RGB Guided ToF Imaging System: A Survey of Deep Learning-Based Methods

Xin Qiao[1] · Matteo Poggi[2,3] · Pengchao Deng[1] · Hao Wei[1] · Chenyang Ge[1] · Stefano Mattoccia[2,3]

Abstract

Integrating an RGB camera into a ToF imaging system has become a significant technique for perceiving the real world. The RGB guided ToF imaging system is crucial to several applications, including face anti-spoofing, saliency detection, and trajectory prediction. Depending on the distance of the working range, the implementation schemes of the RGB guided ToF imaging systems are different. Specifically, ToF sensors with a uniform field of illumination, which can output dense depth but have low resolution, are typically used for close-range measurements. In contrast, LiDARs, which emit laser pulses and can only capture sparse depth, are usually employed for long-range detection. In the two cases, depth quality improvement for RGB guided ToF imaging corresponds to two sub-tasks: guided depth super-resolution and guided depth completion. In light of the recent significant boost to the field provided by deep learning, this paper comprehensively reviews the works related to RGB guided ToF imaging, including network structures, learning strategies, evaluation metrics, benchmark datasets, and objective functions. Besides, we present quantitative comparisons of state-of-the-art methods on widely used benchmark datasets. Finally, we discuss future trends and the challenges in real applications for further research.

Keywords RGB-guided ToF imaging · Deep learning · Depth super-resolution · Depth completion · Multimodal fusion

1 Introduction

The real world is three-dimensional, yet usually recorded with 2D vision sensors such as color cameras. To faithfully portray it, a diffuse trend is to employ a depth camera during collection, widely used nowadays in downstream tasks related to computer vision and pattern recognition, such as face anti-spoofing (Liu et al., 2021b; Deng et al., 2022), saliency detection (Cong et al., 2018), autonomous driving (Häne et al., 2017; Wang et al., 2019) and vir-

Communicated by Yasuyuki Matsushita.

✉ Chenyang Ge
  cyge@mail.xjtu.edu.cn

1   Institute of Artificial Intelligence and Robotics, Xi'an
    Jiaotong University, No.28, West Xianning Road, Xi'an
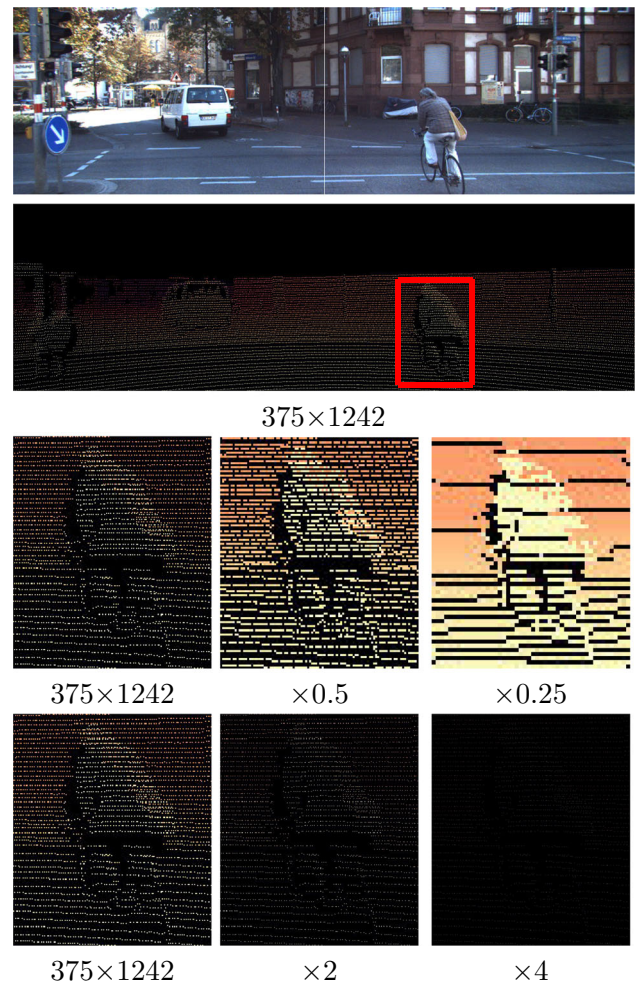   710049, Shaanxi, China

2   Department of Computer Science and Engineering (DISI),
    University of Bologna, Viale Risorgimento, 2, 40136
    Bologna, Emilia-Romagna, Italy

3   Advanced Research Center on Electronic System (ARCES),
    University of Bologna, Via Vincenzo Toffano, 2/2, 40125
    Bologna, Emilia-Romagna, Italy

tual/augmented reality (VR/AR) (Holynski & Kopf, 2018; Kalia et al., 2019). Among depth cameras, time-of-flight (ToF) technology emerges as the preferred choice for many applications thanks to its compact structure, high precision, and low cost. Recently, the iPad Pro and iPhone Pro/Pro Max from Apple use LiDAR based on direct ToF (dToF) technology, often adopted in fields such as biometrics, photography, games, modeling, virtual reality, and augmented reality. The ToF-only imaging system is usually preferred when cost efficiency, energy efficiency, fast real-time, or privacy sensitivity are the primary considerations.

While ToF-only cameras might be the preferred choice in scenarios where depth sensing is paramount, some drawbacks still limit their performance in applications such as scene understanding, mixed reality, human-computer interaction, and facial recognition, where accuracy, versatility, and image content cues play a crucial role. In particular, the resolution of existing consumer-grade ToF cameras is about 300K pixels, with the LiDAR sensors – again based on the time of flight principle – used, for instance, by autonomous vehicles capturing even sparser measurements. In contrast, the resolution of RGB cameras used in mobile phones (e.g., Huawei P60) can reach dozens of MegaPixels, i.e., orders of

magnitude higher, yet at a much lower cost. One solution is to directly enhance the low-quality (LQ) ToF depth map, i.e., to upsample (Riegler et al., 2016b) or densify (Uhrig et al., 2017) it, relying on hand-crafted or neural network-learned priors. However, these priors are likely to cause the predicted High-Resolution (HR) depth to contain blurred edges and other undesired artifacts. To face these issues, *RGBD cameras* integrate a HR RGB imager and a ToF sensor. Then, under the guidance of HR color images, sparse or Low-Resolution (LR) depth maps can be densified or super-solved by exploiting the structural details available from the high-resolution image content.

In practical applications, ToF cameras based on different technical routes are deployed in different scenarios, mainly divided into the following three categories: 1) for short-ranging acquisition in applications like face/gesture recognition, typically using an indirect ToF (iToF) camera with a uniform field of illumination, obtaining a dense depth map, yet at a low resolution; 2) for long-range measurements, as in the case of LiDAR sensors emitting laser pulses, which produces a scattered point cloud of the sensed scene: when coupled with a color camera, the sparsity of the depth map obtained by projecting the point cloud on the image plane depends on the resolution of the image, the higher, the sparser – i.e., containing only 5% of pixels encoding valid depth measurements (Ma et al., 2018) on $\sim$ 0.3Mpx images; 3) for medium-ranging, as in the case of the ToF cameras used in augmented reality and automated guided vehicle robots, either iToF or dToF can be used. However, even when iToF is used for short-ranging, depth values around the corners or edges may be missing.

In the three cases, improving the quality of original depth maps translates into 1) upsampling the original, spatial resolution through Guided Depth Super-Resolution (GDSR) (Ferstl et al., 2013; He et al., 2021), or 2) densifying the sparse depth map through Guided Depth Completion (GDC) (Uhrig et al., 2017; Chodosh et al., 2019). The goal of GDSR is to utilize RGB images to increase pixel density for low-resolution depth and provide richer visual details, while GDC aims to fill in missing depth values in sparse depth maps obtained when using laser pulses, as well as increase the overall depth map resolution – as the sparsity of the depth map depends on the resolution of the image over which the sensed point cloud is projected (the lower the image resolution, the denser the projected depth map; the higher the image resolution, the sparser the depth map). In Fig. 1, we show an example of LiDAR point cloud from the KITTI dataset reprojected at the original resolution used in the dataset (375×1242) as well as both at lower and higher resolution (respectively ×0.5, ×0.25, ×2 and ×4). We can observe how assuming different camera resolutions produces depth maps whose density changes sensibly, with fewer holes present at lower resolutions and much sparser depth points at

Fig. 1 An example of LiDAR point cloud from the KITTI dataset reprojected on images at different resolutions

higher resolutions. This evidence confirms that the density of a depth map obtained from LiDAR scans is a function of the resolution at which we wish to obtain a depth map and that, to some extent, the completion process has analogies with the super-resolution task. The key difference with GDSR is that the laser spot emitted by LiDAR in GDC is tiny and typically cannot cover the scene corresponding to one pixel in the sensor. Therefore, GDC usually requires perceiving a larger neighborhood (e.g., 7 × 7) than GDSR in shallow feature extraction of the input to estimate depth values in areas where the information is missing. Generally, RGB guided ToF imaging systems involve one of the two sub-tasks, depending on the use case and the depth sensing technologies. Consequently, different solutions must be deployed depending on the specific sub-task. Nonetheless, both tasks aim to yield high-quality depth maps with the guidance of RGB information, that is, transfer high-frequency structural information and fine-grained features from HR color images to

LR/sparse depth maps while avoiding texture-copying artifacts. Furthermore, it is worth noting that the model design of the two tasks exhibits intriguing similarities, which inspired us to analyze them together as one topic for investigation, and their specific classification will be elaborated upon in Sec. 2.4.

At first, traditional methods extract meaningful features from color images through guided filtering-based (Petschnigg et al., 2004; He et al., 2012; Liu et al., 2012) or optimization-based (Diebel & Thrun, 2005; Ferstl et al., 2013; Lu et al., 2014) algorithms. These hand-crafted approaches rely on the assumption that depth discontinuities statistically co-occur in correspondence with RGB image edges. In real-world cases, however, this prior may fail to extract informative features and introduce artifacts. For instance, in the presence of rich texture, these methods convey essential structural details but also transfer the appearance of RGB images to depth maps as a side effect. Therefore, this task remains challenging as it requires properly distinguishing between high-frequency image information corresponding to discontinuities in the depth map and those not. Over the past decades, solutions for enhancing RGB guided ToF imaging have shifted from model-driven to data-driven paradigms (Deng & Dragotti, 2019; Wong et al., 2021), in particular thanks to deep learning methods (Cheng et al., 2020; Zuo et al., 2020b; Zhao et al., 2022) that have rapidly advanced the field. Even with a simple network structure, such as a few stacked convolutional layers (Uhrig et al., 2017) or the vanilla encoder-decoder (Li et al., 2016) architecture, it is straightforward to recover HR/dense depth maps with much higher accuracy compared to what obtained through hand-crafted methods. Consequently, RGB guided ToF imaging via deep learning received increasing research interest from academia and industry, motivating us to survey recent progress in this field.

Specifically, given the substantial development of deep-learning solutions in the last decade, this survey aims to provide a complete overview of solutions designed for enhancing RGB guided ToF imaging in both use cases. Although we are aware of previous review papers on guided depth super-resolution (Zhong et al., 2023b) and depth completion (Xie et al., 2022; Hu et al., 2022) treated as standalone tasks, we highlight the lack of interconnection between the two, despite the shared insights they bring to the community of researchers working with RGB guided ToF sensors, which we aim to envelop in a comprehensive survey. Purposely, we will review the literature concerning deep learning frameworks developed to enhance the depth maps perceived from this family of devices. We will study and classify them according to several aspects, such as the framework design, the learning paradigm, and the objective functions minimized during training. Moreover, we introduce the reader to the standard datasets used as benchmarks for evaluating existing methodologies, reporting the performance of state-of-the-art methods as a reference.

