



City Research Online

City St George's, University of London

Citation: Shi, Y., Cao, H., Xia, B., Zhu, R., Liao, Q. & Yang, W. (2024). DSR-Diff: Depth map super-resolution with diffusion model. Pattern Recognition Letters, 184, pp. 225-231. doi: 10.1016/j.patrec.2024.06.025

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/36257/>

Link to published version: <https://doi.org/10.1016/j.patrec.2024.06.025>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

DSR-Diff: Depth Map Super-Resolution with Diffusion Model

Yuan Shi^{a,1}, Huiyun Cao^{a,1}, Bin Xia^a, Rui Zhu^b, Qingmin Liao^a, Wenming Yang^{a,*}

^aShenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China

^bBayes Business School, City, University of London, London EC1V 0HB, United Kingdom

Abstract

Color-guided depth map super-resolution (CDSR) improves the spatial resolution of a low-quality depth map with the corresponding high-quality color map, benefiting various applications such as 3D reconstruction, virtual reality, and augmented reality. While conventional CDSR methods typically rely on convolutional neural networks or transformers, diffusion models (DMs) have demonstrated notable effectiveness in high-level vision tasks. In this work, we present a novel CDSR paradigm that utilizes a diffusion model within the latent space to generate guidance for depth map super-resolution. The proposed method comprises a guidance generation network (GGN), a depth map super-resolution network (DSRN), and a guidance recovery network (GRN). The GGN is specifically designed to generate the guidance while managing its compactness. Additionally, we integrate a simple but effective feature fusion module and a transformer-style feature extraction module into the DSRN, enabling it to leverage the guiding prior in the extraction, fusion, and reconstruction of multi-model images. Taking into account both accuracy and efficiency, our proposed method has shown superior performance in extensive experiments compared to state-of-the-art methods. Specifically, our proposed DSR-Diff demonstrates remarkable improvements of 18.7% and 15.9% in terms of RMSE at scales 8 and 16 on the NYU v2 dataset while saving nearly 30% of the runtime, compared to existing SOTA method. Our codes will be made available at <https://github.com/shiyuan7/DSR-Diff>.

1. Introduction

Depth information is a crucial element in understanding three-dimensional spaces, complementing color data and playing a pivotal role in various computer vision applications, such as 3D reconstruction (Abate et al., 2021), SLAM (Tateno et al., 2017) and semantic segmentation (Chen et al., 2022). However, the practical use of modern depth sensors in consumer devices faces limitations due to budget and low-power constraints, resulting in significantly lower resolutions compared to their RGB counter-

parts. For example, the resolution of depth and color maps acquired by Kinect v2 is 512×424 pixels and 1920×1080 pixels, respectively. This resolution disparity hinders the effectiveness of the depth modality. Considering the accessibility of high-resolution (HR) color images, color-guided depth map super-resolution (CDSR) aims to recover a HR depth map from its low-resolution (LR) counterpart under the guidance of HR color images, benefiting various 3D related downstream tasks.

Traditional CDSR methods usually suffer from limited generalization performance in complex situations. With the development of deep learning, convolutional neural networks (CNNs) have been applied in CDSR (Hui et al., 2016; Wang et al., 2020; Tang et al., 2021; Zhong et al., 2021) and have achieved impressive success. However, CNN-based and Transformer-based approaches are usually optimized with L_1/L_2 loss functions, which tend to produce overly smooth results and lose high-frequency

*Corresponding author.

Email addresses: shiy22@mails.tsinghua.edu.cn (Yuan Shi), caohy22@mails.tsinghua.edu.cn (Huiyun Cao), zjbixia@gmail.com (Bin Xia), rui.zhu@city.ac.uk (Rui Zhu), liaoqm@tsinghua.edu.cn (Qingmin Liao), yang.wenming@sz.tsinghua.edu.cn (Wenming Yang)

¹These authors contributed equally to this work.

information. Limited by the inherent capabilities of foundational models CNN and Transformer, the research of CDSR has encountered the bottleneck and there is an urgent need for new breakthroughs.

Recently, due to potent generation ability and data fitting ability, diffusion models (DMs) (Ho et al., 2020) have exhibited pronounced effectiveness in image synthesis (Ho et al., 2020; Song et al., 2020) and image restoration tasks (Li et al., 2022; Dhariwal & Nichol, 2021; Shang et al., 2023; Xia et al., 2023). DM encompasses a forward process and a reverse process. In the forward process, the input is gradually perturbed by adding Gaussian noise, which defers to a parameterized Markov chain. In the reverse process, retrorse transitions of this chain are learned to recover the original input from pure Gaussian white noise. Compared to other generative models, DMs are capable of generating samples with higher quality and diversity. DMs also exhibit stronger data fitting capabilities in comparison to CNNs and transformers.

In the field of CDSR, existing models are still based on CNN, which have approached the upper limit of CNN. CDSR haven't yet benefited from the powerful capabilities of diffusion models. We aim to design a DM-based CDSR framework, making the most of DMs' powerful distribution mapping and data estimation abilities to further promote the reconstruction performance of CDSR. Nevertheless, the utilization of DMs for depth map super-resolution is confronted with a series of impediments. Firstly, the intrinsic iterative nature of DMs results in a temporal protraction. It usually takes thousands of iterations for DMs to restore images, consuming massive computational resources. Secondly, as generative models, DMs are prone to introduce artifacts which are high-quality perceptually but misaligned with the input image, leading to limited image fidelity.

