of scene structures, whereas CAC is designed to effectively model co-occurrence correlations between depth and guidance features.

Although the preceding CNN-based methods have achieved significant success, there are still some limitations. First, existing CNN-based methods extract multi-modal features by applying cascaded convolutional layers on the input images. The local receptive field of convolutions makes it difficult for these methods to capture long-range spatial dependencies. Second, the weights of these convolutional layers are fixed during the inference process, which may hinder these methods from modeling the dynamic relationships between depth and guidance images. To mitigate the limitations mentioned earlier regarding CNN-based methods, some Transformer-based approaches are proposed in the area of GDSR. For example, Yao et al. [142] propose a texture-depth Transformer to explore useful texture information for depth super-resolution. Specifically, the authors first use a pretrained VGG model to extract multi-scale features. Then, the extract features are sent to the proposed texture-depth Transformer module to transfer the structures from the guidance features to the depth features. Finally, these enhanced features are used to generate the desired depth map. For depth map super-resolution, it is vital to maintain HR representations. However, the computational complexity of self-attention is quadratic to the size of the image. Directly applying the Transformer to the GDSR task would result in a great waste

of computational resources. To overcome this issue, Ariav and Cohen [2] propose a cascaded Transformer-based module for GDSR where they calculate self-attention in a local window and capture long-range dependencies with a shift operation. The complexity of this window-based attention is linear. Thus, their model is capable of handling large inputs.

Loss Function. In the deep learning based depth super-resolution methods, the most widely used loss functions are the pixel-wise $\mathcal{L}_1$ and $\mathcal{L}_2$ losses, since they result in lower MAE/RMSE values. The $\mathcal{L}_1$ and $\mathcal{L}_2$ losses are defined as follows:

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N}|\boldsymbol{X}_i - \boldsymbol{Y}_i^*|, \quad \mathcal{L}_2 = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{X}_i - \boldsymbol{Y}_i^*)^2, \tag{47}$$

where N is the number of pixels, $\boldsymbol{X}$ means the ground truth, and $\boldsymbol{Y}^*$ denotes the reconstructed depth map. However, since these pixel-wise losses treat every pixel equally, models trained with these functions tend to generate over-smooth results. Motivated by this, various loss functions are proposed to force their model to generate more realistic results. For example, Sun et al. [109] propose to use SSIM loss to improve the structure similarity between the reconstructed depth map and the ground truth. The SSIM loss is as follows:

$$\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(\boldsymbol{X}, \boldsymbol{Y}^*), \tag{48}$$

where SSIM means the SSIM metric, and it is defined in Equation (3). Zhao et al. [152] propose an eight-neighboring gradient difference loss to force the network to pay more attention to the boundaries, and the gradient difference loss is formulated as follows:

$$\mathcal{L}_{\text{GD}} = \frac{1}{N}\sum_{i}\left(\sum_{k\in\Omega}|\nabla_k\boldsymbol{X}_i - \nabla_k\boldsymbol{Y}_i^*|_1\right), \tag{49}$$

where Ω means the pixel's neighborhood, and ∇_k denotes the k-th gradient between each pixel and k-th pixels among eight-neighboring pixels. Dong et al. [26] propose a cycle-consistent loss for self-supervised depth map restoration:

$$\mathcal{L}_{\text{C}} = |f_{\text{down}}(f_{\text{net}}(\hat{\boldsymbol{Y}}, \boldsymbol{G})) - \hat{\boldsymbol{Y}}|_1, \tag{50}$$

where $\hat{\boldsymbol{Y}}$ means the interpolated LR depth map, f_{down} is average pooling downsampling, and f_{net} denotes the reconstruction network.

6 DEPTH-RELATED APPLICATIONS

In the following, we introduce depth-related application briefly, including salient object detection and semantic segmentation.

6.1 Salient Object Detection

RGB images contain rich color and content information, whereas depth images can provide supplementary depth information. In light of this, many works explore the potential of depth cue in salient object detection. To explore depth cue and inherit the performance RGB salient object detection, Cong et al. [14] develop a transformation model that helps existing RGB models work well in RGB-D scenarios. Considering that RGB and depth images belong to two different modalities, Zhang et al. [146] build RGB-induced and depth-induced enhancement modules to differentially model the dependence of two modalities. Cong et al. [15] couple interaction and refinement to explore cross-modality information, whereas Yang et al. [141] perform salient object detection in a bi-directional progressive manner. Zhang et al. [149] propose to extract intra- and inter-saliency cues that are used to promote salient object detection.

6.2 Semantic Segmentation

Depth images contain geometric information that is important for semantic segmentation, and depth image has been introduced into semantic segmentation. Chen et al. [9] propose a separation-and-aggregation gate to fuse multi-modal information and utilize bi-direction multi-step propagation to promote information interaction. Li et al. [67] design a collaborative optimization method to generate segmentation results. Zhou et al. [157] develop two encoders to extract information from RGB and depth images and carefully design a co-attention mechanism to fuse color and depth information. Shi et al. [101] propose to utilize depth and RGB images to generate 2D semantic segmentation results and then obtain refined 3D semantic segmentation results with 2D ones.

7 EXPERIMENTS

In this section, we conduct experiments to evaluate the performance of representative GDSR approaches. First, we introduce the datasets and evaluation metrics used in this article. Additionally, both the quantitative and qualitative results for the compared methods are presented. Based on these results, we analyze the advantages and disadvantages between different types of methods. Finally, we conduct experiments with several tricks widely used in other vision tasks and explore whether they can improve the performance of GDSR.

Datasets and Metrics. Three widely used datasets are selected, including the Middlebury dataset [51, 94–97], NYU v2 dataset [104], and RGB-D-D dataset [47]. For the Middlebury dataset, following the experimental settings of Guo et al. [40] and Hui et al. [54], we utilize 82 RGB-D image pairs collected by Hui et al. [54] as the training set, and 6 RGB-D image pairs from Middlebury 2005 [95] as the testing set. However, this dataset only contains 6 RGB-D image pairs for evaluation, which cannot fully demonstrated the performance of deep learning based methods. To this end, we further conduct experiments on a large-scale NYU v2 dataset [104] to compare recently proposed deep learning based methods. Specifically, following Kim et al. [58] and Li et al. [70], we use the first 1,000 RGB-D image pairs from the this dataset as the training data, and the testing data are (1) the rest of the 449 image pairs from the NYU v2 dataset [104], (2) 1,064 image pairs from the Sintel dataset [7], (3) 325 image pairs from the DIDOE indoor test set [117], (4) the first 500 image pairs from the SUN RGBD test set [106], (5) 405 image pairs from the RGB-D-D test set [47], and (6) 503 image pairs from the DIML indoor test set [12]. Additionally, we conduct experiments on the RGB-D-D dataset [47], and this is a newly proposed real-world dataset.

Among these dataset, the Middlebury and NYUv2 datasets are synthetic datasets, we use Bicubic interpolation to generate the LR depth maps, and the downsampling factors are set as 4×, 8×, and 16×. The RGB-D-D dataset is a real-world dataset that contains paired LR-HR depth maps, and we use the official split training/testing data to train and evaluate the GDSR models. For both datasets, we follow existing state-of-the-art methods [24, 47, 54, 70, 80] that employ RMSE values as the image quality assessment metric.

Experimental Results. We first show the experimental results on the Middlebury dataset [95]. For this dataset, we select 13 state-of-the-art methods as competitors, which includes two local filtering-based methods (i.e., GF [44] and SDF [148]), four optimization-based methods (i.e., TGV [91], AR [138], SRAM [121], and JGIE [80]), and seven learning-based methods (i.e., SRCNN [23], DJFR [70], DMSG [54], DepthSR [40], CGN [159], MFR-SR [162], and AHMF [156]). We set Bicubic interpolation as the baseline model. The quantitative results for 4×, 8×, and 16× depth map super-resolution are presented in Table 6. It can be observed that the learning-based methods obtain the best performance due to the powerful learning ability of the deep convolution neural network. The optimization-based methods are more robust than the filtering-based methods. The single depth map super-resolution method (i.e., SRCNN [95]) has a complex network architecture and achieves better performance than some traditional methods for the small up-scale factor.