We outline our major contributions as follows:

- To the best of our knowledge, we are the first to provide an in-depth investigation for RGB guided ToF imaging through deep learning, including guided depth super-resolution and guided depth completion as the main tasks to be dealt with according to the specific use case.
- In each subtask, this paper reviews recent deep learning-based methods under several aspects to illustrate the motivations behind their design, contributions, and performance.
- We discuss the challenges and future trends of RGB guided ToF imaging.

The rest of this survey is structured as follows. In Sect. 2, we describe preliminaries for RGB guided ToF imaging. Sections 3 and 4 present deep learning-based methods for GDSR and GDC, respectively. In Sect. 5, we introduce the datasets and loss functions. Section 6 provides comparison results of some recent methods. Finally, we discuss future trends and challenges in Sect. 7. Figure 2 draws a timeline and reports for each year from 2017 to 2023, any method we will review in this paper, and the publication venue.

2 Preliminaries and Taxonomy

In this section, we first introduce the foundation of RGB guided ToF imaging, then we present GDSR and GDC as the two main research trends aimed at enhancing ToF depth maps and the commonly used objective functions in both.
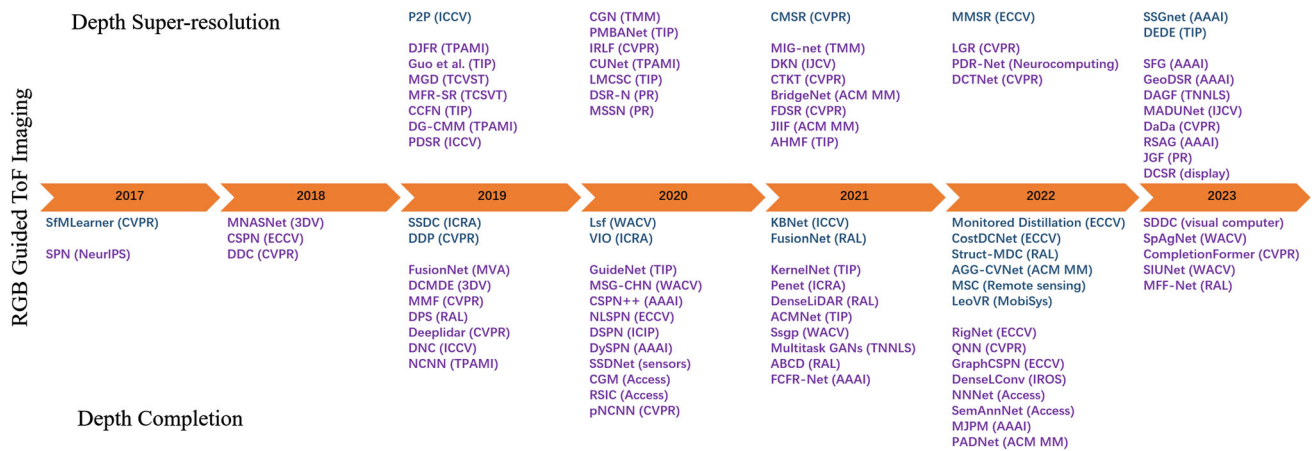
2.1 Principles of ToF Imaging

Here, we describe the imaging principles of iToF and dToF, respectively. In Fig. 3, we provide a schematic diagram of the principles commonly used by iToF and dToF.

2.1.1 Indirect ToF

Indirect ToF (iToF) determines distance by capturing reflected light and calculating the phase delay between emitted and reflected light. Currently, most of the off-the-shelf ToF cameras emit signals $s(t)$ using sinusoidal amplitude modulation:

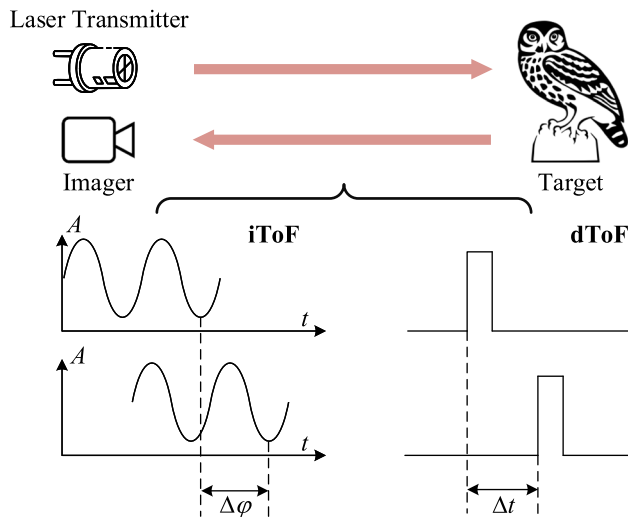$$s(t) = s_1 \cos(\omega t) + s_0 \tag{1}$$

Since the baseline between the ToF transmitter and the receiver is so small (usually only a few millimeters), they can generally be considered coaxial. Assuming that the target is static, and the signal is only reflected once before being

Fig. 2 Timeline showing the advances in RGB guided ToF Imaging. Methods are marked in blue or purple, depending on the learning paradigm being unsupervised or supervised

Table 1 Characteristics of different types of LiDAR

Type	Technology	Advantages	Disadvantages
Mechanical	Rotating mirrors or prisms	High accuracy, long range	Large, expensive
MEMS	Micromirrors	Small, cheap	Lower FoV, range
Flash	Arrays of lasers firing simultaneously	Fast, high resolution	Short range
OPA	Optical phased arrays	Small, cheap, versatile	Still in development

Fig. 3 iToF and dToF working principles. The former computes distances based on the phase shift between emitted and reflected light; the latter measures the time required for a laser light to bounce over an object

received by the sensor, the reflected signal of a single pixel measured at a specific fixed frequency can be expressed as a function of time t:

$$r(t) = \frac{\rho}{c^2 \tau_0^2}[s_1 \cos(\omega t - 2\omega\tau_0) + s_0] + e_0 \qquad (2)$$

where ρ is the surface reflectivity of the target, ω denotes angular frequency, τ_0 is the delay of signal reflection from the target back to a pixel at the speed of light $c \approx 3 \times 10^8$ m/s, and e_0 is the offset caused by ambient light. According to Heide et al. (2015), Eq. 1 can be rewritten as:

$$A_1 = \frac{\rho s_1}{c^2 \tau_0^2}, \quad A_0 = \frac{\rho s_0}{c^2 \tau_0^2} + e_0 \qquad (3)$$

$$r(t) = A_1 \cos(\omega t - 2\omega\tau_0) + A_0 \qquad (4)$$

The received signal remains sinusoidal, but the phase, which contains scene depth information, is changed. After a high-frequency periodic signal $B_1 \cos(\omega t - \phi)$ modulates the incident signal, the modulated signal can be expressed as:

$$
\begin{aligned}
\tilde{I}_{\phi,\omega}(t) = & r(t) B_1 \cos(\omega t - \phi) \\
= & \frac{A_1 B_1 \cos(\phi - 2\omega\tau_0)}{2} + \\
& \frac{A_1 B_1 \cos(2\omega t + \phi + 2\omega\tau_0)}{2} + \\
& A_0 B_1 \cos(\omega t - \phi)
\end{aligned}
\qquad (5)
$$

Considering the quantum efficiency of ToF sensors, the sensor exposure time T set to capture a sufficient number of photons is usually much larger than $\pi c/\omega$. During the exposure, the measurement is equivalent to integrating over the time range T, so the period terms in Eq. 5 all vanish in the

calculation, and the result can be computed as:

$$I_{\phi,\omega} = \int_{-T/2}^{T/2} \tilde{i}_{\phi,\omega}(t)dt$$
$$\approx A_1 B_1 \cos(\phi - 2\omega\tau_0) = A_1 g_{\phi,\omega} \qquad (6)$$

where A_1 is the scene response that represents the intensity of the encoded corresponding pixel at time τ_0 after the laser transmitter sends a light pulse, $g_{\phi,\omega} = B_1 \cos(\phi - 2\omega\tau_0)$ is the camera function, and $I_{\phi,\omega}$ is the raw measurement with correlation at phase ϕ, angular frequency ω, i.e. ToF raw data. If a single modulation frequency f_m is used to collect multiple ($K \geq 2$) raw measurements in one cycle, the distance between the scene and camera can be calculated as:

$$d = \frac{c}{2\omega} \arctan\left(\frac{\sin\phi \cdot I_{\phi,\omega}}{\cos\phi \cdot I_{\phi,\omega}}\right) \qquad (7)$$

In practice, if the modulation frequency $f_m = \frac{\omega}{2\pi}$ is fixed, ToF cameras usually adopt the four-phase method to sample ToF raw data, i.e., sampling four times in one cycle, with a phase step of $90°$. For ease of reading, the four raw data can be abbreviated as $I_i = I_{\frac{\pi}{2\omega}i}, i = 0, 1, 2, 3$.

$$A = \frac{\sqrt{(I_3 - I_1)^2 + (I_0 - I_2)^2}}{2} \qquad (8)$$

$$B = \frac{I_0 + I_1 + I_2 + I_3}{4} \qquad (9)$$

$$\phi = \arctan\frac{I_3 - I_1}{I_0 - I_2} \qquad (10)$$

$$d = \frac{c}{2\pi}(\frac{\phi}{2f} + N) \qquad (11)$$

where N is an integer that represents the number of phase wraps. Based on the above principle, some works (Su et al., 2018; Chugunov et al., 2021; Gutierrez-Barragan et al., 2021; Li et al., 2022b) directly process the raw signal of the iToF, thus improving the depth quality.

At present, most iToF for consumer use sensors based on complementary metal-oxide semiconductor (CMOS) or charge-coupled device (CCD), with a resolution of 320×240 (QVGA) or 640×480 (VGA) pixels. However, the resolution is still significantly lower than that of an RGB image using the same technology.

2.1.2 Direct ToF

DToF is a method that measures the time it takes for laser light to be fired directly at an object and returned to the sensor. The moment the laser is fired, the electronic clock is activated. Then, the pulse bounces off the target and is partially picked up by the photodetector. By measuring the time lapse ΔT, the

distance d between the outgoing light pulse and the reflected light pulse is calculated with