To address these issues, we propose DSR-Diff, the first DM-based framework for CDSR. We implement diffusion models in the latent space and apply the diffusion process to a compact representation that acts as a guidance prior for depth map SR, rather than directly denoising the entire depth map or its latent embedding. The advantages of our work are twofold. On one hand, the application of DM in a compact latent space significantly reduces the number of iterations required for denoising. On the other hand, the compact representation restored by DM is fused into the SR process as guidance, providing information

regarding depth map degradation and the real data distribution. Since the representation synthesized by DM does not directly affect the output HR depth map, the artifacts can be significantly reduced.

Specifically, we adopt a two-stage training approach. In the first stage, to obtain as much useful information for reconstruction as possible, we compress the ground truth depth map, color image and low-resolution depth map to generate a guiding prior using the proposed guidance generation network (GGN). And an effective and efficient depth map super-resolution network (DSRN) is designed to reconstruct the depth map under the guidance of the prior. In the second stage, the guiding prior is generated by a guidance recovery network (GRN) with low-resolution depth map, relying on the powerful generative capability of diffusion models. Generally, the main contributions of our work are made as follows:

- We present a novel method, DSR-Diff, as the pioneering effort in applying diffusion models in latent space to enhance the resolution of depth maps, yielding noteworthy results.
- We devise the GGN to systematically generate and control compressed guiding priors. To fully harness these guiding priors, we introduced the DSRN, which integrates a guidance embedding module with our proposed simple yet efficient feature fusion module.
- Extensive experiments show that our proposed method achieves state-of-the-art performance, considering both precision and efficiency.

2. Related Work

2.1. Color-guided Depth Map Super-Resolution

Existing CDSR methodologies fall into three categories: filtering-based methods, optimization-based methods and learning-based methods. In filtering-based methods (Tomasi & Manduchi, 1998; Liu et al., 2013), the target image is weighted averaged by local filters to reconstruct depth map. Optimization-based methods (Zuo et al., 2016; Omar & Abd El-Hafeez, 2024) optimize the SR problem globally and employ different conditions to constrain the solution space of the ill-posed problem.

With the rapid development of deep learning, learning-based methods have been proposed to address the limitations of hand-designed objective functions and filters

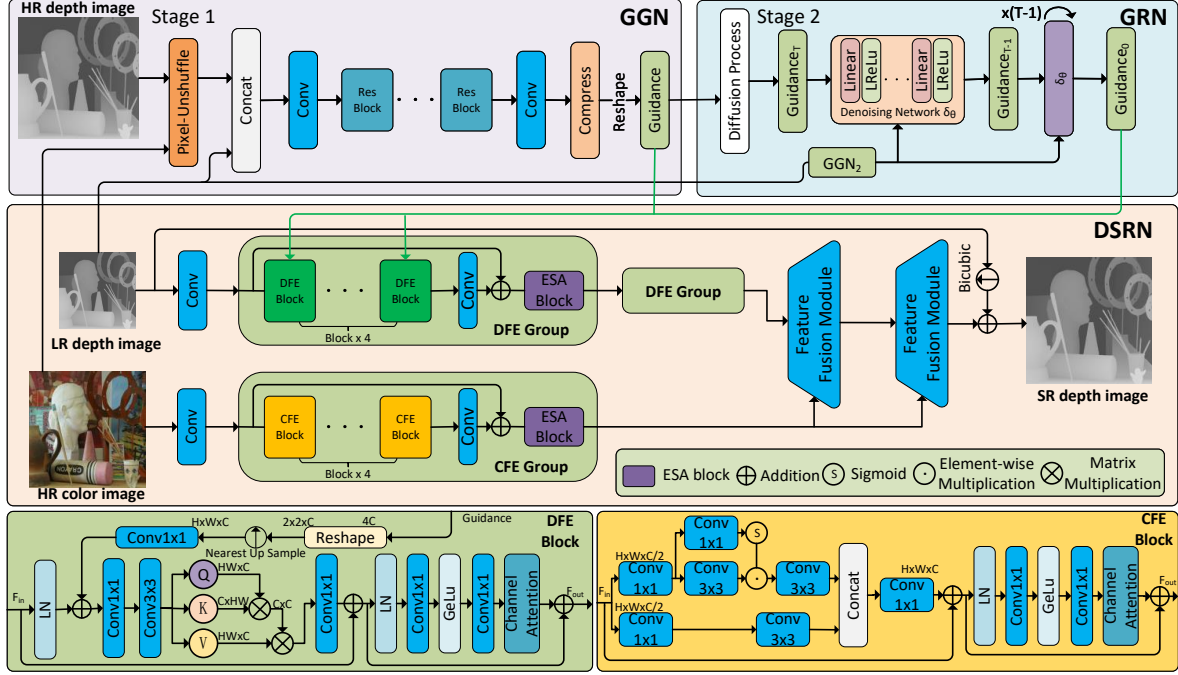


Figure 1: The architecture of the proposed DSR-Diff. (a) DSRN utilizes two branches to extract features from LR depth images and HR color images, followed by feature fusion and reconstruction of the depth map, respectively. (b) In the first stage, LR depth images, HR depth images, and HR color images are input into GGN to generate guidance, directing the feature extraction and reconstruction in DSRN. (c) In the second stage, GRN is designed to estimate the guidance generated by GGN, thanks to the powerful estimation abilities of the diffusion model.