Table 6. Quantitative Comparison with the State-of-the-Art Methods on Middlebury Datasets [95]

Method	Art			Book			Dools			Laundry			Mobeius			Reindeer		
	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×
Bicubic	3.87	5.46	8.17	1.27	2.34	3.34	1.31	1.86	2.62	2.06	3.45	5.07	1.33	1.97	2.85	2.42	3.99	5.86
GF [44]	3.78	5.41	8.04	1.29	2.26	3.34	1.30	1.85	2.61	2.01	3.40	5.01	1.32	1.83	2.79	2.32	3.88	5.86
SDF [148]	3.55	5.30	7.84	1.24	2.31	3.33	1.29	1.84	2.61	1.91	3.37	5.04	1.27	1.91	2.83	2.33	3.93	5.84
TGV [91]	3.34	4.10	6.43	1.47	1.82	2.63	1.31	1.61	2.22	2.39	2.64	4.17	1.22	1.64	2.41	2.71	3.15	4.60
AR [138]	2.80	3.65	5.90	1.54	1.82	2.92	1.38	1.66	2.34	2.22	2.69	4.57	1.01	1.46	2.53	2.45	2.96	4.11
SRAM [121]	2.57	3.20	4.87	1.33	1.46	2.51	1.07	1.19	1.90	2.00	2.11	4.07	0.85	1.10	1.98	2.07	2.47	3.44
JGIE [80]	2.02	3.38	5.58	0.93	1.57	2.47	0.97	1.27	1.76	1.33	2.00	2.37	0.97	1.42	2.19	1.39	2.23	3.56
SRCNN [23]	1.87	3.70	7.31	0.85	1.59	3.12	0.88	1.46	2.42	1.08	2.31	4.60	0.86	1.48	2.68	1.35	2.74	5.33
DJFR [70]	1.62	3.08	5.81	0.54	1.11	2.24	0.78	1.27	2.02	0.90	1.83	3.65	0.68	1.22	2.21	1.25	2.38	4.22
DMSG [54]	1.47	2.46	4.57	0.67	1.03	1.60	0.69	1.05	1.60	0.79	1.51	2.63	0.66	1.02	1.63	0.98	1.76	2.92
DepthSR [40]	1.20	2.22	**3.90**	0.60	0.89	1.51	0.84	1.14	**1.52**	0.78	1.31	2.26	0.96	1.19	1.58	0.96	1.57	**2.47**
CGN [159]	1.50	2.69	4.14	0.60	0.97	1.73	0.88	1.20	1.80	0.98	1.57	2.57	0.69	1.06	1.69	1.22	2.02	3.60
MFR-SR [162]	1.54	2.71	4.35	0.63	1.05	1.78	0.89	1.22	1.74	1.23	2.06	3.74	0.72	1.10	1.73	1.23	2.06	3.74
AHMF [156]	**1.09**	**2.14**	4.20	**0.38**	**0.72**	**1.49**	**0.62**	**1.03**	1.66	**0.64**	**1.22**	**2.14**	**0.54**	**0.88**	**1.53**	**0.85**	**1.56**	2.84

We use the RMSE metric (the lower the better). The best performance is shown in **bold**, whereas the second and the third best performance are represented by underscored and waved lines, respectively.

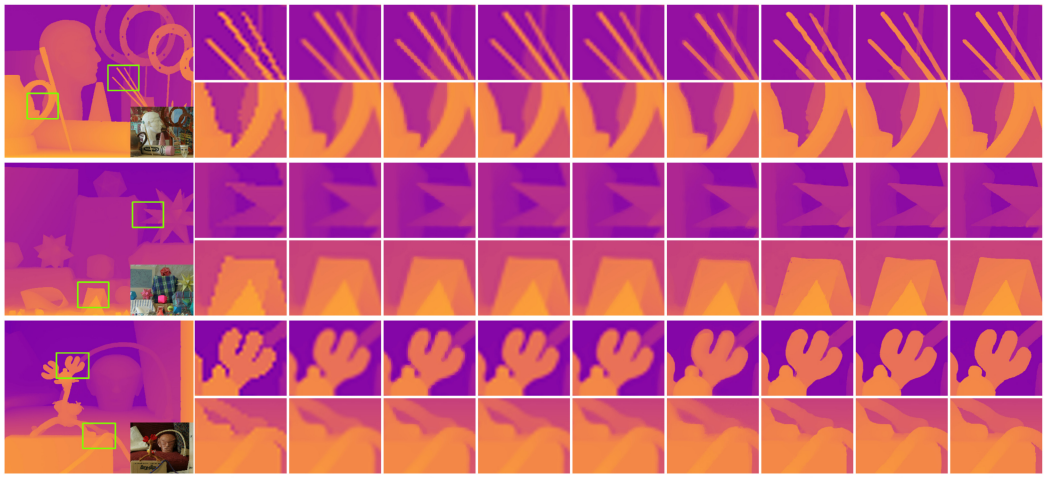


Fig. 2. Visual comparisons for 8× GDSR on *Art*, *Moebius*, and *Reindeer* from the Middlebury dataset: GT and RGB images (a), Bicubic (b), GF [44] (c), SDF [148] (d), TGV [91] (e), JGIE [80] (f), DJFR [70] (g), MFR-SR [162] (h), AHMF (i), GT (j). Please enlarge this Figure for more detail.

However, when the up-scale factor increases, the LR depth maps are seriously damaged, making it impossible to generate accurate depth maps. Thus, the performance of SRCNN [23] is worse than those of the compared methods. By introducing the guidance information, the performance of DJFR [70] is significantly improved over SRCNN [23]. Compared to DJFR [70], the subsequent methods (e.g., DepthSR [40] and CGN [159]) adopt powerful network architectures and elaborately designed guidance strategies, thus the depth map reconstruction performance is further improved.

Figure 2 shows the upsampled depth maps (8×) for *Art*, *Moebius*, and *Reindeer*. The visual comparisons are generally consistent with the quantitative results that the learning-based methods

Table 7. Quantitative Comparison with the State-of-the-Art Methods on NYU v2 [104], MPI Sintel Depth [7], DIDOE [117], SUN RGB-D [106], RGB-D-D [47], and DIML [12] Datasets for ×4, ×8, and ×16 Scale Factors, Respectively