$$d = \frac{c\Delta T}{2} \qquad (12)$$

To achieve higher sensitivity and lower power consumption, dToF usually adopts single photon avalanche diodes (SPADs) instead of Avalanche Photodiodes (APDs) as photodetectors. Due to limitations in the size of the chip and its auxiliary circuits, the pixel count of SPAD arrays is much lower than that of iToF.
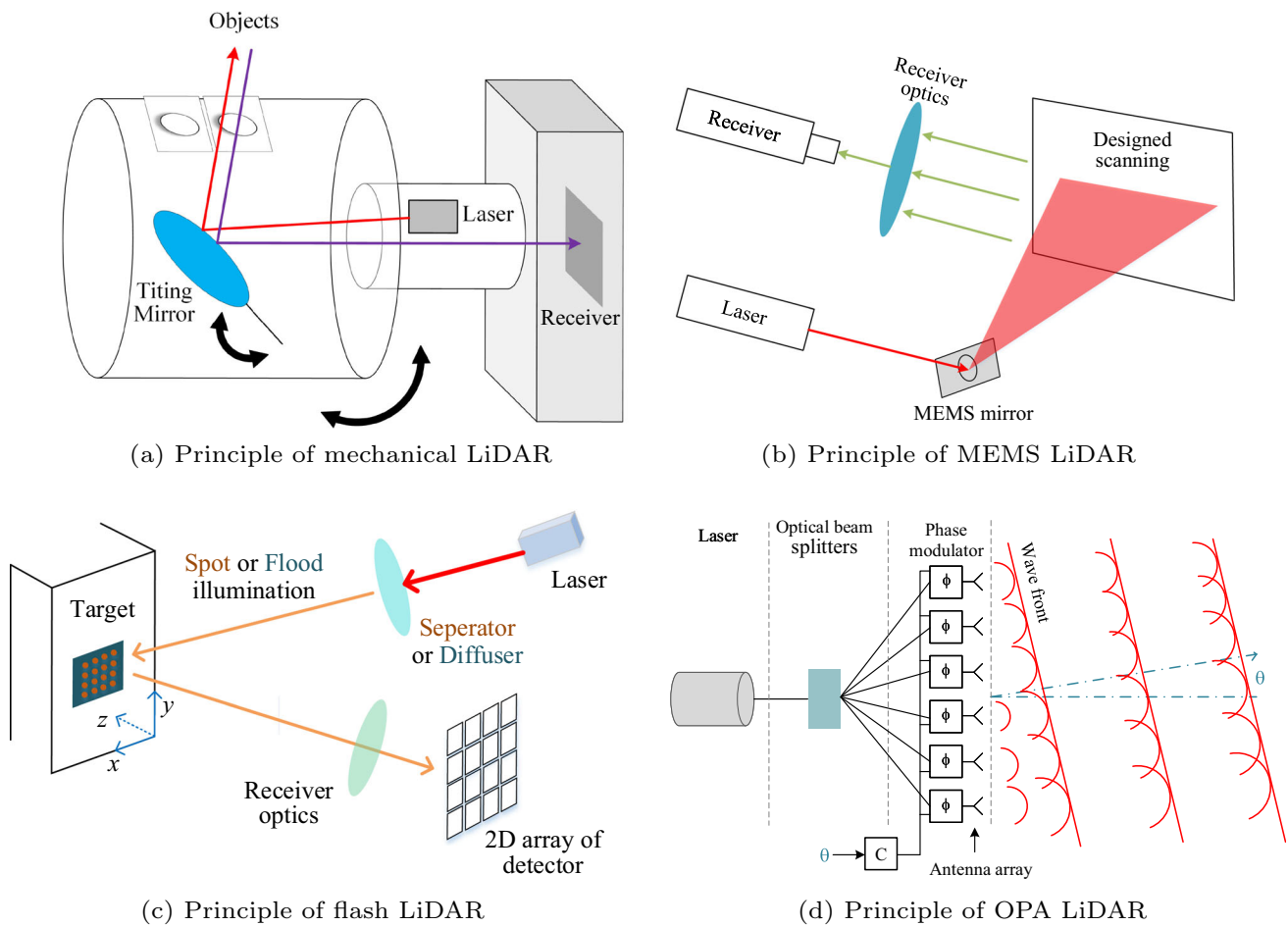
LiDAR, an acronym for Light Detection and Ranging, despite being based on the same working principle as dToF, is frequently listed as a distinct depth sensing technology due to some of its peculiar design and setup. For instance, conventional LiDAR continuously scans the environment utilizing mechanical components (e.g., spinning mirrors). Another type is MEMS LiDAR, which uses micromirrors controlled by micro-electromechanical systems (MEMS) to scan the environment. This type of LiDAR is smaller and cheaper than mechanical LiDARs but has lower FoV and range. As a still developing technology, OPA LiDARs use optical phased arrays (OPAs) to steer the laser beam electronically, and they have the potential to be small, cheap, and versatile. Different from the techniques mentioned above, global shutter flash LiDARs, also referred to as flash LiDARs or ToF LiDARs, use arrays of lasers firing simultaneously to create a 3D image of the surrounding environment, just like iToF, thereby improving frame rate and imaging efficiency for applications such as autonomous driving. This technology enables the integration of dToF into consumer electronic devices. We provide a comparison summarizing the different types of LiDAR in Table 1 and the schematic diagram of each method in Fig. 4.

In essence, LiDAR still employs the dToF principle, and the two terms of dToF and LiDAR are even interchangeable in many scenarios (e.g., Apple refers to the dToF sensor in iPad Pro and iPhone 12 Pro as a LiDAR scanner). However, LiDAR technology is rapidly evolving, and new state-of-the-art LiDARs are constantly being developed, thus potentially having more applications in various scenarios. For instance, the InnovizOne MEMS LiDAR, manufactured by EDOM technology, achieves an angular resolution of up to $0.1° \times 0.1°$, producing much denser point clouds than before.

2.1.3 Comparative Analysis of dToF and iToF

Here, we provide their characteristics in Table 2. Besides, we conduct a comparative analysis on the following aspects.

Accuracy. iToF cameras typically provide accurate distance measurements within a specific range. Its measurement error is theoretically positively correlated with the target

(a) Principle of mechanical LiDAR

(b) Principle of MEMS LiDAR

(c) Principle of flash LiDAR

(d) Principle of OPA LiDAR

Fig. 4 Schematic diagram of LiDAR systems with different imaging principles

Table 2 Characteristics of iToF and dToF

Parameter	iToF	dToF
Principle	Using phase shift between emitted and reflected light to determine distance.	Using stop watch method to calculate time lapse.
Detector type	PMDs: $6 \sim 100 \mu m$ pixel size	SPADs / APDs
Depth calculation	In-pixel calculation	Histogram analysis
Performance	Long integration time	Fast acquisition
Range	Short-medium	Longer
Range ambiguity	Yes	No
Pixel count	Large	Smaller
Accuracy	Linearly related to distance	Higher
Power consumption	Low	Higher
Portability	Generally compact	Traditionally larger
Cost	Medium	Higher

distance. Currently, the accuracy of mass-produced iToF cameras can be controlled within 1%, e.g., Helios2 ToF 3D Camera with Sony's IMX556 DepthSense ToF has an accuracy of less than 5 mm and a precision of less than 2 mm at 1 m camera distance. At the same time, dToF (LiDAR) can achieve millimeter-level accuracy even over longer distances. Therefore. for short-range applications, both iToF and dToF can achieve comparable accuracy, although they may fall

short of structured light cameras. However, a key advantage of LiDAR is that its accuracy substantially remains constant regardless of distance, as discussed in Sec 2.1.3. This attribute positions LiDAR measurements as reliably accurate in long-range applications, distinguishing them from other technologies.

Range and Power consumption. iToF cameras are generally suitable for short to medium-range applications. The effective range may vary but typically within a few to tens of meters, depending on the specific design and implementation. Because of these characteristics, ToF cameras are often designed to be power-efficient and, therefore, suitable for consumer-grade devices. In contrast, LiDAR systems can achieve longer ranges, making them suitable for applications like autonomous vehicles and long-range mapping. Consequently, they are commonly employed in scenarios where power consumption is less critical.

Portability. iToF cameras are inherently compact and lightweight, making them suitable for integration into portable devices such as smartphones, tablets, and wearable gadgets. Meanwhile, LiDAR, traditionally larger, is undergoing miniaturization trends, especially with the development of solid-state LiDAR, making it increasingly applicable in portable and handheld devices.

Applications. Due to the properties peculiar to each, the two families of sensors are usually involved in different downstream applications. Specifically, iToF sensors are the best fit for AR/VR experiences, which mainly focus on narrow, close-range environments or, in general, for any applications running in a constrained space (e.g., bin picking with a robotic arm). On the contrary, dToF is preferable in contexts requiring long-range perception, e.g., the automotive one with ADAS or autonomous driving systems. It is worth noting that there is an overlap in the operating ranges of iToF and dToF, i.e., both can be used for applications such as AR/VR (e.g., Kinect v2 with iToF and iPad Pro with dToF) when performing medium range sensing.

Though both imaging systems currently excel in their respective application domains, advancements in technology are steadily bridging the gap between their performance metrics, leading to a growing overlap in their practical uses.

2.2 Problem Formulation

Although ToF cameras have many advantages, they also bring some inherent limitations, such as multipath interference Gupta et al. (2015); Gutierrez-Barragan et al. (2021), flying pixels Qiao et al. (2020) and wiggling errors Hussmann et al. (2013). Hand-crafted models can deal with some issues, e.g., depth correction, but others require additional cues, such as those available in RGB images.

The goal of RGB guided ToF imaging is to recover a high-quality depth map Z_{hq} from the acquired low-quality depth

map $Z_{lq} : \Omega_Z \subset \Omega \mapsto \mathbb{R}_+$ and high-resolution color image $I : \Omega \subset \mathbb{R}^2 \mapsto \mathbb{R}^3_+$. Depending on the setup of ranging systems, low-quality depth maps can be either dense, yet at low resolution, or sparse. Besides, we assume all LQ depth maps and the corresponding color images are correctly aligned. This is possible if the RGB-D system is calibrated, i.e., the relative pose between the two color and depth cameras is known. If a deep neural network Φ_γ with parameters γ is deployed for RGB guided ToF imaging, the task can be modeled as

$$\hat{Z}_{hq} = \Phi_\gamma(Z_{lq}, I) \tag{13}$$

where $\hat{Z}_{hq}$ is the prediction of latent HR depth map.

The network parameters γ are updated during training of Φ, by solving the following optimization problem:

$$\hat{\gamma} = \arg\min_\gamma \mathcal{L}(Z_{hq}, \hat{Z}_{hq}) \tag{14}$$

where $\mathcal{L}$ is an objective function, usually minimizing the distance between the prediction and the ground truth depth.

2.3 Evaluation Metrics

As for most computer vision tasks, for RGB guided ToF imaging, it is necessary to employ appropriate evaluation metrics. The most commonly used measures for both the two sub-tasks, i.e., GDSR and GDC, are root mean squared error (RMSE) and mean absolute error (MAE):

$$\text{RMSE}(mm) = \sqrt{\frac{1}{N} \sum_{p \in N} (Z_{hq}^p - \hat{Z}_{hq}^p)^2} \tag{15}$$

$$\text{MAE}(mm) = \frac{1}{N} \sum_{p \in N} |Z_{hq}^p - \hat{Z}_{hq}^p| \tag{16}$$

with p being a single pixel in the depth map.

Metrics for GDSR. In addition to the above metrics, mean squared error (MSE) is frequently utilized, which has a similar role to RMSE:

$$\text{MSE} = \frac{1}{N} \sum_{p \in N} |Z_{hq}^p - \hat{Z}_{hq}^p|^2 \tag{17}$$

Furthermore, with a short-range ToF sensor, GDSR focuses on recovering depth maps with desirable details. As a result, peak signal-to-noise ratio (PNSR) and structural similarity index (SSIM) are also sometimes used to assess the quality of depth maps, which are defined as:

$$\text{PSNR} = 10 log_{10} \left(\frac{Z_{max}^2}{\text{MSE}} \right) \tag{18}$$

$$\text{SSIM}(p, q) = \frac{(2\mu_p\mu_q + C_1)(2\sigma_{xy} + C_2)}{(\mu_p^2 + \mu_q^2 + C_1)(\sigma_p^2 + \sigma_q^2 + C_2)} \quad (19)$$

with Z_{max} being the maximum depth value; q is a neighbor of pixel p; μ_p and μ_q denote the mean values of HQ depth maps and the corresponding ground truth, respectively; σ_p^2 and σ_q^2 are the variance; C_1 and C_2 are constants used to maintain the stability of the division. The smaller the pixel value difference between the two depth maps, the higher the PSNR.