(Zuo et al., 2019), which acquire the mapping between low-resolution (LR) and precise high-resolution (HR) depth maps. As the first CNN architecture specifically designed for CDSR, MSG-Net (Hui et al., 2016) employs multi-scale depth and color feature maps to enhance both the structure and texture of the HR depth maps. To leverage the edge information of the color maps, Wang et al. (2020) propose a two-stage method consisting of an edge-aware learning framework and depth restoration modules. Then, JIIF (Tang et al., 2021) is proposed to learn the interpolation weights, informed by the graph attention mechanism and implicit neural representation. Utilizing an efficient CNN architecture, AHMF (Zhong et al., 2021) introduced an attention-based fusion mechanism for the integration of guidance information. Learning-based methods have promoted the performance of CDSR significantly.

2.2. Diffusion Models

Diffusion models (Ho et al., 2020) generate samples of the target distribution from Gaussian noise via a stochastic iterative denoising process. DMs have been successfully applied to various generative tasks, such as image synthesis (Ho et al., 2020; Song et al., 2020; Dhariwal & Nichol, 2021), speech synthesis (Liu et al., 2022) and 3D generation (Luo & Hu, 2021), and demonstrated excellent performance. In the field of image restoration, Li et al. (2022) propose the first DM-based single image super-resolution algorithm SRDiff. SR3 (Saharia et al., 2022) adapts DMs to conditional image generation and performs super-resolution through an iterative denoising process. On the basis of SRDiff and SR3, a large number of DM-based studies have emerged, Dhariwal & Nichol (2021); Shang et al. (2023); Gao et al. (2023). To alleviate the computational burden of DMs, Rombach et al. (2022); Xia et al. (2023) perform DM in latent space to improve

the restoration efficiency. However, DMs haven't been applied to multi-modal restoration tasks such as CDSR. It is pivotal to design a DM-based framework for CDSR according to the characteristics of color image and depth map.

3. Proposed Method

3.1. Overview

Most DM-based image restoration methods perform the diffusion process on the whole image or its feature maps, which require a large number of iterations to denoise and introduce undesirable artifacts. To alleviate these problems, we propose the first diffusion model-based network for depth map super-resolution, DSR-Diff. We compress the HR depth map into the latent space and obtain its compact representation which contains the real data distribution information beneficial to image restoration. Since DMs have powerful distribution mapping and data estimation abilities, we adopt DM in the latent space to estimate the highly compact representation, which is fused into the SR process as a guidance.

Figure 1 shows the overall architecture of our DSR-Diff. The guidance generation network (GGN) and guidance recovery network (GRN) are responsible for extracting and recovering the compact restoration guidance mentioned above and trained in two stages separately. The depth map super-resolution network (DSRN) functions as the main restoration network, employing LR depth images, HR color images, and the restoration guidance as inputs to reconstruct SR depth images. During the first stage, GGN generates the restoration guidance by extracting and compressing beneficial knowledge from HR color images, HR depth images, and LR depth images. In the second stage, GRN utilizes the data estimation abilities of DM to recover the compact restoration guidance. In the forward process, the diffusion model progressively introduces noise into the guidance extracted by GGN. In the reverse process, the diffusion model utilizes LR depth images to estimate the necessary guidance from pure Gaussian noise. In the first stage, GGN and DSRN are optimized together, enabling the GGN to fully learn to extract guidance conducive to depth map restoration. And in the second stage, GRN and DSRN are optimized together while parameters of GGN are frozen, for the sake of training GRN to denoise and recover the restoration guidance.

Additionally, the method we proposed requires the input to be a two-dimensional image. For other types of multimodal information, modifications need to be made to a portion of the network input.

To provide a more intuitive understanding of our algorithm, we present the training pseudocode for Stage 1 and Stage 2 in algorithm 1 and 2, respectively.

Algorithm 1 Training for Stage 1

Input: HR depth images dataset $\mathcal{D}_D^{HR}$, LR depth images dataset $\mathcal{D}_D^{LR}$, HR color images dataset $\mathcal{D}_C^{HR}$, total training steps N

Output: SR depth images D_{SR}

```

1: Initialization: network parameters  $\omega_{DSRN}, \omega_{GGN}$ 
2: repeat
3:   for step  $n = 1$  to  $N$  do
4:     Sample  $D_{HR}^n$  from  $\mathcal{D}_D^{HR}$ ,  $D_{LR}^n$  from  $\mathcal{D}_D^{LR}$ ,  $C_{HR}^n$  from  $\mathcal{D}_C^{HR}$ 
5:      $G^n = GGN(D_{HR}^n, D_{LR}^n, C_{HR}^n)$ 
6:      $D_{SR}^n = DSRN(G^n, D_{LR}^n, C_{HR}^n)$ 
7:     Take gradient descent step on
            $\nabla_{\omega_{DSRN}, \omega_{GGN}} \|D_{SR}^n - D_{HR}^n\|$ 
8:   end for
9: until converged

```

3.2. Guidance Generation Network (GGN)

In the first stage of training, GGN is designed to extract a compact representation from the HR depth image, HR color image, and LR depth image and utilize the essential information in the representation to provide accurate and concise guidance for DSRN.