Method	NYU v2			Sintel			DIDOE			SUN RGB			RGB-D-D			DIML		
	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×
Bicubic	4.27	7.17	11.37	6.41	8.79	11.90	5.67	8.41	10.01	1.23	2.01	3.19	2.75	4.47	6.98	1.92	3.20	5.14
GF [44]	5.05	6.97	11.13	7.15	8.66	11.63	7.80	8.78	9.98	1.75	2.18	3.21	2.72	4.02	6.68	2.72	3.40	5.13
DGF [130]	3.21	5.92	10.45	5.91	8.02	11.17	5.79	8.41	9.94	1.16	1.95	3.14	2.11	3.56	6.31	2.01	3.22	5.10
DMSG [54]	3.02	5.38	9.17	4.73	6.26	8.36	6.05	8.47	9.83	0.91	1.53	2.62	1.80	3.04	5.10	1.39	2.34	4.02
DJF [69]	2.80	5.33	9.46	5.30	7.53	10.41	6.04	8.29	10.15	0.89	1.65	2.83	1.72	2.96	5.26	1.39	2.49	4.27
DepthSR [40]	3.00	5.16	8.41	4.49	6.53	9.28	6.07	8.44	9.73	0.92	1.45	2.48	1.82	2.85	4.60	1.40	2.23	3.75
DJFR [70]	2.38	4.94	9.18	4.90	7.39	10.33	5.63	8.24	9.89	0.81	1.54	2.80	1.50	2.72	5.05	1.27	2.34	4.13
CUNet [19]	1.65	3.35	6.64	4.27	5.96	8.51	3.80	6.98	9.44	0.63	1.13	2.20	1.20	1.86	3.27	1.18	1.88	3.25
SVLRM [24]	1.51	3.21	6.98	4.05	5.83	8.60	3.58	6.96	9.55	0.59	1.10	2.33	1.22	1.88	3.55	1.19	1.93	3.49
FDKN [58]	1.86	3.58	6.96	4.23	5.92	8.73	4.04	7.07	9.50	0.67	1.13	2.23	1.18	1.91	3.41	1.13	1.84	3.29
DKN [58]	1.62	3.26	6.51	4.38	5.89	8.40	3.49	6.96	9.31	0.63	1.10	2.16	1.31	1.87	3.26	1.27	1.86	3.22
GraphSR [17]	1.79	3.04	6.02	4.29	5.56	7.93	3.93	6.37	9.06	0.67	1.05	2.03	1.30	1.83	3.12	1.25	1.79	3.03
FDSR [47]	1.61	3.18	5.86	4.14	5.67	7.86	3.62	6.54	8.85	0.64	1.05	1.97	1.16	1.82	3.06	**1.10**	1.71	2.87
AHMF [156]	1.40	2.89	5.64	3.84	5.62	7.55	**2.93**	**6.14**	**8.54**	0.57	0.99	1.82	**1.10**	**1.73**	3.04	**1.10**	**1.70**	**2.83**
JIIF [112]	**1.37**	**2.76**	**5.27**	**3.82**	**5.50**	**7.46**	2.94	6.17	8.58	**0.54**	**0.95**	**1.79**	1.15	1.77	**2.79**	1.17	1.79	2.86

We use the RMSE metric (the lower the better). The best performance is shown in **bold**, whereas the second and the third best performance are represented by underscored and waved lines, respectively.

achieve better performance than the conventional handcrafted methods. GF [44] suffers from halo artifacts due to its local characteristic. The results of SDF [148] suffer from jagging artifacts at the pencial region. The results of TGV [91] are too smooth, which means that the hand-designed regularizer cannot deal with a large up-scale factor. JGIE [80] combines both the local structure and non-local low-rank regularization, and obtains better performance than TGV [91], which implies that combining multiple regularizers is beneficial for GDSR. The results of DJFR [70] are more faithful to the ground-truth image. However, they still suffer from texture copying artifacts in some regions, since DJFR [70] uses a naive method to fuse the guidance information. AHMF [156] proposes an attention-based fuse mechanism to fuse the guidance information, achieving the best performance.

We mainly use the NYU v2 [104] and RGB-D-D datasets [47] to compare the recently proposed learning-based methods, which include GF [44], DGF [130], DMSG [54], DJF [69], DJFR [70], DepthSR [40], CUNet [19], SVLRM [24], JIIF [112], FDKN [58], DKN [58], AHMF [156], FDSR [47], DCTNet [153], and GraphSR [17]. The RMSE values for different methods are presented in Table 7 and Table 8. From these tables, we can see that the learning-based methods achieve significantly better performance than traditional GDSR models, which demonstrates the powerful learning ability of deep neural networks. In addition to quantitative comparisons, we present computation complexity comparison in Figure 3 and show the qualitative results of different methods in Figure 4. As can be seen, the results of GF [44] are over-smooth. To improve the performance of GF [44], Wu et al. [130] propose to utilize two networks to modify the guidance and LR depth images. The modified guidance and depth images are more suitable than the original ones. However, DGF [130] still cannot generate satisfactory results, since it still follows the basic framework of GF [44]. For comparison, the deep learning based methods can produce results with sharper boundaries and fewer artifacts. From Figure 4 and Table 7, we can also find that the RMSE values do not always consist of subjective assessment results. Thus, it is of vital importance to develop a more reasonable evaluation metric for the GDSR task. Finally, we report the parameter number, maximum GPU

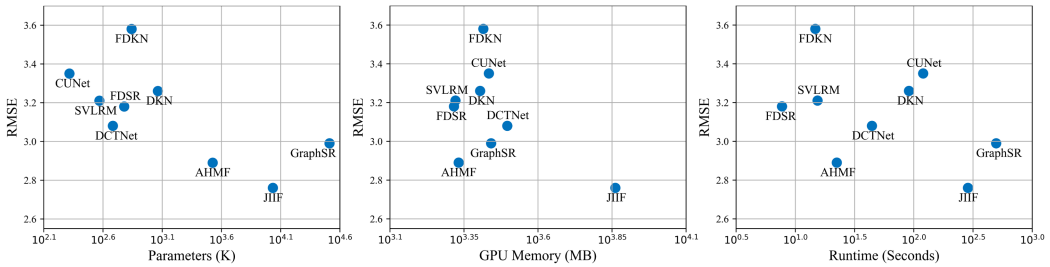


Fig. 3. Comparison of the state of the art for 8× GDSR on the NYU v2 [104] dataset in terms of RMSE, number of parameters (a), peak GPU memory consumption (b), and average running times (c). The experiments are evaluated on an NVIDIA 3090 GPU with HR depth map size 480 × 640.

Table 8. Quantitative Comparison with the State-of-the-Art Methods on the RGB-D-D [47] Dataset

Method	Bicubic	GF [44]	DJFR [70]	SVLRM [24]	DKN [58]	FDSR [47]	AHMF [156]	GraphSR [17]	DCTNet [153]
RMSE	7.19	6.97	6.24	5.81	5.74	5.49	5.46	5.52	**5.38**

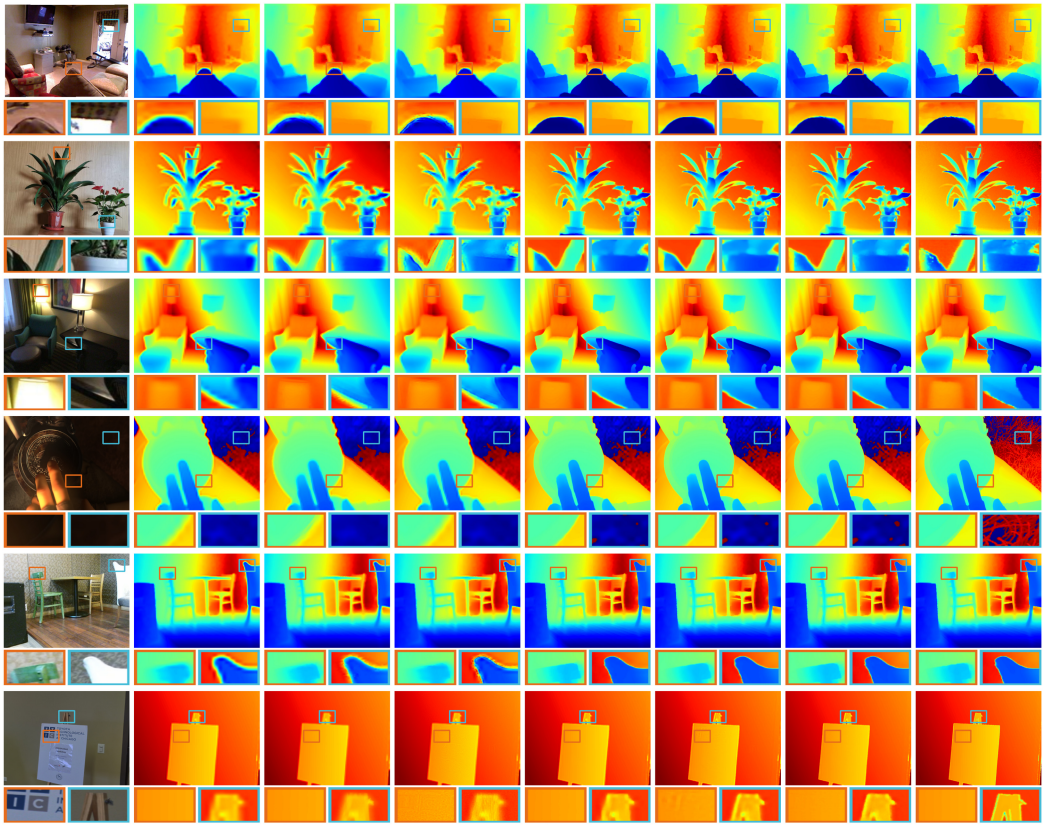
We use the RMSE metric (the lower the better). The best performance is shown in **bold**, whereas the second and the third best performance are represented by underscored and waved lines, respectively.