Metrics for GDC. Sparse depth measurements are often captured from a long-range ToF LiDAR, so there are several task-specific metrics for GDC, including RMSE of the inverse depth (iRMSE), MAE of the inverse depth (iMAE), which are defined by

$$\text{iRMSE}(\frac{1}{km}) = \sqrt{\frac{1}{N} \sum_{p \in N} (Z_{hq}^p - \hat{Z}_{hq}^p)^2} \quad (20)$$

$$\text{iMAE}\left(\frac{1}{km}\right) = \frac{1}{N} \sum_{p \in N} |\frac{1}{Z_{hq}^p} - \frac{1}{\hat{Z}_{hq}^p}| \quad (21)$$

When the evaluation of deep models is performed on indoor datasets, e.g., NYU-v2 (Silberman et al., 2012), mean absolute relative error (REL) and thresholded accuracy are more popular
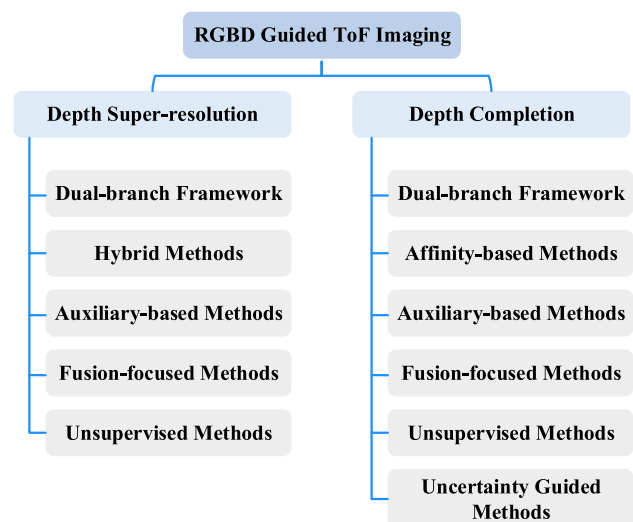
$$\text{REL}(mm) = \frac{1}{N} \sum_{p \in N} \frac{|Z_{hq}^p - \hat{Z}_{hq}^p|}{Z_{hq}^p} \quad (22)$$

$$\sigma = \max\left(\frac{Z_{hq}^p}{\hat{Z}_{hq}^p}, \frac{\hat{Z}_{hq}^p}{Z_{hq}^p}\right) < th \quad (23)$$

with th being a given threshold.

2.4 Taxonomy

Figure 5 anticipates the proposed taxonomy for methods we will survey in this paper. On the one hand, from the figure, we can already highlight some common trends in GDSR and GDC tasks. Specifically, we can identify the category of *Dual-branch Frameworks* in both, i.e., those deploying two main branches for processing RGB and depth cues separately, as well as the category of *Auxiliary-based Methods* – those exploiting multi-task learning to improve the accuracy of the predicted depth map – *Fusion-focused Methods* which mainly study the fusion strategy for combining color and depth features, and the *Unsupervised Methods* not requiring ground-truth labels at training time. On the other hand, we identify some trends that are peculiar to one of the two tasks: specifically, for GDSR, we have *Hybrid Methods*, focusing on obtaining explainable models to run the super-resolution

Fig. 5 Taxonomy of RGB guided ToF imaging techniques. We identify 5 and 6 main categories for GDSR and GDC methods

process; for GDC, *Affinity-based Methods* developed are characterized by the used of Spatial Propagation Networks (SPNs) which learn affinity matrices for propagating depth across neighboring pixels, and *Uncertainty-guided Methods* explicitly involve uncertainty modeling to improve the results. Each category will be introduced and discussed in detail in Sects. 3 and 4, for GDSR and GDC, respectively.

3 Guided Depth Super-Resolution

In this section, we present and discuss recent deep learning-based approaches. Initially, researchers exploited general architectures with two branches to extract features from RGB and depth separately, which are fused in a naive manner, i.e., by means of concatenation or summation. Numerous novel architectures and fusion schemes have been proposed to boost network accuracy. From the perspective of data requirements, depth super-resolution methods are divided into supervised and unsupervised learning. The difference between the two paradigms is that the former requires labeled input data, while the latter does not. For what concerns supervised learning, we classify existing works into three categories, depending on the main contribution they bring to the literature, in terms of (i) Dual-branch framework; (ii) optimization-inspired architecture; (iii) auxiliary learning; (iv) multi-modal fusion scheme. For unsupervised learning, the emphasis is on the fusion strategy and loss functions. These five categories are listed in the top branch of Fig. 5.
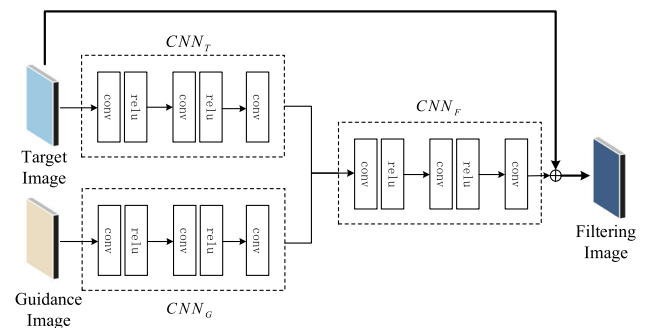
Supervised methods are supplemented by labeled data so that the task can be viewed as a regression problem. Thus, deep models aim to learn depth predictions as close as possible to the ground truth depth labels.

3.1 Dual-Branch Framework

Since the depth sensor captures the same scene as the RGB camera, the obtained two images are geometrically similar and complementary. Therefore, how to extract meaningful features from the two modalities through designed algorithms is the key to depth super-resolution. Intuitively, Li et al. (2019) design a deep joint filter (DJFR) which contains two sub-networks for extracting depth and RGB features, respectively. After concatenating the two streams, the fused information is decoded by another sub-network and outputs the predicted depth, as seen in Fig. 6.
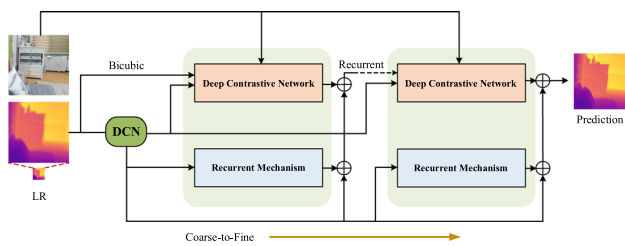
Multi-scale information is crucial for many low-level vision tasks, including GDSR, which can recover images from different scales through rich hierarchical features. The multi-scale guided convolutional network (MSG-Net) (Hui et al., 2016) is the first work to super-solve depth maps using multi-scale guidance, which progressively enhances depth details from high- to low-level. Inspired by this idea, numerous methods utilizing multi-scale guidance have been proposed and attained significantly improved performance. For instance, a novel deep network for depth map super-resolution (DepthSR) proposed by Guo et al. (2019) leverages a residual U-Net architecture (Ronneberger et al., 2015), where hierarchical guidance is introduced at each scale, to recover HR depth maps. Zuo et al. (2019a) design a multi-scale architecture with a guide similar to MSG-Net (Hui et al., 2016) where dense layers (Huang et al., 2017) are employed to revisit the features from all higher levels at a given scale. Also, guided by RGB images, the multi-scale fusion residual network for GDSR (MFR-SR) proposed by Zuo et al. (2019b) progressively upsamples depth maps in multiple scales with global and local residual learning. Li et al. (2020b) employ a multi-scale strategy with the proposed symmetric unit (SU) as the basic component of the network. With SU, it can effectively deal with textureless and edge features in depth images. Unlike previous works, Zuo et al. (2021) propose a two-stream multi-scale network, MIG-net, including three branches, i.e., intensity, depth, and gradient branches. At each scale, the depth and gradient features are iteratively refined by the guide. Zhong et al. (2021) introduce an attention-based hierarchical multi-modal fusion network (AHMF) where a bi-directional hierarchical feature collaboration module is proposed to make full use of multi-scale features. In this module, features from different levels are aggregated so that low and high level spatial information can collaborate to improve each other. Wang et al. (2023b) present a novel method that exploits pyramid structure to capture multi-scale features so that HR depth maps can be progressively recovered. Yuan et al. (2023b) propose a structure flow-guided framework for GDSR, whose core is to learn a structure flow map to guide structure representation transfer. In this method, a flow-enhance pyramid edge attention

Fig. 6 DJFR architecture (Li et al., 2019) belonging to dual-branch framework. It employs a typical dual-branch auto-encoder framework, where the guide and source features are extracted from the two sub-networks in the encoder

network is introduced to learn multi-scale features for sharp edge reconstruction.

Using a coarse-to-fine scheme is another common solution for GDSR, consisting of building a multi-stage network to enhance depth maps progressively. The definitions of the multi-scale strategy and the coarse-to-fine scheme may overlap in some approaches. Thus, we distinguish them according to the focus of the works. Wen et al. (2019) propose a coarse-to-fine neural network with two stages. In the coarse stage, the model learns large filter kernels to obtain less accurate results, further refined by smaller filter kernels in the fine stage. In Zuo et al. (2020b), the filtering and refinement of depth-guided intensity features and intensity-guided depth features are iteratively conducted in multiple stages through the proposed depth-guided affine transformation. The network can reduce artifacts produced by distribution gaps between depth maps and the corresponding guide. Ye et al. (2020) introduce a multi-branch aggregation network with multiple stages to reconstruct sharp depth boundaries. Specifically, they design three parallel branches – the reconstruction, the color, and the multi-scale branches – in each stage. Song et al. (2020) develop a coarse-to-fine framework consisting of several sub-modules that can gradually extract the high-frequency features from depth maps. Each sub-module employs a channel attention strategy to obtain informative features. Besides, the framework further improves the depth quality through Total Generalized Variation (TGV) term and input loss. In He et al. (2021), a high-frequency guidance branch is developed to adaptively extract the HF information and reduce the LF components. The HF components are then fused with depth features by means of concatenation. Yuan et al. (2023a) propose a recurrent structure attention guidance framework (RSAG) that utilizes both multi-scale and multi-stage strategies, shown in Fig. 7. In this framework, a deep contrastive network with multi-scale filters is designed to adaptively separate HF and LF features.
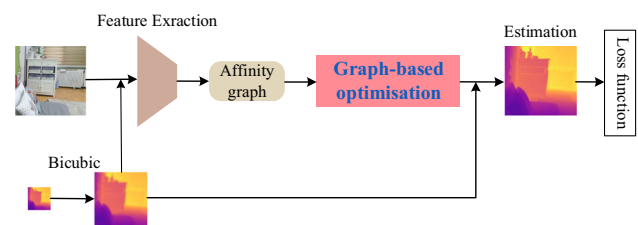
Fig. 7 RSAG architecture (Yuan et al., 2023a) belonging to dual-branch framework. It progressively refines the depth using a coarse-to-fine strategy
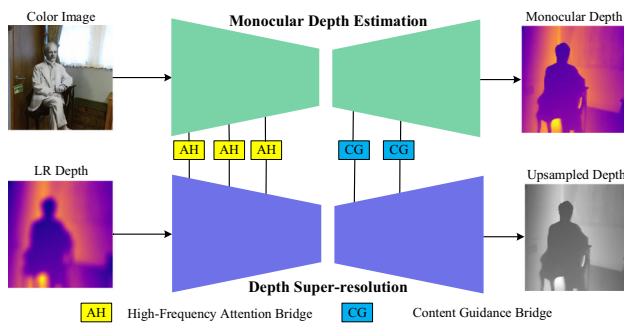

Fig. 8 LGR architecture (De Lutio et al., 2022) belonging to optimization-inspired architecture. The optimization layer is plugged into the end of the network to predict the high-resolution depth

Instead of deploying spatially-invariant kernels in vanilla convolutional neural networks (CNNs), some methods design learnable kernels to boost the GDSR performance. Kim et al. (2021) propose a deformable network that provides a custom convolution kernel for each pixel by computing neighborhood weights. The kernel weights can be obtained not only by calculating the pixels in regular positions (e.g., 8-neighborhood), but also by calculating its sub-pixels after interpolation. Wang et al. (2022c) introduce a continuous depth representation whose core is the proposed geometric spatial aggregator (GSA). The GSA comprises two parts: (1) the geometric encoder using the scale-modulated distance field to build the correlation between pixels and (2) the learnable kernel learning typical texture pattern processing prior. Thanks to the continuous representation, the model can super-solve depth at an arbitrary scale. Zhong et al. (2023a) design a kernel generation network that can cope with inconsistent structures between RGB and depth.