As shown in figure 1, HR depth image and HR color image are initially downsampled using pixel-unshuffle according to the scale factor. Subsequently, the downsampled images are concatenated with the LR depth image and pass through several residual blocks to extract preliminary guidance feature $G' \in \mathbb{R}^{H \times W \times C}$,

$$G' = Res(Concat(D_{LR}, Down(C_{HR}, D_{HR})), \quad (1)$$

where D_{LR} , C_{HR} and D_{HR} denote the LR depth image, HR color image and HR depth image respectively. $Down(\cdot)$ denotes the downsampling operation, and we use the PixelUnshuffle to realize it in our work. $Concat(\cdot)$ denotes

concatenation operation. $Res(\cdot)$ is the residual block for feature extraction.

The restoration guidance is going to be recovered by DM in GRN. In order to reduce the iterations of denoising as well as the difficulty of generating accurate guidance for DM, we divide the feature map of each channel in G' into $K \times K$ blocks and then employ pooling within each block to obtain a highly compressed guidance. We can adjust the compression degree of the guidance by modifying the value of K , which allows the compressed guidance to adapt to different tasks.

As the value of K increases, the compression level of the guiding prior information decreases, resulting in both a higher computational demand and richer guiding information. For more challenging tasks, we recommend selecting a larger value of K . In the context of the CDSR problem, since the depth image is guided by a color image, there is no need for excessive additional guiding priors. In order to strike a balance between accuracy and computational complexity, we set the value of K to 2. And we reshape the feature map into a vector to generate the final restoration guidance $G \in \mathbb{R}^{K^2C}$. This process can be expressed as,

$$G = \text{Reshape}(\text{Compress}(G')) \quad (2)$$

where $\text{Compress}(\cdot)$ and $\text{Reshape}(\cdot)$ denote compression and reshape operation, respectively.

3.3. Guidance Recovery Network (GRN)

In the second stage, we use GRN to generate guidance from the LR depth image. The guidance $G \in \mathbb{R}^{K^2C}$ is initially generated by the GGN trained in the first stage. And we apply the diffusion model on G to sample G_T and estimate $\hat{G}_0$ from G_T .

In the forward process, each iteration can be described as:

$$q(G_t|G_{t-1}) = \mathcal{N}(G_t; \sqrt{1-\beta_t}G_{t-1}, \beta_t\mathbf{I}) \quad (3)$$

where G_t is the noised depth map at time-step t , β_t is the variance schedule, $\mathcal{N}$ denotes the Gaussian distribution. Under the property of Markov chain, noised G_t at an arbitrary timestep t can be sampled directly,

$$q(G_t|G_0) = \mathcal{N}(G_t; \sqrt{\bar{\alpha}_t}G_0, (1-\bar{\alpha}_t)\mathbf{I}) \quad (4)$$

Algorithm 2 Training for Stage 2

Input: HR depth images dataset $\mathcal{D}_D^{HR}$, LR depth images dataset $\mathcal{D}_D^{LR}$, HR color images dataset $\mathcal{D}_C^{HR}$, total training steps N , diffusion steps T

Output: SR depth images D_{SR}

```

1: Initialization:  $\omega_{DSRN}, \omega_{GRN}$ 
2: repeat
3:   for step  $n = 1$  to  $N$  do
4:     Sample  $D_{HR}^n$  from  $\mathcal{D}_D^{HR}$ ,  $D_{LR}^n$  from  $\mathcal{D}_D^{LR}$ ,  $C_{HR}^n$  from  $\mathcal{D}_C^{HR}$ 
5:      $G^n = \text{GGN}(D_{HR}^n, D_{LR}^n, C_{HR}^n)$ 
6:     Diffusion Process:
7:      $G_0^n = G^n$ 
8:      $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ 
9:      $G_T^n = \sqrt{\bar{\alpha}_T}G_0^n + \sqrt{1-\bar{\alpha}_T}\epsilon$ 
10:    Reverse Process:
11:     $\hat{G}_T^n = G_T^n$ 
12:     $C_G^n = \text{GGN}_2(D_{LR}^n)$ 
13:    for  $t = T, \dots, 1$  do
14:       $\hat{G}_{t-1}^n = \frac{1}{\sqrt{\alpha_t}}(\hat{G}_t^n - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(\hat{G}_t^n, t, C_G^n))$ 
15:    end for
16:    Depth Map Super-Resolution:
17:     $\hat{G}^n = \hat{G}_0^n$ 
18:     $D_{SR}^n = \text{DSRN}(\hat{G}^n, D_{LR}^n, C_{HR}^n)$ 
19:    Take gradient descent step on
       $\nabla_{\omega_{DSRN}, \omega_{GRN}} \|D_{SR}^n - D_{HR}^n\| + \|\hat{G}^n - G^n\|$ 
20:    end for
21: until converged

```

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=0}^t \alpha_i$. Thus given $G \in \mathbb{R}^{K^2C}$, the final noised guidance G_T can be obtained by

$$q(G_T|G) = \mathcal{N}(G_T; \sqrt{\bar{\alpha}_T}G, (1 - \bar{\alpha}_T)\mathbf{I}) \quad (5)$$

where T is the total number of iterations.

In the reverse process, DM learns to transfer G_t to G_{t-1} and generates samples of the target distribution from pure Gaussian white noise. Given G_t and G_0 , the posterior distribution of G_{t-1} is formulated as:

$$q(G_{t-1}|G_t, G_0) = \mathcal{N}(G_{t-1}; \mu_t(G_t, G_0), \sigma_t^2 \mathbf{I}), \quad (6)$$

$$\mu_t(G_t, G_0) = \frac{1}{\sqrt{\alpha_t}}(G_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon), \sigma_t^2 = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t. \quad (7)$$

where ϵ represents the noise in G_t , which is the only variable in the reverse process. We adopt a denoising network ϵ_θ to estimate ϵ . In our GRN, the LR depth image is initially input into GGN_2 , which has almost the same architecture with GGN except for the input, to generate the condition C_G . Then C_G , G_t and t are input to the denoising network ϵ_θ to estimate noise, which can be described as:

$$\hat{G}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\hat{G}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta(\hat{G}_t, t, C_G)), \quad (8)$$

After T times iterations, we get the estimated guidance $\hat{G}_0 \in \mathbb{R}^{K^2C}$.