memory consumption, and average running times of different methods in Figure 3. As can be seen, models with more parameters can obtain better performance, although at the expense of high computational complexity. These models may not be suitable for real-world scenarios, especially for resource-limited devices such as mobile phones. Therefore, it is also important to design light-weight and efficient models for the GDSR task.

Bag of Tricks for GDSR. Recently, with the development of deep convolution neural networks, a lot of deep learning-based methods have been proposed and have dominated the area of GDSR. By reviewing these methods, we find that most of them pay more attention to designing effective network architectures, and only a few works concern the training strategies that have been shown to be important for various tasks, such as image classification [48] and image super-resolution [74]. To this end, we survey several tricks widely used in other vision tasks and explore whether they can improve the performance of GDSR. Specifically, we use the recently proposed AHMF [156] as the baseline model and perform the following strategies one by one: (1) use geometric self-ensemble [73] during inference; (2) employ mix-up as the data augmentation method; (3) replace PReLU [45] with SiLU [32, 49] activation; (4): during training, increase the patch size from 256×256 to 384×384 to enlarge the input *field of view*; (5) augment the guidance image by randomly change brightness, contrast, saturation, and hue; and (6) use cosine annealing learning rate decay to adjust the learning rate. The ablation studies of different training strategies are presented in Table 9. We can observe that almost all the tricks contribute to the final performance, and by stacking all the tricks, the final model ("+All of the above"in Table 9) can even outperform the current SOAT method JIIF [112].

8 DISCUSSION AND CONCLUSION

In this article, we presented, to the best of our knowledge, the first comprehensive overview of GDSR methods. We first reviewed the widely used datasets and evaluation metrics for GDSR. Then, we roughly classified existing GDSR methods into three categories: filtering-based methods, prior-based methods, and learning-based methods. In each category, we investigated representative and milestone methods and discussed their contributions, benefits, and weaknesses. Moreover, based



(a) Guidance (b) Bicubic (c) GF [44] (d) DGF [130] (e) GraphSR [17] (f) AHMF [156] (g) JIIF [112] (h) GT

Fig. 4. Visual comparisons for 8× GDSR. (Top to bottom) Each row shows the super-resolved depth maps on the NYU v2 [104], RGB-D-D [47], SUN RGB-D [106], MPI Sintel Depth [7], DIML [12], and DIDOE [117] datasets, respectively. Please enlarge this figure for more details.

Table 9. Ablation Studies of Different Training Strategies on the NYU v2 Dataset [104] for 8× Depth Map Super-Resolution

Configuration	NYU v2		Sintel		DIDOE		SUN RGB		RGB-D-D		DIML	
	RMSE	Δ	RMSE	Δ	RMSE	Δ	RMSE	Δ	RMSE	Δ	RMSE	Δ
Baseline	2.887	–	5.624	–	6.143	–	0.985	–	1.733	–	1.699	–
+Self-ensemble	2.827	−0.060	5.577	−0.027	6.088	−0.055	0.954	-0.031	1.705	-0.028	1.677	-0.022
+MixUp	2.847	−0.030	5.592	−0.032	6.185	+0.042	0.945	−0.040	**1.695**	**−0.038**	1.672	−0.027
+SiLU activation	2.883	−0.004	5.616	−0.008	6.080	−0.063	0.984	−0.001	1.741	+0.008	1.685	−0.014
+Large patches	2.825	-0.062	5.613	−0.011	6.059	−0.084	0.981	−0.004	1.756	+0.023	1.683	−0.016
+Color augmentation	2.892	+0.005	5.594	-0.030	6.098	−0.045	0.983	−0.002	1.730	−0.003	1.687	−0.012
+Cosine annealing	2.809	−0.078	5.612	−0.012	6.090	−0.053	0.968	−0.017	1.728	−0.005	1.685	−0.014
+All of the above	**2.731**	**−0.156**	**5.558**	**−0.066**	**5.965**	**−0.178**	**0.916**	**−0.069**	1.701	−0.032	**1.628**	**−0.071**

We use AHMF [156] as the baseline. For the RMSE, lower values means better performance. The best performance is shown in **bold**, whereas the second and the third best performance are represented by underscored and waved lines, respectively.

on the unified experimental configurations, we provided a comprehensive evaluation of existing state-of-the-art GDSR methods.

It is true that the methods mentioned previously have achieved gratifying performance and have significantly promoted the development of GDSR. However, we cannot neglect that there are still some challenging problems in GDSR. In the following, we list some of the challenges and present prospects for future research.

Lightweight Models. Some newly released smartphones are equipped with a range sensor to capture depth maps for face recognition, AR games, and so forth. Therefore, an accurate and efficient model is highly desirable to improve the quality of captured depth maps. However, current state-of-the-art GDSR methods are designed to be complicated to increase the model capacity for higher performance, making them hard to apply directly on the mobile device. To address this problem, model compression and neural architecture search can be adopted to learn a compact GDSR model with promising reconstruction accuracy.

Evaluation Metrics. Evaluation metrics play a fundamental role in computer vision tasks. First, the choice of objective functions is typically influenced by the evaluation metric. For instance, $\mathcal{L}_1$ and $\mathcal{L}_2$ losses have been extensively used in GDSR, as they are particularly relevant with the MAE and RMSE metrics. Newly proposed algorithms are needed to compare with existing methods using these metrics. However, previous studies [57, 65] show that these metrics cannot accurately reflect the visual quality of the restored image. In addition, MAE and RMSE are full-reference image quality assessment scores, which cannot be used in real-world applications. Hence, it is an incredibly urgent and important task to develop appropriate evaluation metrics for GDSR.

Scale Arbitrary GDSR. Depth information has been widely used in a variety of tasks. Therefore, it is desirable to develop a GDSR model that can be adopted to upsample a given LR depth map to any scale. Current GDSR methods can only be applied to one or a limited number of integer upsampling factors. Recently, the implicit neural representation technique has been employed in single-image super-resolution [11] for arbitrary-scale super-resolution. The question of how to design a framework that incorporates the implicit neural representation technique with GDSR models may be a valuable topic for future research.

Interact with High-Level Tasks. In most cases, the purpose of GDSR is not only to restore an HR depth map with lower MAE values, but it is our hope that the reconstructed depth map can facilitate the relevant high-level tasks, such as image segmentation and scene understanding. Therefore, we suggest employing the accuracy of high-level tasks as evaluation metrics to evaluate the performance of the GDSR method. Moreover, we can design a hybrid model that combines the GDSR model with other high-level tasks to make them learn from each other.