To increase the interpretability, Tang et al. (2021a) regards GDSR as a neural implicit interpolation problem, which can map continuous coordinates of LR depth and HR color images into latent codes. Meanwhile, they learn the interpolation weights to establish a unified framework to yield the interpolation weights and values. To promote the deployment of dToF cameras in mobile devices, Li et al. (2022c) perform depth estimation from a lightweight ToF sensor and RGB image (DELTAR). In this work, PointNet (Qi et al., 2017) and Efficient B5 (Tan & Le, 2019) are used to extract depth and RGB features, respectively, fused at the decoder stage through the proposed transformer-based fusion module.

3.2 Hybrid Architecture

As neural networks are usually described as black-box models, some researchers combine optimization methods to obtain explainable frameworks. Riegler et al. (2016a) present a novel network consisting of two sub-networks, i.e., a fully-convolutional network and a primal-dual network, with the first yielding HR depth maps and weights and the latter producing the final results with a non-local variational method.

As an optimization-inspired network, a weighted analysis sparse representation (WASR) model (Gu et al., 2019) is designed for GDSR. It leverages a neural network to learn filter parameterizations and non-linear functions to build more flexible stage-wise operations. Inspired by the multi-modal convolutional sparse coding model (Song et al., 2019; Deng & Dragotti, 2020) develop a network that can adaptively separate the shared features between different modalities from the unique features existing in a single modality. Marivani et al. (2020) propose a deep unfolding network to efficiently compute the convolutional sparse coding with the guide.

Zhao et al. (2022) introduce discrete cosine transform (DCT) into the proposed neural network to make the model explainable. This model utilizes DCT to tackle the optimization issue for GDSR by reconstructing the multi-channel HR depth features. Another advantage of using DCT is that it can make network design easier because it does not need to learn the mapping function between LR/HR image pairs, as it explicitly models it. De Lutio et al. (2022) propose a novel architecture learning graph regularization (LGR) with a neural network to perform GDSR, as shown in Fig. 8. First, a network extracts features from the source and the guide. A graph constructed using such features is then sent to a differentiable optimization layer to regularize it. Combining the advantages of model-based and learning-based methods, Zhou et al. (2023) propose a memory-augmented deep unfolding network in which local implicit prior and global implicit prior are introduced based on a maximal posterior view. To further prevent information loss, long short-term unit (LSTM) is employed in the persistent memory mechanism. Metzger et al. (2022) propose integrating guided anisotropic diffusion into a CNN to enhance depth discontinuities in the depth images. It has a fixed inference time and memory footprint, given a scale factor.

3.3 Auxiliary-Based Methods

Auxiliary learning, also known as multi-task learning, aims to find or design auxiliary tasks that can improve the performance of one or several primary tasks. Driven by its recent progress, several works attempt to exploit auxiliary task

Fig. 9 BridgeNet architecture (Tang et al., 2021b) belonging to auxiliary-based methods. The BridgeNet consists of a monocular depth estimation subnetwork and a guided depth upsampling subnetwork

learning to upsample depth maps. Sun et al. (2021) propose a knowledge distillation approach where depth estimation is employed as an auxiliary task during training to improve the results. As a representative method shown in Fig. 9, BridgeNet (Tang et al., 2021b) explores a paradigm where the association between monocular depth estimate (MDE) and depth super-resolution (DSR) is exploited deeply. To make the two tasks work together, they design two auto-encoders, namely DSRNet and MDENet, with information interaction at the encoder stage. In MDENet, a high-frequency attention bridge is proposed to capture HF information which can guide depth upsampling in the other task. In contrast, the content guidance is provided by DSRNet to guide monocular depth estimates. Considering the distinct learning curves of the two sub-tasks, they optimize the two sub-networks separately.

In addition to auxiliary learning, some researchers consider using auxiliary guidance from color images to improve depth quality. Based on the importance of edge information in depth upsampling, Wang et al. (2020) propose an edge-guided depth upsampling framework that leverages edge maps as the guide to recover HR depth maps. For training, the ground truth edge map is computed by the Canny operator on the corresponding HR depth map.

Differently from most existing GDSR methods, Voynov et al. (2019) attempt to gauge the upsampled depth quality through renderings of surface normal maps. More specifically, a visual appearance-based loss is proposed, which can assist a baseline network in yielding more visually pleasing results.
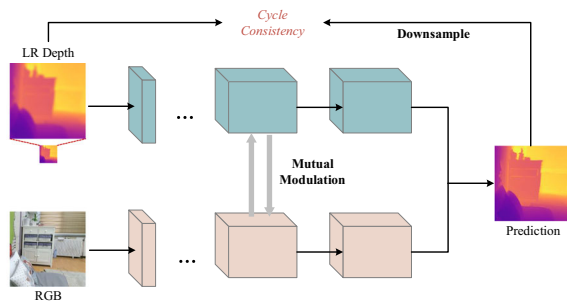
3.4 Fusion-Focused Methods

How to effectively fuse information from different modalities is critical to obtain high-quality depth maps. Intuitively, we can perform multi-modal fusion with a simple operation, such as concatenation (Wang et al., 2020; Tang et al., 2021a; De Lutio et al., 2022). However, these fusion schemes

can not selectively transfer HF features from the guide to the target, and may even result in texture-copying artifacts. Therefore, researchers design different fusion schemes to alleviate the issue, mainly grouped into three categories: early, late, and multi-level fusion. Li et al. (2020b) present a correlation-controlled color guidance block (block) to fuse the multi-modal information. Zhong et al. (2021) design a multi-modal attention-based fusion strategy, including a feature enhancement block and a feature recalibration block, with the former paying more attention to meaningful features extracted from depth maps and color images and the latter aiming at rescaling multi-modal features. This strategy can effectively avoid the texture-copying effect in the final prediction. Zhao et al. (2022) deploy an enhanced spatial attention block to transfer meaningful structural information from RGB to depth. Yuan et al. (2023a) develop a recurrent structure attention block, where the latest depth estimate and its corresponding HR color image are taken as the input, to obtain useful HF features of the color image. Wang et al. (2023b) propose a multi-perspective cross-guided fusion filter block, to improve depth features gradually by fusing the structure details in RGB images. In this block, spatial representations from various views are learned to capture depth salient structures further. Moreover, a color-depth cross-attention module is employed to achieve edge preservation. Zhong et al. (2023a) propose a multi-scale guided filtering module to refine the depth map in a coarse-to-fine manner. To obtain stable depth predictions, Sun et al. (2023) conduct cross-modal fusion in a multi-frame manner, i.e., dToF depth video super-resolution (DVSR) and histogram video super-resolution (HVSR), to upsample LR depth maps. Qiao et al. (2023b) propose an Adaptive Feature Fusion Module (AFFM), which enables the recovery of fine details from the HR guide and the LR depth maps. Zhao et al. (2023) propose a Spherical Space feature Decomposition network (SSDNet), which projects encoded features onto the spherical space. This strategy allows for separating and aligning domain-shared and domain-private features, respectively.

3.5 Unsupervised Methods

Since collecting training data paired with annotations is labor-intensive and time-consuming, researchers seek to solve the issue by leveraging self-supervised learning, allowing training models without labels. In Lutio et al. (2019), GDSR is regarded as a pixel-wise mapping from the source to the target, implemented as a multi-layer perceptron. This formulation can be trained in a fully unsupervised manner with the constraint of having the LR source. Shacht et al. (2021) present a single-pair method that can upsample depth maps even when the input pairs are misaligned. During training, patches cropped from the input pairs are utilized as pseudo-label data to enable weakly supervised learning. To

Fig. 10 MMSR framework (Dong et al., 2022) belonging to unsupervised methods. The MMSR adopts the cycle consistency loss to train the network in a self-supervised manner

improve generalization, Dong et al. (2022) propose a mutual modulation super-resolution model (MMSR), where a cross-modality modulation strategy using adaptive filters transfers meaningful features from one modality to the other, as shown in Fig. 10. In this mutual modulation, the spatial relationship between the corresponding pixels of the two modalities is fully exploited. Moreover, a cycle consistency loss is adopted to enforce the target faithful to the source. Differently from previous unsupervised methods for GDSR, Shin et al. (2023) introduce a pretext task into this field. Through the proposed scene structure guidance network, explicit structure features of color images can be obtained, which, together with the corresponding LR depth maps, serve as the input of a baseline network to produce high-resolution depth maps. Wang et al. (2023a) construct an adversarial network where the dependency between RGB images and depth maps is fully exploited to enhance the depth. Further, the optimal transport theory is introduced into the framework to boost depth enhancement performance. Qiao et al. (2023c) exploit a contrast learning scheme to extract unique features from the guidance, which can boost the GDSR performance based on a baseline network.

To conclude this section, Table 3 collects any of the methods discussed so far, divided into five categories. For each method, the venue and year are reported, together with a short description of the key idea behind it.

4 Guided Depth Completion

Pivotal works on depth completion (Uhrig et al., 2017; Chodosh et al., 2019) only take depth maps as input. However, due to the sparsity of depth maps obtained from long-range LiDAR measurements, a substantial amount of details, such as object boundaries, are lost. Thus, restoring these details is difficult on densified depth maps, even using deep learning. Therefore, researchers started exploiting additional modalities capturing the same scene to guide the completion of sparse depth maps. In practice, color images can provide rich

texture, scene structure, and object details, thus allowing for the extraction of sufficient cues to enhance depth completion. Besides, color images can also offer various auxiliary cues, such as semantics, monocular depth, and surface normal. Similar to GDSR, we divide GDC into two categories: supervised and unsupervised methods. Supervised methods can be further classified into dual-branch architectures, spatial propagation networks, auxiliary-based methods, and uncertainty guidance. These six categories are listed in the bottom branch of Fig. 5.

Most existing approaches employ supervised learning for this task, which typically constructs image pairs from dense maps and their sparse version for training. With dense depth information being used as annotation, deep neural networks are trained to solve a regression problem, as done for GDSR.

4.1 Dual-Branch Architecture

The dual-branch architecture is common in GDC. It utilizes two encoders to extract features from RGB and sparse depth maps, and then the cross-modal information is sent to a single decoder after aggregation. In this latter, a designed information fusion strategy can also be incorporated. Thus, two folds must be focused on: effective representation learning and information fusion.

Jaritz et al. (2018) employ a dual-branch architecture based on a modified version of the neural architecture search network (NASNet) (Zoph et al., 2018). First, the features from RGB images and depth maps are extracted using two NASNet encoders. Then, the intermediate features are fused in a late fusion scheme through channel-wise concatenation and fed into a decoder, yielding the final dense estimate. In this framework, validity masks are demonstrated to fail further to improve performance for GDC in large neural networks. To address the issue of model performance degradation caused by blurry guidance in the color images and unclear structures in the depth maps, Yan et al. (2022) exploit a repetitive design based on the hourglass network (Tang et al., 2020). To achieve multi-scale training, Li et al. (2020a) introduce the cascaded architecture into the hourglass network, extended by Fan et al. (2022) with denser interaction between the two modalities.