Since the guidance in our DSR-Diff is highly compressed, we can use a straightforward denoising network ϵ_θ to estimate noise, making the guidance estimation simple and efficient. Additionally, because the number of iterations required for denoising is significantly reduced, we don't have to randomly sample a time-step and optimize ϵ_θ at that step like traditional DMs. We can run all denoising iterations to obtain the estimated guidance and send it to DSRN for joint optimization.

3.4. Depth Map Super-Resolution Network (DSRN)

DSRN can utilize the guidance generated by GGN or GRN to guide the feature extraction of depth images and obtain the reconstructed SR depth maps. DSRN primarily consists of the depth image feature extraction group (DFE group), the color image feature extraction group (CFE group), and the feature fusion module. In our pursuit of reconstructing depth images with a focus on both efficiency and accuracy, we have chosen to utilize two

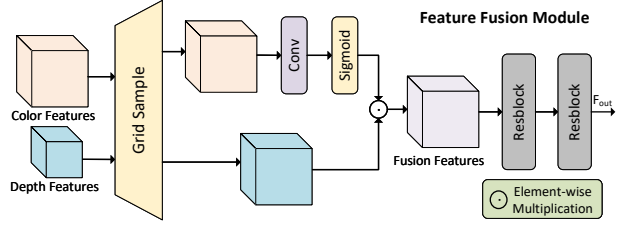


Figure 2: The architecture of our proposed feature fusion module (FFM).

DFE groups to process depth features, along with a dedicated CFE group specifically designed for handling color features.

By concatenating multiple DFE Blocks and incorporating an Enhanced Spatial Attention (ESA) block (Liu et al., 2020) in a sequential manner, a feature extraction architecture denoted as the DFE Group is obtained. We devise the DFE block based on the transformer block in Zamir et al. (2022). And to leverage the guidance, the input guidance is first reshaped, upsampled to match the size of the feature map, and then added to the feature map after a 1x1 convolution. Given an input feature $F_{in} \in \mathbb{R}^{H \times W \times C}$ and guidance $G \in \mathbb{R}^{K^2C}$. This process can be formulated as:

$$\hat{F} = \text{Conv}(\text{Up}(\text{Reshape}(G))) + \text{LN}(F_{in}), \quad (9)$$

where $\hat{F} \in \mathbb{R}^{H \times W \times C}$ is the guidance-adjusted feature map, $\text{Reshape}(\cdot)$ is the reshape operation and $\text{Reshape}(G) \in \mathbb{R}^{K \times K \times C}$, $\text{Up}(\cdot)$ is the nearest upsampling, $\text{Conv}(\cdot)$ refers to a convolution with a kernel size of 1, and $\text{LN}(\cdot)$ is the layer normalization.

The structure of the CFE Group is consistent with that of the DFE Group. Due to the fact that the color image feature extraction branch does not require parsing information from the guidance, to enhance the efficiency of the model, we employ a simpler approach to design the CFE block inspired by Zhao et al. (2020) as shown in figure 1.

The extracted deep features and color features are upsampled to the same size and then fused within the feature fusion module (FFM), as illustrated in figure 2. The reconstructed SR depth image is obtained after passing through two FFM. Depth and color features are upsampled through grid sampling, allowing the sampling of features at specific coordinates, thereby mitigating the impact of positional offsets in depth and color features.

3.5. Feature Fusion Module (FFM)

Traditional learning-based DSR methods such as Tang et al. (2021) and Hui et al. (2016) often directly concatenate the color and depth features which tend to introduce unnecessary features in color image and interfere with the SR of depth map, as well as resulting in additional computation.

In our DSR-Diff, we introduce a new feature fusion module FFM, which fuses the features of the color image and depth image through element-wise multiplication and gated convolution. Specifically, color features are first upsampled to the same size as the depth features. To alleviate the problem of position deviation between depth image and color image, we perform the upsampling operation according to coordinates: a set of coordinates in the range of $[-1, 1]$ is generated according to the number of pixels in HR color image and then upsample the color features according to the coordinates. After upsampling, we apply convolution and the Sigmoid activation function to the complex color features for the adaptive fusion of color and depth features. Finally, the output color features are multiplied by upsampled depth features with two residual blocks performing further refinement,

$$F_{fuse} = \text{Sigmoid}(\text{Conv}(Up(F_c))) \cdot Up(F_d) \quad (10)$$

$$F_{out} = \text{Res}(\text{Res}(F_{fuse})) \quad (11)$$

where F_c , F_d , F_{fuse} , F_{out} and F_d denote the input color features, depth features, fused features and output features, respectively.

The proposed FFM is a simple and effective method for the fusion of color features and depth features. Element-wise multiplication allows the effective components for reconstruction in depth and color features to be retained, while weak features in the depth map impair those of the corresponding position in the color image. In this way, the interference of non-essential texture in the color image to the depth image is effectively solved.