Weak/Unsupervised Learning. It is known to all that deep learning based approaches do not perform well in real-world cases. This is mainly because most of these methods require pairs of LR and HR for training. However, the degradation process in the real-world scenario is so complex that it cannot be simulated by Bicubic or Nearest downsampling methods, which leads to a big gap between these synthetic depth maps and real-world LR depth maps. In addition, it is time consuming to develop a large-scale dataset that contains real LR and HR depth maps for GDSR models. Although some weak/unsupervised GDSR methods are proposed, the performance of them is still worse than that of supervised methods. Therefore, employing the weak/unsupervised learning methods for GDSR appears to be a promising research direction.

RGB-D Datasets. Generally, the paired RGB-D data play an important role in the GDSR task, especially for the learning methods. In recent years, several RGB-D datasets have been released, greatly promoting the development of the GDSR algorithm. However, compared with the dataset for other computer vision tasks, such as object detection and classification, the size of these datasets is very small. In addition, most of the RGB-D datasets are captured from indoor scenes. Therefore,

it is of vital importance to construct large-scale and realistic RGB-D datasets that can serve as benchmark datasets for future research.

REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein. 2006. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing* 54, 11 (2006), 4311–4322.

[2] I. Ariav and I. Cohen. 2022. Depth map super-resolution via cascaded transformers guidance. *Frontiers in Signal Processing* 2 (2022), Article 847890, 12 pages.

[3] S. Bae, S. Parisand, and F. Durand. 2006. Two-scale tone management for photographic look. *ACM Transactions on Graphics* 25, 3 (2006), 637–645.

[4] K. Bredies, K. Kunisch, and T. Pock. 2010. Total generalized variation. *SIAM Journal on Imaging Sciences* 3, 3 (2010), 492–526.

[5] H. Bristow, A. Eriksson, and S. Lucey. 2013. Fast convolutional sparse coding. In *Proceedings of CVPR*. 391–398.

[6] A. Buades, B. Coll, and J.-M. Morel. 2005. A non-local algorithm for image denoising. In *Proceedings of CVPR*, Vol. 2. IEEE, Los Alamitos, CA, 60–65.

[7] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. 2012. A naturalistic open source movie for optical flow evaluation. In *Proceedings of ECCV*. 611–625.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders. 2001. Atomic decomposition by basis pursuit. *SIAM Review* 43, 1 (2001), 129–159.

[9] X. Chen, K. Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng. 2020. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In *Proceedings of ECCV*. 561–577.

[10] Y. Chen, H. Fan, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, S. Yan, and J. Feng. 2019. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of ICCV*. 3435–3444.

[11] Y. Chen, S. Liu, and X. Wang. 2021. Learning continuous image representation with local implicit image function. In *Proceedings of CVPR*. 8628–8638.

[12] J. Cho, D. Min, Y. Kim, and K. Sohn. 2021. Deep monocular depth estimation leveraging a large-scale outdoor stereo dataset. *Expert Systems with Applications* 178 (2021), 114877.

[13] P. Clifford. 1990. Markov random fields in statistics. In *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, Geoffrey Grimmett and Dominic Welsh (Eds.). Oxford University Press, 19–32.

[14] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong. 2020. Going from RGB to RGBD saliency: A depth-guided transformation model. *IEEE Transactions on Cybernetics* 50, 8 (2020), 3627–3639.

[15] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao. 2022. CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection. *IEEE Transactions on Image Processing* 31 (2022), 6800–6815.

[16] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. 2017. Deformable convolutional networks. In *Proceedings of ICCV*. 764–773.

[17] R. de Lutio, A. Becker, S. D'Aronco, S. Russo, J. D. Wegner, and K. Schindler. 2022. Learning graph regularisation for guided super-resolution. In *Proceedings of CVPR*. 1979–1988.

[18] X. Deng and P. L. Dragotti. 2019. Deep coupled ISTA network for multi-modal image super-resolution. *IEEE Transactions on Image Processing* 29 (2019), 1683–1698.

[19] X. Deng and P. L. Dragotti. 2020. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 10 (2020), 3333–3348.

[20] X. Deng and P. L. Dragotti. 2020. Deep coupled ISTA network for multi-modal image super-resolution. *IEEE Transactions on Image Processing* 29 (2020), 1683–1698.

[21] X. Deng, P. Song, M. R. D. Rodrigues, and P. L. Dragotti. 2020. RADAR: Robust algorithm for depth image super resolution based on FRI theory and multimodal dictionary learning. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 8 (2020), 2447–2462.

[22] J. Diebel and S. Thrun. 2005. An application of Markov random fields to range sensing. In *Advances in Neural Information Processing Systems*. 291–298.

[23] C. Dong, C. C. Loy, K. He, and X. Tang. 2015. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 2 (2015), 295–307.

[24] J. Dong, J. Pan, J. S. Ren, L. Lin, J. Tang, and M. H. Yang. 2021. Learning spatially variant linear representation models for joint filtering. In *Proceedings of CVPR*.

[25] W. Dong, G. Shi, X. Li, K. Peng, J. Wu, and Z. Guo. 2017. Color-guided depth recovery via joint local structural and nonlocal low-rank regularization. *IEEE Transactions on Multimedia* 19, 2 (2017), 293–301.

[26] X. Dong, N. Yokoya, L. Wang, and T. Uezato. 2022. Learning mutual modulation for self-supervised cross-modal super-resolution. In *Proceedings of ECCV*. 1–18.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, Mostafa Dehghani, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*.
[28] L. Du, X. Ye, X. Tan, E. Johns, B. Chen, E. Ding, X. Xue, and J. Feng. 2022. AGO-Net: Association-guided 3D point cloud object detection network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2022), 8097–8109.
[29] F. Durand and J. Dorsey. 2002. Fast bilateral filtering for the display of high-dynamic-range images. In *Proceedings of CGIT*. 257–266.
[30] I. Eichhardt, D. Chetverikov, and Z. Janko. 2017. Image-guided ToF depth upsampling: A survey. *Machine Vision and Applications* 28, 3 (2017), 267–282.
[31] E. Eisemann and F. Durand. 2004. Flash photography enhancement via intrinsic relighting. *ACM Transactions on Graphics* 23, 3 (2004), 673–678.
[32] S. Elfwing, E. Uchibe, and K. Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks* 107 (2018), 3–11.
[33] K. Engan, S. O. Aase, and J. Hakon Husoy. 1999. Method of optimal directions for frame design. In *Proceedings of ICASSP*, Vol. 5. 2443–2446.
[34] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski. 2008. Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Transactions on Graphics* 27, 3 (2008), 1–10.
[35] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof. 2013. Image guided depth upsampling using anisotropic total generalized variation. In *Proceedings of ICCV*. 993–1000.
[36] F. Gao, X. Deng, M. Xu, J. Xu, and P. L. Dragotti. 2022. Multi-modal convolutional dictionary learning. *IEEE Transactions on Image Processing* 31 (2022), 1325–1339.
[37] S. Gu, S. Guo, W. Zuo, Y. Chen, R. Timofte, L. Van Gool, and L. Zhang. 2019. Learned dynamic guidance for depth image reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 10 (2019), 2437–2452.
[38] S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen, and L. Zhang. 2017. Learning dynamic guidance for depth image enhancement. In *Proceedings of CVPR*. 3769–3778.
[39] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang. 2015. Convolutional sparse coding for image super-resolution. In *Proceedings of ICCV*. 1823–1831.
[40] C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han. 2019. Hierarchical features driven residual learning for depth map super-resolution. *IEEE Transactions on Image Processing* 28, 5 (2019), 2545–2557.
[41] X. Guo, Y. Li, J. Ma, and H. Ling. 2020. Mutually guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 3 (2020), 694–707.
[42] B. Ham, M. Cho, and J. Ponce. 2018. Robust guided image filtering using nonconvex potentials. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 1 (2018), 192–207.
[43] X. F. Han, H. Laga, and M. Bennamoun. 2019. Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 5 (2019), 1578–1604.
[44] K. He, J. Sun, and X. Tang. 2010. Guided image filtering. In *Proceedings of ECCV*. 1–14.
[45] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of ICCV*. 1026–1034.
[46] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*. 770–778.
[47] L. He, H. Zhu, F. Li, H. Bai, R. Cong, C. Zhang, C. Lin, M. Liu, and Y. Zhao. 2021. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *Proceedings of CVPR*. 9229–9238.
[48] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li. 2019. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of CVPR*. 558–567.
[49] D. Hendrycks and K. Gimpel. 2016. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415* (2016).
[50] D. Herrera, J. Kannala, and J. Heikkilä. 2012. Joint depth and color camera calibration with distortion correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 10 (2012), 2058–2064.
[51] H. Hirschmuller and D. Scharstein. 2007. Evaluation of cost functions for stereo matching. In *Proceedings of CVPR*. IEEE, Los Alamitos, CA, 1–8.
[52] P. L. Huber. 1973. Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics* 1, 5 (1973), 799–821.
[53] B. Huhle, T. Schairer, P. Jenke, and W. Straßer. 2010. Fusion of range and color images for denoising and resolution enhancement with a non-local filter. *Computer Vision and Image Understanding* 114, 12 (2010), 1336–1345.
[54] T. W. Hui, C. C. Loy, and X. Tang. 2016. Depth map super-resolution by deep multi-scale guidance. In *Proceedings of ECCV*. 353–369.
[55] R. J. Jevnisek and S. Avidan. 2017. Co-occurrence filter. In *Proceedings of CVPR*. 3184–3192.
[56] Z. Jiang, Y. Hou, H. Yue, J. Yang, and C. Hou. 2018. Depth super-resolution from RGB-D pairs with transform and spatial domain regularization. *IEEE Transactions on Image Processing* 27, 5 (2018), 2587–2602.