In the two-branch framework proposed by Van Gansbeke et al. (2019), global and local information are extracted with two encoders based on ERFNet (Romera et al., 2017), respectively. Notably, the local branch uses two U-nets to reconstruct sharp edges and structural details. Finally, the predictions are weighted by the outputs of each branch. Hu et al. (2021) propose a network that can perform precise and efficient depth completion (PENet), as depicted in Fig. 11. It gradually enhances the accuracy of depth prediction in a coarse-to-fine manner. The first branch, named the color-dominant branch, takes a color image and a sparse depth map

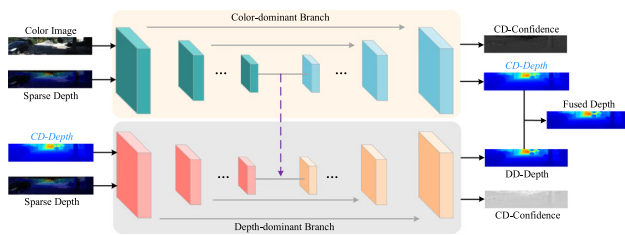**Table 3** List of Guided Depth Super-Resolution methods, divided according to the five outlined categories

Paradigm	Method	Reference	Key Idea
Dual-branch framework	DJFR (Li et al., 2019)	TPAMI-2019	Introduce two sub-networks for extracting features from the target and guidance images. Following the concatenation operation, the fused information is fed into another sub-network for the predicted depth.
	DMSG (Hui et al., 2016)	ECCV-2016	First work to super-solve depth maps using multi-scale guidance, which progressively enhances depth details from high- to low-level
	DepthSR (Guo et al., 2019)	TIP-2019	Exploit a residual U-Net structure, where hierarchical guidance is used at each scale, to recover HR depth maps.
	MGD (Zuo et al., 2019a)	TCSVT-2019	Propose a multi-scale architecture based on dense layers that aim to revisit the features from all higher levels at a given scale.
	MFR-SR (Zuo et al., 2019b)	TCSVT-2019	Combine global and local residual learning to upsamples depth maps from coarse to fine via multi-scale frequency synthesis.
	MIG-net (Zuo et al., 2021)	TMM-2021	Introduce a two-stream multi-scale network to extract features in the image and gradient domains, where the depth features and gradient features are alternatively complemented with each other.
	SFG (Yuan et al., 2023b)	AAAI-2023	Propose a structure flow-guided framework for GDSR, whose key point is to learn a structure flow map to guide structure representation transferring.
	CCFN (Wen et al., 2019)	TIP-2019	Propose a coarse-to-fine neural network with two stages. In the first stage, the model learns large filter kernels to obtain coarse results, which are further refined by smaller filter kernels in the second stage.
	CGN (Zuo et al., 2020b)	TMM-2020	With the proposed depth-guided affine transformation, the intensity-guided depth feature filtering and refinement are carried out iteratively in multiple stages.
	PMBANet (Ye et al., 2020)	TIP-2020	Introduce a multi-branch aggregation network with multiple stages to reconstruct sharp depth boundaries.
	IRLF (Song et al., 2020)	CVPR-2020	Develop a coarse-to-fine framework consisting of several sub-modules with channel attention strategies.
	DKN (Kim et al., 2021)	IJCV-2021	Design a deformable kernel network that obtains a kernel for each pixel based on the neighborhood weights.
	FDSR (He et al., 2021)	CVPR-2021	Explore high-frequency (HF) information via an HF guidance branch, where the HF information is fused with depth features to improve the performance.
	GeoDSR (Wang et al., 2022c)	AAAI-2023	Introduce a continuous depth representation to effectively implement scale-continuous and spatial-continuous upsampling in guided depth super-resolution.
	DAGF (Zhong et al., 2023a)	TNNLS-2023	Design a kernel generation network that can cope with inconsistent structures between RGB and depth, where a multi-scale guided filtering module is proposed to refine the depth map in a coarse-to-fine manner.
	DCSR (Wang et al., 2023d)	Display-2023	Proposes a depth map continuous SR framework that can achieve resolution adaptation at arbitrary SR ratios.

Table 3 continued

Paradigm	Method	Reference	Key Idea
	PAC (Su et al., 2019)	CVPR-2019	Present a pixel-adaptive convolution operation for deep joint image upsampling.
	DELTAR (Li et al., 2022c)	ECCV-2022	Propose a transformer-based architecture with two branches, which can extract and fuse depth and RGB features efficiently. Besides, a cross-modal calibration is performed to align RGB and depth.
Hybrid model	DPDN (Riegler et al., 2016a)	BMVC-2016	Use two sub-networks. one based on a fully-convolutional structure produces HR depth maps and weights. Another uses a non-local variational method to get final results.
	DG-CMM (Gu et al., 2019)	TPAMI-2019	Propose a weighted analysis sparse representation (WASR) model and a neural network to learn filter parameterizations and non-linear functions to build flexible stage-wise operations.
	CUNet (Deng & Dragotti, 2020)	TPAMI-2020	Develop a multi-modal convolutional sparse coding (MCSC) models to solve the general multi-modal image restoration and multi-modal image fusion problems.
	LMCSC (Marivani et al., 2020)	TIP-2020	Propose a deep unfolding network using guided image SR that exploits information from two modalities.
	LGR (De Lutio et al., 2022)	CVPR-2022	Propose a novel architecture combining a neural network and graph-based optimization to perform GDSR.
	DCTNet (Zhao et al., 2022)	CVPR-2022	Utilizes discrete cosine transform (DCT) to tackle the optimization issue for GDSR by reconstructing the multi-channel HR depth features.
	DaDa (Metzger et al., 2022)	CVPR-2023	Suggest incorporating guided anisotropic diffusion into a CNN to improve depth discontinuities in depth images. Given a scale factor, the inference time and memory demands of the method are fixed.
	MADUNet (Zhou et al., 2023)	IJCV-2023	Propose a memory-augmented deep unfolding network where the introduction of local and global implicit priors is based on a maximal a posterior view.
Auxiliary-based	CTKT (Sun et al., 2021)	CVPR-2021	Propose a knowledge distillation method that uses the depth estimate auxiliary task during training for better results.
	BridgeNet (Tang et al., 2021b)	ACM MM-2021	Explore a paradigm where the association between monocular depth estimate (MDE) and depth super-resolution are used to advance the performance of depth map super-resolution.
	DSR-N (Wang et al., 2020)	PR-2020	Propose a depth upsampling framework that leverages edge information as the guide to recover HR depth maps.
	PDSR (Voynov et al., 2019)	ICCV-2019	Attempt to measure the upsampled depth quality using renderings of surface normal maps.
	PDR-Net (Liu et al., 2022b)	Neurocomputing-2022	Proposes a progressive depth reconstruction network to further enhance the performance of DMSR.

Table 3 continued

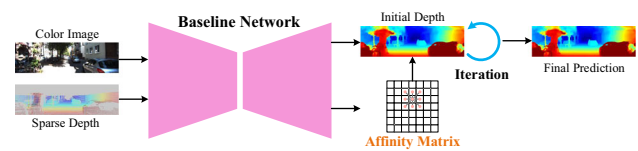| Paradigm	Method	Reference	Key Idea
Fusion-focused	SVLRM (Dong et al., 2021)	TPAMI-2021	Propose a new joint filtering method based on a spatially variant linear representation model where the target image is linearly represented by the guidance image.
	JIIF (Tang et al., 2021a)	MM-2021	Considers GDSR to be a neural implicit interpolation problem that can convert latent codes from continuous coordinates in LR depth and HR color images.
	MSSN (Li et al., 2020b)	PR-2020	Present a multi-scale strategy with the proposed symmetric unit as the basic component of the network, which can effectively deal with textureless and edge features in depth images.
	AHMF (Zhong et al., 2021)	TIP-2021	Propose a hierarchical network based on a bi-directional hierarchical feature collaboration module, where low- and high-level spatial information can work together to improve each other.
	RSAG (Yuan et al., 2023a)	AAAI-2023	Propose a recurrent structure attention guidance framework that employs both multi-scale and multi-stage strategies. A deep contrastive network with multi-scale filters is used to separate HF and LF features.
	JGF (Wang et al., 2023b)	PR-2023	Present a novel method that uses pyramid structure to capture multi-scale features, allowing HR depth maps to be recovered progressively.
	DVSR (Sun et al., 2023)	CVPR-2023	Propose a multi-frame scheme to upsample LR dToF videos with corresponding HR RGB sequences.
	SSDNet (Zhao et al., 2023)	ICCV-2023	Propose a spherical contrast refinement module to enhance depth details from RGB features.
	DSR-EI (Qiao et al., 2023b)	CVIU-2023	Propose AFFM to adaptively fuse discriminative cross-modality features.
Unsupervised	P2P (Lutio et al., 2019)	ICCV-2019	GDSR is regarded as a pixel-wise mapping from source to target and implemented as a multi-layer perceptron.
	CMSR (Shacht et al., 2021)	CVPR-2021	Present a single-pair method that can upsample depth maps even when the input pairs are misaligned, where patches cropped from the input pairs are utilized as pseudo-label data to enable weakly supervised learning.
	MMSR (Dong et al., 2022)	ECCV-2022	A cross-modality modulation strategy using adaptive filters is developed to transfer significant features from one modality to the other in a proposed mutual modulation super-resolution model.
	DEDE (Wang et al., 2023a)	TIP-2023	Design an adversarial network where the relationship between depth maps and RGB images are fully exploited to improve the depth.
	SSGnet (Shin et al., 2023)	AAAI-2023	Explicit structure features of color images can be extracted using the suggested scene structure guidance network, and these features-along with the corresponding LR depth maps-serve as the input of a base network.
	CMPNet (Qiao et al., 2023c)	NN-2023	Develop an autoencoder-based framework with contrastive and reconstruction losses to reduce information redundancy.

Fig. 11 PENet (Hu et al., 2021) framework belonging to dual-branch architectures. It consists of two branches that can respectively produce depth estimates from color-dominant and depth-dominant information

Fig. 12 CSPN framework (Cheng et al., 2018) belonging to affinity-based methods. At each step, CSPN propagates a local region in all directions at the same time

as input to produce a dense depth map that inherits sharp edges and structural details in the color image. The second branch, i.e., the depth dominant branch, employs a similar network, whose inputs are the previous dense and sparse depth map, to output another dense map. Finally, the two depth maps are fused using the strategy proposed by Fusion-Net (Van Gansbeke et al., 2019). Notably, they concatenate the 3D position maps to the convolutional layer, forming a geometric convolutional layer, to lift the performance further. For indoor depth completion, Jiang et al. (2022) use a dual-branch structure to achieve an efficient solution with low power consumption.

Using a dense pseudo-depth map as a coarse reference for the final prediction, Gu et al. (2021) can output stable and accurate dense depth maps. Different from the vanilla point-estimate methods with a dual-branch architecture, Yang et al. (2019) develop a conditional prior network (CPN) to calculate a posterior probability over the depth of each pixel, combined with a likelihood term using sparse measurements. The standard convolution operation applies a kernel to the pixel grid in an image, which, however, cannot handle sparse depth maps with uneven distribution of depth values. To address this issue, Zhao et al. (2021c) propose to leverage graph propagation to extract spatial contexts. Considering the variability of the graph structure, it applies the co-attention module (Lu et al., 2016) on the guided graph propagation for efficient multi-modal representation extraction. Combining traditional methods with a dual-branch deep network, Xu et al. (2023) achieve promising results. In the most recent works, Chen et al. (2023) employ two parallel Unet-shape networks to extract RGB and depth features at different scales. Then, feature fusion is performed in the decoding stage through the proposed attention-guided skip connection (AG-SC) module, yielding the desired results. In contrast, Yu et al. (2023) propose a network named PointDC that leverages 2D and 3D information to conduct robust depth completion.