3.6. Loss Function

In the first stage, our objective is to generate an effective guidance and get high-quality depth map through DSRN. So the training loss is defined as follows:

$$L_{img} = \|D_{GT} - D_{SR}\|_1 \quad (12)$$

where D_{GT} is the ground-truth depth image, D_{SR} is the reconstructed depth image, and $\|\cdot\|_1$ denotes the L_1 norm.

In the second stage, we need to estimate the guidance in the first stage and reconstruct the SR depth image. So we jointly train the DRN and DSRN using loss function L_{com} :

$$L_{com} = \|G - \hat{G}_0\|_1 + \|D_{GT} - D_{SR}\|_1 \quad (13)$$

where G and $\hat{G}_0$ are the guidance generated in the first and the second stage, respectively.

4. Experiments

4.1. Implementation Details

4.1.1. Dataset

We select three benchmark RGB-D datasets, NYU V2 (Silberman et al., 2012), Middlebury (Hirschmuller & Scharstein, 2007; Scharstein & Pal, 2007), and Lu datasets (Lu et al., 2014) to evaluate our method. NYU V2 dataset includes 1449 RGB-D image pairs, with 1000 image pairs for training and the remaining 449 pairs for evaluation. Middlebury and Lu datasets have 30 and 6 RGB-D image pairs, respectively, all of which are utilized for evaluation.

4.1.2. Training Setting

We apply a batch-mode learning method with a batch size of 1. Both the first and second stage require training for 300 epochs each. And the initial learning rate is set to 2×10^{-4} and decays by half every 80 epochs. The Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$ are used for optimization. Since we only need to utilize diffusion models to recover compressed compact priors, we can achieve a good fit using a small number of time steps, thereby saving computational time. In training the diffusion model, time step T are set to 4 and β_t ($\alpha_t = 1 - \beta_t$) linearly increase from 0.1 to 0.99. And all the experiments are performed with PyTorch on a PC with one GTX 3090 GPU, taking almost one day.

4.2. Comparisons with State-of-the-Art Methods

We compare DSR-Diff with state-of-the-art methods. Quantitative comparison results at scales of 4x, 8x, and 16x are presented in Table 1. We use the average RMSE as the evaluation metric. Following Tang et al. (2021), the

Table 1: Quantitative comparison with state-of-the-art methods on NYU v2, Middlebury, and Lu datasets in terms of RMSE (the lower, the better). The best performance is shown in **bold**, while the second best performance is the underscored ones.

Method	Runtime (ms)	NYU V2			Middlebury			Lu		
		4×	8×	16×	4×	8×	16×	4×	8×	16×
Bicubic	\	4.28	7.14	11.58	2.28	3.98	6.37	2.42	4.54	7.38
MSG-Net (Hui et al., 2016)	58	3.02	5.38	9.17	1.88	3.45	6.28	2.30	4.17	7.22
DJF (Li et al., 2016)	\	3.54	6.20	10.21	2.14	3.77	6.12	2.54	4.71	7.66
DG (Gu et al., 2017)	\	3.68	5.78	10.08	1.97	4.16	5.27	2.06	4.19	6.90
DJFR (Li et al., 2019)	122	2.38	4.94	9.18	1.32	3.19	5.57	1.15	3.57	6.77
CUNet (Deng & Dragotti, 2020)	173	1.92	3.70	6.78	1.10	2.17	4.33	0.91	2.23	5.19
FDKN (Kim et al., 2021)	27	1.86	3.55	6.96	1.09	2.17	4.51	0.82	2.09	5.03
DKN (Kim et al., 2021)	182	1.62	3.26	6.51	1.23	2.12	4.24	0.96	2.16	5.11
JiIF (Tang et al., 2021)	191	<u>1.37</u>	<u>2.76</u>	<u>5.27</u>	1.09	1.82	3.31	<u>0.85</u>	1.73	4.16
AHMF (Zhong et al., 2021)	63	1.40	2.89	5.64	<u>1.07</u>	1.63	<u>3.14</u>	0.88	<u>1.66</u>	3.71
DCTNet (Zhao et al., 2022)	107	1.59	3.16	5.84	1.10	2.05	4.19	0.88	1.85	4.39
Ours	76	1.25	2.57	4.91	1.04	<u>1.67</u>	3.10	0.91	1.65	<u>3.92</u>

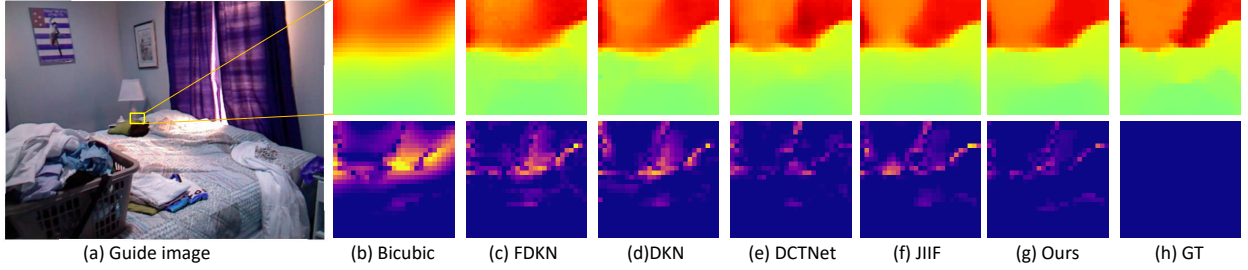


Figure 3: Qualitative comparisons for $\times 8$ CDSR on the NYU v2 dataset. The first row is a zoomed-in view of the reconstructed depth map. The second row displays the error map between the results and the ground truth and brighter areas indicate more significant errors.

average RMSE is measured in centimeters for the NYU v2 dataset. For the other two datasets, RMSE computations are carried out after scaling the depth values within the $[0, 255]$ range.