[57] J. Johnson, A. Alahi, and L. Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of ECCV*. 694–711.
[58] B. Kim, J. Ponce, and B. Ham. 2021. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision* 129, 2 (2021), 579–600.
[59] S. Kim, D. Min, B. Ham, S. Kim, and K. Sohn. 2017. Deep stereo confidence prediction for depth estimation. In *Proceedings of ICIP*. IEEE, Los Alamitos, CA, 992–996.
[60] Y. Kim, B. Ham, C. Oh, and K. Sohn. 2016. Structure selective depth superresolution for RGB-D cameras. *IEEE Transactions on Image Processing* 25, 11 (2016), 5227–5238.
[61] Y. Kim, H. Jung, D. Min, and K. Sohn. 2018. Deep monocular depth estimation via integration of global and local predictions. *IEEE Transactions on Image Processing* 27, 8 (2018), 4131–4144.
[62] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele. 2007. Joint bilateral upsampling. *ACM Transactions on Graphics* 26, 3 (2007), 96–es.
[63] F. Kou, W. Chen, C. Wen, and Z. Li. 2015. Gradient domain guided image filtering. *IEEE Transactions on Image Processing* 24, 11 (2015), 4528–4539.
[64] R. Lange and P. Seitz. 2001. Solid-state time-of-flight range camera. *IEEE Journal of Quantum Electronics* 37, 3 (2001), 390–397.
[65] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of CVPR*. 4681–4690.
[66] H. Lee, A. Battle, R. Raina, and A. Ng. 2006. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, Bernhard Scholkopf, John Platt, and Thomas Hoffman (Eds.). MIT Press, Cambridge, MA, 801–808.
[67] H. Li, R. Cong, S. Kwong, C. Chen, Q. Xu, and C. Li. 2021. Stereo superpixel: An iterative framework based on parallax consistency and collaborative optimization. *Information Sciences* 556 (2021), 209–222.
[68] T. Li, H. Lin, X. Dong, and X. Zhang. 2020. Depth image super-resolution using correlation-controlled color guidance and multi-scale symmetric network. *Pattern Recognition* 107 (2020), 107513.
[69] Y. Li, J. B. Huang, N. Ahuja, and M. H. Yang. 2016. Deep joint image filtering. In *Proceedings of ECCV*. 154–169.
[70] Y. Li, J. B. Huang, N. Ahuja, and M. H. Yang. 2019. Joint image filtering with deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 8 (2019), 1909–1923.
[71] Y. Li, L. Zhang, Y. Zhang, H. Xuan, and Q. Dai. 2014. Depth map super-resolution via iterative joint-trilateral-upsampling. In *Proceedings of VCIP*. 386–389.
[72] Z. Li, J. Zheng, Z. Zhu, W. Yao, and S. Wu. 2014. Weighted guided image filtering. *IEEE Transactions on Image Processing* 24, 1 (2014), 120–129.
[73] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of CVPRW*. 136–144.
[74] Z. Lin, P. Garg, A. Banerjee, S. A. Magid, D. Sun, Y. Zhang, L. Van Gool, D. Wei, and H. Pfister. 2022. Revisiting RCAN: Improved training for image super-resolution. *arXiv preprint arXiv:2201.11279* (2022).
[75] M. Y. Liu, O. Tuzel, and Y. Taguchi. 2013. Joint geodesic upsampling of depth images. In *Proceedings of CVPR*. 169–176.
[76] P. Liu, Z. Zhang, Z. Meng, and N. Gao. 2021. Deformable enhancement and adaptive fusion for depth map super-resolution. *IEEE Signal Processing Letters* 29 (2021), 204–208.
[77] W. Liu, X. Chen, J. Yang, and Q. Wu. 2017. Robust color guided depth map restoration. *IEEE Transactions on Image Processing* 26, 1 (2017), 315–327.
[78] W. Liu, P. Zhang, Y. Lei, X. Huang, J. Yang, and M. K. P. Ng. 2021. A generalized framework for edge-preserving and structure-preserving image smoothing. *arXiv:2107.07058* (2021).
[79] X. Liu, D. Zhai, R. Chen, X. Ji, D. Zhao, and W. Gao. 2019. Depth restoration from RGB-D data via joint adaptive regularization and thresholding on manifolds. *IEEE Transactions on Image Processing* 28, 3 (2019), 1068–1079.
[80] X. Liu, D. Zhai, R. Chen, X. Ji, D. Zhao, and W. Gao. 2019. Depth super-resolution via joint color-guided internal and external regularizations. *IEEE Transactions on Image Processing* 28, 4 (2019), 1636–1645.
[81] K. H. Lo, K. L. Hua, and Y. C. F. Wang. 2013. Depth map super-resolution via Markov random fields without texture-copying artifacts. In *Proceedings of ICASSP*. 1414–1418.
[82] K. H. Lo, Y. C. F. Wang, and K. L. Hua. 2018. Edge-preserving depth map upsampling by joint trilateral filter. *IEEE Transactions on Cybernetics* 48, 1 (2018), 371–384.
[83] J. Lu, D. Min, R. S. Pahwa, and M. N. Do. 2011. A revisit to MRF-based depth map super-resolution and enhancement. In *Proceedings of ICASSP*. IEEE, Los Alamitos, CA, 985–988.
[84] S. Lu, X. Ren, and F. Liu. 2014. Depth enhancement via low-rank matrix completion. In *Proceedings of CVPR*. 3390–3397.
[85] I. Marivani, E. Tsiligianni, B. Cornelis, and N. Deligiannis. 2020. Multimodal deep unfolding for guided image super-resolution. *IEEE Transactions on Image Processing* 29 (2020), 8443–8456.