Guided image filtering. In GDC, deep-guided image filtering can be viewed as a type of dual-branch architecture that utilizes features from RGB images to predict a kernel. This kernel is often learned by a deep network and applied to depth for feature fusion or dense depth prediction.

Inspired by classical guided image filtering (Tang & Manduchi, 1998; He et al., 2012; Tang et al., 2020) propose a guided network consisting of a GuideNet and a Depth-Net to recover dense depth maps. Inside the GuideNet, content-dependent and spatial-variant convolution kernels are predicted to capture geometric structural features in color images consistent with real-world scenes. Furthermore, a convolution factorization is introduced into the method to reduce computational overhead.

Liu et al. (2021c) present a two-stage model where depth interpolation and refinement are performed sequentially. Specifically, they first interpolate the sparse depth map via the proposed differentiable kernel regression layer, proving more effective than hand-designed filters. Then, a residual network based on U-net is adopted to refine the coarse depth map.

4.2 Affinity-Based Methods

In the form of a generic matrix, the affinity matrix expresses the pairwise proximity information between a set of data points. Since the spatial propagation network (SPN) for learning the affinity matrix was first employed for depth completion (Liu et al., 2017), many significant efforts have been made to improve it, yielding impressive results. In this pioneering work, the affinity matrix, which can establish associations between any two pixels over a whole image, is composed of the spatial transformation matrix. Given a hidden representation H^s at iteration t, the spatial propagation with affinity matrix w can be expressed as:

$$\mathbf{H}_{p,q}^{t+1} = (I - d_t)\mathbf{H}_{p,q}^t + \sum_{p,q \in N_{p,q}} w_{p,q}^{i,j}\mathbf{H}_{i,j}^t \tag{24}$$

where (p, q) and (i, j) are the positions of reference and neighbor pixels, respectively, $N_{p,q}$ denote the neighborhood of (p, q), and $d_t = 1 - \sum_{i,j \in N_{p,q}} w_{p,q}^{i,j}$.

Since spatial propagation is only performed along a certain direction at a time, the time cost of this method, which is run in sequence, is high. In contrast, Cheng et al. (2018) formulate the task as anisotropic diffusion filtering (Weickert et al., 1998; Liu et al., 2016), and implement it through a convolutional network named CSPN, depicted in Fig. 12. In detail, they first employ an Unet-based network that
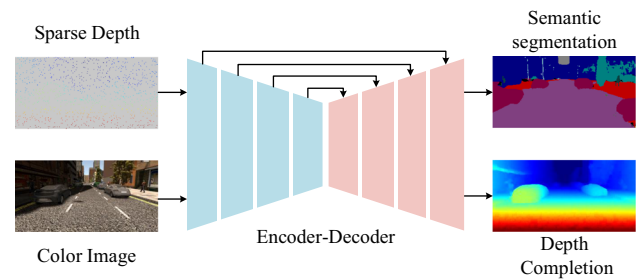
inputs an RGB image and a sparse depth map, and outputs a blurred depth map and an affinity matrix. Then, the propagation is carried out with recurrent convolutional operation. However, both approaches fix the receptive field in the propagation, which may yield suboptimal results for the task. To solve the problem, some methods utilizing flexible propagation schemes, such as CSPN++ (Cheng et al., 2020), non-local SPN (NLSPN) (Park et al., 2020), deformable SPN (DSPN) (Xu et al., 2020a) and dynamic SPN (DySPN) (Lin et al., 2022), are proposed.

To further improve the model performance for depth completion, Cheng et al. (2020) propose CSPN++, which includes two variants, i.e., context-aware CSPN (CA-CSPN) and resource-aware CSPN (RA-CSPN). By converting two constants in CSPN – the number of iterations and kernel size – into variables, CA-CSPN can adaptively assemble results from different steps in the propagation. However, in practice, the computational complexity is too high to be deployed in real applications. To speed up the algorithm, RA-CSPN picks the optimal kernel size and the number of iterations for each pixel depending on the learned hyper-parameters. With the improvements, the method can dynamically allocate the context and computing resources required by each pixel.

Unlike CSPN++, NLSPN (Park et al., 2020) is proposed with a two-stage network based on non-local neighbors instead of a fixed receptive field. In the first stage, a U-net is employed to produce a coarse depth, a confidence map, and non-local neighbors prediction with their raw affinities. Then, non-local spatial propagation computed by deformable convolutions (Zhu et al., 2019) is performed iteratively. In this manner, irrelevant features are avoided, while only relevant non-local neighborhood features are emphasized. Similar to NLSPN (Park et al., 2020), DSPN (Xu et al., 2020a) also conducts spatial propagation with deformable convolution, which can offer an adaptive receptive field according to the dependencies between pixels.

Later, motivated by dynamic convolution (Chen et al., 2020), Lin et al. (2022) develop a non-linear propagation model, named dynamic spatial propagation network (DySPN), where the affinity matrix is decoupled into parts depending on the distance. For neighborhoods with different distances, the affinity is assigned different weights to improve CSPN. Moreover, they propose a diffusion suppression operation, which adaptively stops iterations on specific pixels as directed by the attention matrix to avoid over-smoothing issues. According to different computational budgets, they design three variants of adaptive affinity matrix to achieve a balance between performance and complexity.

Liu et al. (2022d) integrate SPN and graph neural networks into one framework, allowing to learn both neighboring and long-range features. To adapt to practical applications, Schuster et al. (2021); Conti et al. (2023) propose novel approaches that can achieve good generalization. Zhang et al.

Fig. 13 MNASnet framework (Jaritz et al., 2018) belonging to auxiliary-based methods. It can perform depth completion and semantic segmentation simultaneously with minor changes in the last layer

(2023) adopt the non-local spatial propagation network for improving the depth quality after obtaining an initial densified depth map. As for the backbone that produces an initial depth, they propose a novel single-branch network, which leverages both the advantages of CNNs and transformers. In the network, they design a joint convolutional attention and transformer (JCAT) block that contains two paths – the transformer layer and the convolutional attention layer – and achieves more efficient performance than pure transformer-based methods. To learn the neighborhood affinity, Lee et al. (2022) propose a network that leverages multi-scale and local features. (Wang et al., 2023c) propose a long-short range recurrent updating (LRRU) network that leverages an iterative scheme and employs fewer parameters to achieve state-of-the-art performance.

4.3 Auxiliary-Based Methods

As described in Sect. 3.3, auxiliary learning employs several related tasks to improve the performance of the main task (Zhang & Yang, 2021b). Many methods using auxiliary learning also emerge in GDC. Atapour et al. (2019) propose a multi-task framework, which allows conducting monocular depth estimation and sparse depth completion. This network is implemented as two cascaded sub-networks based on encoder-encoder architecture, with the first yielding a sparse depth map from a color image and the second estimating a dense depth map, taking the previous sparse depth as input. To obtain sharp depth boundaries in the final depth maps, they adopt adversarial training on both synthetic and real-world datasets.

In addition to monocular depth estimation, semantic segmentation also contributes to the performance improvement of depth completion. Jaritz et al. (2018) perform multi-task learning, namely depth completion and semantic segmentation as shown in Fig. 13 by designing different heads at the end of a common network, e.g., NASNet (Zoph et al., 2018). When depth density varies, the method can still fuse features across modalities effectively. Unlike the previous method, which does not utilize results from the auxiliary task, Zhang

et al. (2021a) use semantic maps to improve the accuracy of depth completion. In order to integrate semantic segmentation and depth completion into a single framework, they build multi-task generative adversarial networks in combination with a semantic-guided smoothness loss, which is inspired by Zhao et al. (2020). In addition to semantic segmentation and depth completion, edge detection, which serves as a bridge between the two tasks, is introduced into the simultaneous semantic segmentation and depth completion multi-task network (SSDNet) (Zou et al., 2020a). Based on the multi-task learning of semantic segmentation and depth completion, Hirata et al. (2019) utilize boundary class labeling to adjust the effects of adjacent pixels. In Liang et al. (2019), the authors perform multiple perception tasks jointly, including 2D and 3D object detection, ground estimation and depth completion, to obtain better representations. Although this work is aimed at object detection, it demonstrates that the selected tasks are complementary, so helpful cues provided by the other tasks can also be applied to depth completion. Motivated by Lu et al. (2020), Ramesh et al. (2023) design an auxiliary task to generate depth contours while outputting dense depth maps at the end of the network.

Researchers leverage some further auxiliary information to enhance the precision of densified depth maps. Through a modified auto-encoder framework, Zhang and Funkhouser (2018); Qiu et al. (2019b) propose to estimate surface normal maps from color images to improve the depth accuracy. In Qiu et al. (2019b), the surface normals are regarded as intermediate representations since they can result in higher errors in distance measures. Considering the limitation of surface normals, the dense depth map estimated by the color branch is also considered in the generation of the final dense depth. Instead of generating a surface normal from a color map, Liu and Jung (2022a) take the normal map produced by sparse depth maps as an intermediate constraint during training. Xu et al. (2019) introduce a unified two-stage framework based on the assumption that natural scenes can be represented by piecewise planes. They build a diffusion module that incorporates the geometric constraints between depth and surface normals. This module is differentiable, allowing the diffusion conductance to be adjusted flexibly depending on the similarity in the guide. Inspired by Van Gansbeke et al. (2019); Hu et al. (2021); Nazir et al. (2022) added a semantic-guided branch to highlight depth discontinuities.

Instead of using surface normals as auxiliary information, some works utilize a point cloud to enhance model performance. In order to cope with unpredictable real-world environments, Jeon et al. (2021) propose a point-cloud-centric method based on the proposed attentive bilateral convolutional layer. This 3D convolution layer consists of four steps, i.e., splat, convolve, attention, and slice, which can directly perform convolution operations on point clouds and focus on capturing helpful features for GDC.

Fig. 14 The framework of FusionNet (Van Gansbeke et al., 2019) belonging to the category of fusion strategy. This method fuses global and local features based on the confidence maps in a late fusion manner

Sometimes, depth completion as auxiliary information can also improve the performance of other tasks. For instance, Carranza-García et al. (2022); Wu et al. (2022) integrate depth completion networks into object detection models to achieve more accurate results.

4.4 Fusion-Focused Methods

Similar to GDSR, fusion strategies are also crucial for GDC. Compared with straightforward fusion methods, such as concatenation (Eldesokey et al., 2019; Zhu et al., 2022) or summation (Shivakumar et al., 2019; Ryu et al., 2021), researchers design various complicated and effective methods.

Lee et al. (2020) propose a multi-scale network that exchanges information with the attention mechanism at each encoder scale. Yan et al. (2020b) employ a similar strategy to fuse cross-modality features at multiple scales. Zhao et al. (2021c) conduct multi-modal information fusion via the proposed symmetric gated fusion strategy consisting of two paths, each aiming to modulate the current modality features with adaptively extracted features from the other modality. Using this strategy, features can interact across modalities, thus allowing the model to obtain accurate depth estimates. Liu et al. (2021d) develop a novel coarse-to-fine framework that emphasizes the fusion of cross-modal information. It introduces a channel shuffle extraction operation to blend features from different modalities. Next, they propose an energy-based fusion operation, which selects feature values with higher regional energy to fuse the features effectively. Ke et al. (2021) propose a multi-modal deep aggregation network (MDANet), containing multiple connection and aggregation pathways to implement a deeper fusion of the two modalities. Liu et al. (2023a) design a multi-feature channel shuffle extraction, which utilizes the channel shuffle operation to fuse features from color images and depth maps in the encoding stage. Then, a multi-level weighted combination leveraging multi-scale features is introduced into the decoding process to enhance the information fusion effectiveness further.