The results in Table 1 clearly indicate that our proposed methods outperform alternative approaches in terms of reconstruction accuracy. This superiority is particularly evident in the NYU v2 and Middlebury datasets, both of which include more test images. Compared to DCTNet (CVPR2022), our proposed method demonstrates remarkable improvements of **18.7%** and **15.9%** in terms of RMSE at scales 8 and 16 on the NYU v2 dataset while saving nearly **30%** of the runtime. To further illustrate the superiority of our proposed method, we compared its runtime, parameters, and accuracy with the current state-of-the-art methods, as depicted in Figure 4. It is evident that, considering both efficiency and accuracy, our proposed method achieves state-of-the-art performance.

The superiority in accuracy benefits from the powerful distribution estimation and data generation abilities of DMs. Owing to the fact that DMs can model highly complex distributions of natural images, the DM in GRN has learned the distribution of HR depth images and provided beneficial prior about the real distribution for SR. In addition, the DM in DSR-Diff is designed to generate a guidance for SR, rather than generating the whole image or its latent embedding directly. The novel paradigm overcomes the negative influence of artifacts introduced by DM and benefits reconstruction accuracy. The high compactness of the guidance reduces the computation cost of DM significantly.

To better demonstrate the performance of DSR-Diff, we provide an intuitive visual comparison in Fig. 3. It can be clearly seen that the depth images reconstructed by DSR-Diff exhibit superior visual results, especially in areas with complex textures and edges.

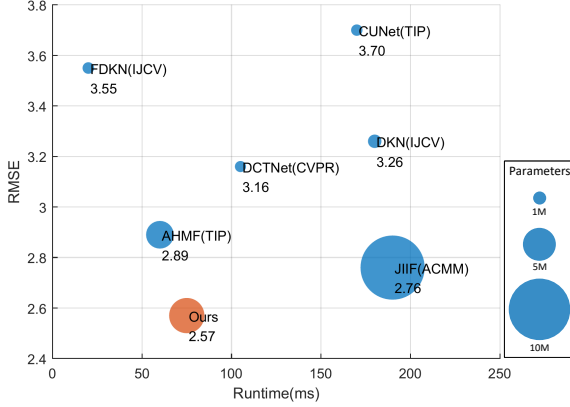


Figure 4: Efficiency comparison with the current state-of-the-art methods for 8x CDSR on the NYU v2 dataset. The RMSE results are indicated below the name of each method, and the circular area serves as a visual representation of the model’s parameters.

4.3. Ablation Studies

To validate the effectiveness of each module in DSR-Diff, we conducted ablation experiments and the results are presented in Table 2.

	M-1	M-2	M-3	M-4	M-5
GGN	×	×	√	√	√
GRN	×	√	×	√	√
FFM	√	√	√	×	√
Average RMSE	2.643	2.597	2.570	2.587	2.572

Effective guidance for depth map SR. In M-1, the guidance is not incorporated into DSRN and we solely utilize the designed DSRN for depth image reconstruction. In M-2, the compression operation in GGN is replaced with a regular pooling operation. M-3 presents the results of the first stage. In conjunction with M-5, it substantiates the efficacy of the diffusion model in the second stage, aligning with the guidance generated by GGN.

Impact of the feature fusion module. In M-4, the up-sampled depth features and color features in FFM are directly concatenated together. Comparing M-5 with M-4, it is obvious that our FFM is able to fuse the color features and depth features more effectively by element-wise multiplication.

5. Discussion

5.1. Strengths and Weaknesses

Our proposed DSR-Diff leverages the powerful fitting capability of diffusion models to estimate the guiding prior information during the super-resolution process. In the first stage, we employ a joint training approach to learn the required compact guiding priors in the super-resolution process. In the second stage, we utilize diffusion models to recover the compact priors obtained in the first stage. We have designed the Guidance Generation Network to achieve more flexible control over the information density of the guiding priors. This enables our proposed method to have higher computational efficiency compared to other diffusion-based approaches. However, due to the limited scale of the training dataset, the generalization performance of our currently trained models still requires further enhancement.

5.2. Future Prospects

Denoising Diffusion Probabilistic Models (DDPM) achieve image reconstruction through stochastic differential equations and Markov chains, exhibiting stronger generative capacity and interpretability compared to transformer and CNN architectures. Our proposed DSR-Diff successfully applies diffusion models to the CDSR problem and achieves promising performance. However, the commonly used datasets for CDSR tasks are often small in scale, which limits the application of more powerful networks. We are currently exploring the establishment of a more comprehensive training dataset. In addition, other advanced architectures such as vision mamba (Shi et al., 2024) are also awaiting further exploration in this problem.