[86] K. Morishita, S. Yamagata, T. Okabe, T. Yokoyama, and K. Hamatani. 1988. Unsharp masking for image enhancement. (Dec. 27, 1988). US Patent 4,794,531.
[87] S. Peng, C. Jiang, Y. Liao, M. Niemeyer, M. Pollefeys, and A. Geiger. 2020. Shape as points: A differentiable poisson solver. In *Proceedings of NeurIPS'20*.
[88] Yiguo Qiao, Licheng Jiao, Wenbin Li, Christian Richardt, and Darren Cosker. 2021. Fast, high-quality hierarchical depth-map super-resolution. In *Proceedings of ACM MM*. 4444–4453.
[89] Y. Qiao, L. Jiao, S. Yang, and B. Hou. 2018. A novel segmentation based depth map up-sampling. *IEEE Transactions on Multimedia* 21, 1 (2018), 1–14.
[90] G. Riegler, D. Ferstl, M. Rüther, and H. Bischof. 2016. A deep primal-dual network for guided depth super-resolution. *arXiv preprint arXiv:1607.08569* (2016).
[91] G. Riegler, M. Rüther, and H. Bischof. 2016. ATGV-Net: Accurate depth super-resolution. In *Proceedings of ECCV*. 268–284.
[92] L. I. Rudin, S. Osher, and E. Fatemi. 1992. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena* 60, 1-4 (1992), 259–268.
[93] F. Santosa and W. W. Symes. 1986. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific Computing* 7, 4 (1986), 1307–1330.
[94] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In *Proceedings of GCPR*. 31–42.
[95] D. Scharstein and C. Pal. 2007. Learning conditional random fields for stereo. In *Proceedings of CVPR*. IEEE, Los Alamitos, CA, 1–8.
[96] D. Scharstein and R. Szeliski. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision* 47, 1 (2002), 7–42.
[97] Daniel Scharstein and Richard Szeliski. 2003. High-accuracy stereo depth maps using structured light. In *Proceedings of CVPR*, Vol. 1. IEEE, Los Alamitos, CA.
[98] G. Shacht, D. Danon, S. Fogel, and D. Cohen-Or. 2021. Single pair cross-modality super resolution. In *Proceedings of CVPR*. 6378–6387.
[99] K. Shankar, M. Tjersland, J. Ma, K. Stone, and M. Bajracharya. 2022. A learned stereo depth system for robotic manipulation in homes. *IEEE Robotics and Automation Letters* (2022).
[100] X. Shen, Z. Chao, X. Li, and J. Jia. 2017. Mutual-structure for joint filtering. *International Journal of Computer Vision* 125, 3 (2017), 1–15.
[101] W. Shi, J. Xu, D. Zhu, G. Zhang, X. Wang, J. Li, and X. Zhang. 2022. RGB-D semantic segmentation and label-oriented voxelgrid fusion for accurate 3D semantic mapping. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 1 (2022), 183–197.
[102] Z. Shi, Y. Chen, E. Gavves, P. Mettes, and C. G. Snoek. 2021. Unsharp mask guided filtering. *IEEE Transactions on Image Processing* 30 (2021), 7472–7485.
[103] N. Silberman and R. Fergus. 2011. Indoor scene segmentation using a structured light sensor. In *Proceedings of ICCVW*. IEEE, Los Alamitos, CA, 601–608.
[104] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. 2012. Indoor segmentation and support inference from RGBD images. In *Proceedings of ECCV*. 746–760.
[105] P. Song, X. Deng, J. F. C. Mota, N. Deligiannis, P. L. Dragotti, and M. R. D. Rodrigues. 2019. Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries. *IEEE Transactions on Computational Imaging* 6 (2019), 57–72.
[106] S. Song, S. P. Lichtenberg, and J. Xiao. 2015. Sun RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of CVPR*. 567–576.
[107] X. Song, Y. Dai, D. Zhou, L. Liu, Wei W. Li, H. Li, and R. Yang. 2020. Channel attention based iterative residual learning for depth map super-resolution. In *Proceedings of CVPR*. 5631–5640.
[108] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of ICCV*. 7262–7272.
[109] B. Sun, X. Ye, B. Li, H. Li, Z. Wang, and R. Xu. 2021. Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution. In *Proceedings of CVPR*. 7792–7801.
[110] Z. Sun, B. Han, J. Li, J. Zhang, and X. Gao. 2020. Weighted guided image filtering with steering kernel. *IEEE Transactions on Image Processing* 29 (2020), 500–508.
[111] H. Takeda, S. Farsiu, and P. Milanfar. 2007. Kernel regression for image processing and reconstruction. *IEEE Transactions on Image Processing* 16, 2 (2007), 349–366.
[112] J. Tang, X. Chen, and G. Zeng. 2021. Joint implicit image function for guided depth super-resolution. In *Proceedings of ACM MM*. 4390–4399.

[113] Q. Tang, R. Cong, R. Sheng, L. He, D. Zhang, Y. Zhao, and S. Kwong. 2021. BridgeNet: A joint learning network of depth map super-resolution and monocular depth estimation. In *Proceedings of ACM MM*. 2148–2157.
[114] C. Tomasi and R. Manduchi. 1998. Bilateral filtering for gray and color images. In *Proceedings of ICCV*. IEEE, Los Alamitos, CA, 839–846.
[115] I. Tosic and S. Drewes. 2014. Learning joint intensity-depth sparse representations. *IEEE Transactions on Image Processing* 23, 5 (2014), 2122–2132.
[116] J. A. Tropp and A. C. Gilbert. 2007. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory* 53, 12 (2007), 4655–4666.
[117] I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, et al. 2019. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463* (2019).
[118] F. Wang, J. Pan, S. Xu, and J. Tang. 2022. Learning discriminative cross-modality features for RGB-D saliency detection. *IEEE Transactions on Image Processing* 31 (2022), 1285–1297.
[119] H. Wang, R. Fan, P. Cai, and M. Liu. 2021. PVStereo: Pyramid voting module for end-to-end self-supervised stereo matching. *IEEE Robotics and Automation Letters* 6, 3 (2021), 4353–4360.
[120] J. Wang, L. Sun, R. Xiong, Y. Shi, Q. Zhu, and B. Yin. 2022. Depth map super-resolution based on dual normal-depth regularization and graph Laplacian prior. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 6 (2022), 3304–3318.
[121] J. Wang, W. Xu, J. F. Cai, Q. Zhu, Y. Shi, and B. Yin. 2019. Multi-direction dictionary learning based depth map super-resolution with autoregressive modeling. *IEEE Transactions on Multimedia* 22, 6 (2019), 1470–1484.
[122] L. Wang, L. Zhang, Y. Zhu, Z. Zhang, T. He, M. Li, and X. Xue. 2021. Progressive coordinate transforms for monocular 3D object detection. In *Proceedings of NeurIPS'21*.
[123] Q. Wang, S. Li, H. Qin, and A. Hao. 2016. Super-resolution of multi-observed RGB-D images based on nonlocal regression and total variation. *IEEE Transactions on Image Processing* 25, 3 (2016), 1425–1440.
[124] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang. 2021. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 6 (2021), 3239–3259.
[125] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. 2018. Scalable online convolutional sparse coding. *IEEE Transactions on Image Processing* 27, 10 (2018), 4850–4859.
[126] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.
[127] Z. Wang, X. Ye, B. Sun, J. Yang, R. Xu, and H. Li. 2020. Depth upsampling based on deep edge-aware learning. *Pattern Recognition* 103 (2020), 107274.
[128] Y. Wen, B. Sheng, P. Li, W. Lin, and D. D. Feng. 2019. Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution. *IEEE Transactions on Image Processing* 28, 2 (2019), 994–1006.
[129] B. Wohlberg. 2015. Efficient algorithms for convolutional sparse representations. *IEEE Transactions on Image Processing* 25, 1 (2015), 301–315.
[130] H. Wu, S. Zheng, J. Zhang, and K. Huang. 2018. Fast end-to-end trainable guided filter. In *Proceedings of CVPR*. 1838–1847.
[131] Y. H. Wu, Y. Liu, L. Zhang, M. M. Cheng, and B. Ren. 2022. EDN: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing* 31 (2022), 3125–3136.
[132] J. Xu, X. Deng, M. Xu, and P. L. Dragotti. 2021. CU-Net+: Deep fully interpretable network for multi-modal image restoration. In *Proceedings of ICIP*. 1674–1678.
[133] S. Xu, J. Zhang, K. Sun, Z. Zhao, L. Huang, J. Liu, and C. Zhang. 2021. Deep convolutional sparse coding network for pansharpening with guidance of side information. In *Proceedings of ICME*. IEEE, Los Alamitos, CA, 1–6.
[134] Y. Xu, X. Yu, J. Zhang, L. Zhu, and Wang D. 2022. Weakly supervised RGB-D salient object detection with prediction consistency training and active scribble boosting. *IEEE Transactions on Image Processing* 31 (2022), 2148–2161.
[135] C. Yan, Z. Li, Y. Zhang, Y. Liu, X. Ji, and Y. Zhang. 2020. Depth image denoising using nuclear norm and learning graph model. *ACM Transactions on Multimedia Computing, Communications, and Applications* 16, 4 (2020), 1–17.
[136] Z. Yan, K. Wang, X. Li, Z. Zhang, G. Li, J. Li, and J. Yang. 2022. Learning complementary correlations for depth super-resolution with incomplete data in real world. *IEEE Transactions on Neural Networks and Learning Systems*. Early access, October 26, 2022.
[137] J. Yang, Z. Jiang, X. Ye, and K. Li. 2019. Depth super-resolution with color guidance: A review. In *RGB-D Image Analysis and Processing*. Advances in Computer Vision and Pattern Recognition. Springer, 51–65.
[138] J. Yang, X. Ye, K. Li, C. Hou, and Y. Wang. 2014. Color-guided depth recovery from RGB-D data using an adaptive autoregressive model. *IEEE Transactions on Image Processing* 23, 8 (2014), 3443–3458.
[139] Q. Yang, R. Yang, J. Davis, and D. Nistér. 2007. Spatial-depth super resolution for range images. In *Proceedings of CVPR*. IEEE, Los Alamitos, CA, 1–8.