6. Conclusion

In this paper, we propose DSR-Diff, the first diffusion model-based framework for color-guided depth map super-resolution. Our proposed DSR-Diff fully leverages the powerful fitting capability of diffusion models and achieves higher accuracy and reduced execution time compared to previous methods based on transformers or CNNs in the field of depth map super-resolution. To address the time-consuming nature of diffusion models, we

apply diffusion models in the latent space to generate a compact prior from the ground truth depth map, which acts as a guidance for depth map SR. Furthermore, we jointly optimize the diffusion model with an effective and efficient depth map super-resolution network (DSRN) to leverage the generated guidance. And a guidance generated network (GGN) is proposed to ensure that the guidance is appropriately generated and compressed. The feature fusion module (FFM) employs a simple yet efficient approach to achieve the integration of multi-modal features. Extensive experiments validate the effectiveness and superiority of our proposed DSR-Diff. We hope that our work can drive the application of diffusion models in the field of depth map super-resolution, thereby further enhancing the performance of depth super-resolution algorithms.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos.62171251&62311530100) and the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen (Nos.CJGJZD2021040809280401&JCYY20200109143010272)

References

- Abate, A. F., De Maio, L., Distasi, R., & Narducci, F. (2021). Remote 3d face reconstruction by means of autonomous unmanned aerial vehicles. *Pattern Recognition Letters*, 147, 48–54.
- Chen, M., Shu, R., Zhu, D., Li, J., & Zhang, X. (2022). Csf: Closed-mask-guided semantic fusion method for semantic perception of unknown scenes. *Pattern Recognition Letters*, 161, 101–107.
- Deng, X., & Dragotti, P. L. (2020). Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 43, 3333–3348.
- Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34, 8780–8794.
- Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., & Zhang, B. (2023). Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10021–10030).
- Gu, S., Zuo, W., Guo, S., Chen, Y., Chen, C., & Zhang, L. (2017). Learning dynamic guidance for depth image enhancement. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3769–3778).
- Hirschmuller, H., & Scharstein, D. (2007). Evaluation of cost functions for stereo matching. In *2007 IEEE conference on computer vision and pattern recognition* (pp. 1–8). IEEE.
- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840–6851.
- Hui, T.-W., Loy, C. C., & Tang, X. (2016). Depth map super-resolution by deep multi-scale guidance. In *European conference on computer vision* (pp. 353–369). Springer.
- Kim, B., Ponce, J., & Ham, B. (2021). Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129, 579–600.
- Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., & Chen, Y. (2022). Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479, 47–59.
- Li, Y., Huang, J.-B., Ahuja, N., & Yang, M.-H. (2016). Deep joint image filtering. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14* (pp. 154–169). Springer.
- Li, Y., Huang, J.-B., Ahuja, N., & Yang, M.-H. (2019). Joint image filtering with deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 41, 1909–1923.
- Liu, J., Li, C., Ren, Y., Chen, F., & Zhao, Z. (2022). Diff-singer: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11020–11028). volume 36.

- Liu, J., Zhang, W., Tang, Y., Tang, J., & Wu, G. (2020). Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2359–2368).
- Liu, M.-Y., Tuzel, O., & Taguchi, Y. (2013). Joint geodesic upsampling of depth images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 169–176).
- Lu, S., Ren, X., & Liu, F. (2014). Depth enhancement via low-rank matrix completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3390–3397).
- Luo, S., & Hu, W. (2021). Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2837–2845).
- Omar, A., & Abd El-Hafeez, T. (2024). Optimizing epileptic seizure recognition performance with feature scaling and dropout layers. *Neural Computing and Applications*, 36, 2835–2852.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684–10695).
- Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., & Norouzi, M. (2022). Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, 4713–4726.
- Scharstein, D., & Pal, C. (2007). Learning conditional random fields for stereo. In *2007 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). IEEE.
- Shang, S., Shan, Z., Liu, G., & Zhang, J. (2023). Resdiff: Combining cnn and diffusion model for image super-resolution. *arXiv preprint arXiv:2303.08714*, .
- Shi, Y., Xia, B., Jin, X., Wang, X., Zhao, T., Xia, X., Xiao, X., & Yang, W. (2024). Vmambair: Visual state space model for image restoration. *arXiv preprint arXiv:2403.11423*, .
- Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from rgb-d images. In *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12* (pp. 746–760). Springer.
- Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, .
- Tang, J., Chen, X., & Zeng, G. (2021). Joint implicit image function for guided depth super-resolution. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 4390–4399).
- Tateno, K., Tombari, F., Laina, I., & Navab, N. (2017). Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6243–6252).
- Tomasi, C., & Manduchi, R. (1998). Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)* (pp. 839–846). IEEE.
- Wang, Z., Ye, X., Sun, B., Yang, J., Xu, R., & Li, H. (2020). Depth upsampling based on deep edge-aware learning. *Pattern Recognition*, 103, 107274.
- Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., & Van Gool, L. (2023). Diffir: Efficient diffusion model for image restoration. *arXiv preprint arXiv:2303.09472*, .
- Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5728–5739).
- Zhao, H., Kong, X., He, J., Qiao, Y., & Dong, C. (2020). Efficient image super-resolution using pixel attention. In *Computer Vision—ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16* (pp. 56–72). Springer.

- Zhao, Z., Zhang, J., Xu, S., Lin, Z., & Pfister, H. (2022). Discrete cosine transform network for guided depth map super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5697–5707).
- Zhong, Z., Liu, X., Jiang, J., Zhao, D., Chen, Z., & Ji, X. (2021). High-resolution depth maps imaging via attention-based hierarchical multi-modal fusion. *IEEE Transactions on Image Processing*, 31, 648–663.
- Zuo, Y., Fang, Y., Yang, Y., Shang, X., & Wu, Q. (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, 4676–4687.
- Zuo, Y., Wu, Q., Zhang, J., & An, P. (2016). Explicit edge inconsistency evaluation model for color-guided depth map enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 28, 439–453.