[140] Y. Yang, Q. Cao, J. Zhang, and D. Tao. 2022. CODON: On orchestrating cross-domain attentions for depth super-resolution. *International Journal of Computer Vision* 130, 2 (2022), 267–284.
[141] Y. Yang, Q. Qin, Y. Luo, Y. Liu, Q. Zhang, and J. Han. 2022. Bi-directional progressive guidance network for RGB-D salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 8 (2022), 5346–5360.
[142] C. Yao, S. Zhang, M. Yang, M. Liu, and J. Qi. 2021. Depth super-resolution by texture-depth transformer. In *Proceedings of ICME*. 1–6.
[143] X. Ye, B. Sun, Z. Wang, J. Yang, R. Xu, H. Li, and B. Li. 2020. PMBANet: Progressive multi-branch aggregation network for scene depth super-resolution. *IEEE Transactions on Image Processing* 29 (2020), 7427–7442.
[144] X. Ye, M. Zhang, J. Yang, X. Fan, and F. Guo. 2020. A sparsity-promoting image decomposition model for depth recovery. *Pattern Recognition* 107 (2020), 107506.
[145] H. Yin, Y. Gong, and G. Qiu. 2019. Side window filtering. In *Proceedings of CVPR*. 8758–8766.
[146] C. Zhang, R. Cong, Q. Lin, L. Ma, F. Li, Y. Zhao, and S. Kwong. 2021. Cross-modality discrepant interaction network for RGB-D salient object detection. In *Proceedings of ACM MM*. 2094–2102.
[147] D. Zhang, J. Han, G. Cheng, and M. H. Yang. 2022. Weakly supervised object localization and detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2022), 5866–5885.
[148] F. Zhang, L. Dai, S. Xiang, and X. Zhang. 2015. Segment graph based image filtering: Fast structure-preserving smoothing. In *Proceedings of ICCV*. 361–369.
[149] N. Zhang, J. Han, and N. Liu. 2022. Learning implicit class knowledge for RGB-D co-salient object detection with transformers. *IEEE Transactions on Image Processing* 31 (2022), 4556–4570.
[150] Y. Zhang, Y. Feng, X. Liu, D. Zhai, X. Ji, H. Wang, and Q. Dai. 2020. Color-guided depth image recovery with adaptive data fidelity and transferred graph Laplacian regularization. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 2 (2020), 320–333.
[151] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia. 2018. PSANet: Point-wise spatial attention network for scene parsing. In *Proceedings of ECCV*. 267–283.
[152] L. Zhao, H. Bai, J. Liang, B. Zeng, A. Wang, and Y. Zhao. 2019. Simultaneous color-depth super-resolution with conditional generative adversarial networks. *Pattern Recognition* 88 (2019), 356–369.
[153] Z. Zhao, J. Zhang, S. Xu, Z. Lin, and H. Pfister. 2022. Discrete cosine transform network for guided depth map super-resolution. In *Proceedings of CVPR*. 5697–5707.
[154] Z. Zhao, J. Zhang, S. Xu, C. Zhang, and J. Liu. 2022. Discrete cosine transform network for guided depth map super-resolution. In *Proceedings of CVPR*. 5697–5707.
[155] H. Zheng, H. Yong, and L. Zhang. 2021. Deep convolutional dictionary learning for image denoising. In *Proceedings of CVPR*. 630–641.
[156] Z. Zhong, X. Liu, J. Jiang, D. Zhao, Z. Chen, and X. Ji. 2022. High-resolution depth maps imaging via attention-based hierarchical multi-modal fusion. *IEEE Transactions on Image Processing* 31 (2022), 648–663.
[157] H. Zhou, L. Qi, H. Huang, X. Yang, Z. Wan, and X. Wen. 2022. CANet: Co-attention network for RGB-D semantic segmentation. *Pattern Recognition* 124 (2022), 108468.
[158] M. Zhou, K. Yan, J. Pan, W. Ren, Q. Xie, and X. Cao. 2023. Memory-augmented deep unfolding network for guided image super-resolution. *International Journal of Computer Vision* 131, 1 (2023), 215–242.
[159] Y. Zuo, Y. Fang, P. An, X. Shang, and J. Yang. 2021. Frequency-dependent depth map enhancement via iterative depth-guided affine transformation and intensity-guided refinement. *IEEE Transactions on Multimedia* 23 (2021), 772–783.
[160] Y. Zuo, Y. Fang, Y. Yang, X. Shang, and Q. Wu. 2019. Depth map enhancement by revisiting multi-scale intensity guidance within coarse-to-fine stages. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 12 (2019), 4676–4687.
[161] Y. Zuo, H. Wang, Y. Fang, X. Huang, X. Shang, and Q. Wu. 2021. MIG-Net: Multi-scale network alternatively guided by intensity and gradient features for depth map super-resolution. *IEEE Transactions on Multimedia* 24 (2021), 3506–3519.
[162] Y. Zuo, Q. Wu, Y. Fang, P. An, L. Huang, and Z. Chen. 2020. Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 2 (2020), 297–306.
[163] Y. Zuo, Q. Wu, J. Zhang, and P. An. 2016. Explicit edge inconsistency evaluation model for color-guided depth map enhancement. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 2 (2016), 439–453.
[164] Y. Zuo, Q. Wu, J. Zhang, and P. An. 2018. Minimum spanning forest with embedded edge inconsistency measurement model for guided depth map enhancement. *IEEE Transactions on Image Processing* 27, 8 (2018), 4145–4159.

Received 18 September 2022; revised 27 January 2023; accepted 14 February 2023