4.5 Uncertainty Guidance

Most of the works surveyed so far do not take into account the reliability of the sparse depth when designing algorithms. Recently, two categories of uncertainty, i.e., aleatoric and epistemic uncertainty, have been considered in GDC with Bayesian deep learning. Aleatoric or stochastic uncertainty represents inherent randomness in each observation of the same experiment. In the task of GDC, aleatoric uncertainty refers to the random noise in raw LiDAR data and the distribution of sparse pixels. Epistemic or systematic uncertainty is often caused by models ignoring certain practical factors. Hence, the models for GDC are usually focused on handling aleatoric uncertainty.
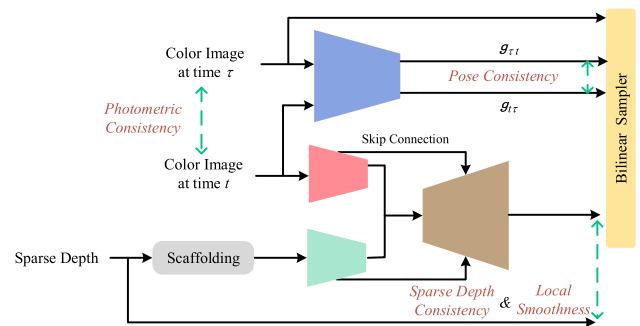
Eldesokey et al. (2018, 2019) introduce a novel framework with a proposed algebraically-constrained normalized convolution layer, where confidence can be computed and propagated to subsequent layers. Here, confidence is used as auxiliary information to improve the model performance. Furthermore, they propose a loss function by maximizing the final confidence to constrain the training. Later, Eldesokey et al. (2020) estimate the confidence of the input depth with the normalized convolution layer in an unsupervised manner, which can significantly improve the accuracy of depth prediction.

As shown in Fig. 14, Van Gansbeke et al. (2019) estimate confidence maps in the global and local branches, respectively. Based on the learned confidence maps calculated by the softmax function in the last layer, the model can pay more attention to reliable pixels. Sharing a similar idea, Qiu et al. (2019b) output confidence maps for occlusion handling and integrate depth predictions from color images and normal as well. Xu et al. (2019) obtain confidence maps through a decoder to suppress noise propagation in depth maps.

Zhu et al. (2022) provide a novel uncertainty formulation with Jeffrey's prior, which can improve the robustness of the model to noise or invalid pixels. In the first stage, this proposed multi-scale joint prediction model outputs coarse depth and uncertainty maps, which are then sent to the uncertainty attention residual learning network for depth refinement. Through their architecture and loss function, outliers and uneven distribution of pixels are well handled. Additionally, Peng et al. (2022) attempt to employ the proposed uncertainty-aware sampling algorithm to identify the reliability of pixels, which is beneficial for the subsequent adaptive prediction of the depth range for each pixel.

4.6 Unsupervised Methods

In practice, dense ground truth depth maps are hard to obtain due to the limitations of LiDAR imaging principles. Even a high-end LiDAR can only generate at most 30% valid pixels of the whole depth map (Uhrig et al., 2017). Thus, some

Fig. 15 VIO (Wong et al., 2020) framework belonging to unsupervised methods. Combining the four losses of photometric consistency, sparse depth consistency, pose consistency and local smoothness enables training the network in a fully unsupervised manner

methods attempt to address the problem with self-supervised learning, among which the representative classes build supervisory signals using monocular sequences or synchronized stereo pairs.

As depicted in Fig. 15, (Zhou et al., 2017) design a self-supervised framework in which only color and depth sequences must be provided instead of sparse/dense image pairs. The method produces single-view depth prediction and multi-view camera pose estimation from two independent networks. This principle is brought to GDC by (Ma et al., 2019) introducing three loss functions, i.e., sparse depth loss, photometric loss, and smoothness loss. Under this framework, the performance even outperforms some methods with semi-dense annotations. To increase the robustness to noise and outliers, (Qu et al., 2020) substituted the proposed least squares fitting model for the last convolutional layer of the self-supervised learning benchmark network (Godard et al., 2019). Wong et al. (2020) propose to use a piecewise planar scaffolding of a scene as supervision to implement self-supervised learning. Using a similar structure, Wong and Soatto (2021) leverage the proposed calibrated backprojection layers to improve the baselines, while Liu et al. (2022c) introduce knowledge distillation into depth completion. Kam et al. (2022) propose a two-branch network inspired by cost volume-based depth estimation, which can obtain informative 2D and 3D representations from cross-modality input with the lightweight structure design. In Jeon et al. (2022), they first perform mesh depth construction by leveraging point and line features, which, together with the color image, are fed into a mesh depth refinement module. Finally, the depth prediction can be obtained through the calibrated backprojection network (KBNet) (Wong & Soatto, 2021). While most networks are trained on real datasets, Wong et al. (2021); Lopez-Rodriguez et al. (2020) train their models on synthetic datasets, alleviating the lack of real image pairs with dense labels.

Table 4 List of Guided Depth Completion methods, divided according to the first three outlined categories

Paradigm	Method	Reference	Key Idea
Dual-branch	MNASNet (Jaritz et al., 2018)	3DV-2018	Leverage fused features from dual streams to yield the final dense estimation via a decoder.
	GuideNet (Tang et al., 2020)	TIP-2020	Propose a hourglass network for depth completion.
	KernelNet (Liu et al., 2021c)	TIP-2021	Present a two-stage model where depth interpolation and refinement are performed sequentially.
	RigNet (Yan et al., 2022)	ECCV-2022	Exploit a repetitive design based on the hourglass network.
	MSG-CHN (Li et al., 2020a)	WACV-2020	Introduce the cascaded architecture into the hourglass network.
	CDCNet (Fan et al., 2022)	BMVC-2022	
	Penet (Hu et al., 2021)	ICRA-2021	Enhances the accuracy of depth prediction in a coarse-to-fine manner.
	FusionNet (Van Gansbeke et al., 2019)	MVA-2019	Exploit confidence maps in the global and local branches to achieve a guided-base method.
	DenseLiDAR (Gu et al., 2021)	RAL-2021	Use a dense pseudo-depth map as a coarse reference for the final prediction.
	SDDC (Xu et al., 2023)	Visual Computer-2023	Propose the semantic-guided branch to achieve better prediction results.
	QNN (Jiang et al., 2022)	CVPR-2022	Adopt a dual-branch structure to achieve an efficient solution with low power consumption.
	ACMNet (Zhao et al., 2021c)	TIP-2021	Conduct multi-modal information fusion via the proposed symmetric gated fusion strategy.
	PointDC (Yu et al., 2023)	ICCV-2023	Propose a dual-branch network to extract 2d and sparse 3D feature point cloud for depth completion.
	AGGNet (Chen et al., 2023)	ICCV-2023	Design a dual-branch autoencoder with attention mechanism to obtain dense depth maps.
Affinity-based	SPN (Liu et al., 2017)	NeurIPS-2017	Exploit the spatial propagation network to learning affinity matrix for depth completion.
	CSPN (Cheng et al., 2018)	ECCV-2018	Formulate the task as anisotropic diffusion filtering, and implement it through a deep network.
	CSPN++ (Cheng et al., 2020)	AAAI-2020	Propose a new structure including context-aware and resource-aware CSPNs.
	NLSPN (Park et al., 2020)	ECCV-2020	Propose a two-stage network that is based on non-local neighbors instead of a fixed receptive field.

Table 4 continued

Paradigm	Method	Reference	Key Idea
	DSPN (Xu et al., 2020a)	ICIP-2020	Conduct spatial propagation with deformable convolution.
	DySPN (Lin et al., 2022)	AAAI-2022	Develop a non-linear propagation model, where the affinity matrix is decoupled into parts depending on the distance.
	GraphCSPN (Liu et al., 2022d)	ECCV-2022	Integrate SPN and graph neural networks to learn both neighboring and long-range features.
	Ssgp (Schuster et al., 2021)	WACV-2021	Propose a unified architecture to perform sparse-to-dense interpolation in different domains.
	SpAgNet (Conti et al., 2023)	WACV-2023	Propose a depth completion method that is independent of their distribution and density.
	CompletionFormer (Zhang et al., 2023)	CVPR-2023	Leverage the non-local spatial propagation network to improve the depth quality after obtaining initial depth with convolutions and transformers.
	DenseLConv (Lee et al., 2022)	IROS-2022	Propose a depth completion method that is independent of their distribution and density.
	LRRU (Wang et al., 2023c)	ICCV-2023	Propose a lightweight network to learn spatially-variant kernels and perform updates iteratively.
Auxiliary-based	DCMDE (Atapour et al., 2019)	3DV-2019	Propose a multi-task framework conducting monocular depth estimation and sparse depth completion.
	Multitask GANs (Zhang et al., 2021a)	TNNLS-2021	Present a multi-task network using semantic images to improve the accuracy of depth completion.
	SSDNet (Zou et al., 2020a)	sensors	Introduce edge detection to serve as a bridge between semantic segmentation and depth completion.
	MMF (Liang et al., 2019)	CVPR-2019	Combine 2D and 3D object detection, ground estimation, and depth completion to obtain better representations.
	DPS (Hirata et al., 2019)	RAL-2019	Present a method for estimating a dense depth from a sparse LIDAR point cloud and an image sequence.
	SIUNet (Ramesh et al., 2023)	WACV-2023	Propose a network regarding depth contours as auxiliary information to obtain dense depth maps.
	DDC (Zhang & Funkhouser, 2018)	CVPR-2018	Propose to estimate surface normal maps from color images to improve the depth accuracy.
	Deeplidar (Qiu et al., 2019b)	CVPR-2019	
	NNNet (Liu & Jung, 2022a)	IEEE Access-2022	
	DNC (Xu et al., 2019)	ICCV-2019	Confidence maps are obtained through a decoder to suppress noise propagation in depth maps.
	SemAttNet (Nazir et al., 2022)	IEEE Access-2022	Add a semantic-guided branch for depth completion to highlight depth discontinuities.
	ABCD (Jeon et al., 2021)	RAL-2021	Propose a point-cloud-centric method that is based on the proposed attentive bilateral convolutional layer.

Table 5 List of Guided Depth Completion methods, divided according to the last three outlined categories

Paradigm	Method	References	Key Idea
Fusion-focused	CGM (Liu et al., 2021c)	IEEE Access-2020	Present a two-stage model where depth interpolation and refinement are performed sequentially.
	RSIC (Yan et al., 2020b)	IEEE Access-2020	Employ a similar strategy to fuse cross-modality features at multiple scales.
	FCFR-Net (Liu et al., 2021d)	AAAI-2021	Develop a novel coarse-to-fine framework that emphasizes the fusion of cross-modal information.
	MDANet (Ke et al., 2021)	ICRA-2021	Propose a multi-modal deep aggregation block that consists of multiple connection and aggregation pathways for deeper fusion.
	MFF-Net (Liu et al., 2023a)	RAL-2023	Propose a multi-feature channel shuffle extraction and a decoding process with multi-scale features.
Uncertainty-guided	MS-Net (Eldesokey et al., 2019)	TPAMI-2019	Introduce a novel framework with a proposed algebraically-constrained normalized convolution layer.
	pNCNN (Eldesokey et al., 2020)	CVPR-2020	Estimate the confidence of the input depth with the normalized convolution layer in an unsupervised manner.
	MJPM (Zhu et al., 2022)	AAAI-2022	Provide Jeffrey's prior uncertainty formulation that can increase the model's resistance to noise and invalid pixel input.
	FusionNet (Wong et al., 2021)	RAL-2021	Leverage synthetic datasets to alleviate the lack of real image pairs.
	PADNet (Peng et al., 2022)	ACM MM-2022	Employ the proposed uncertainty-aware sampling algorithm to identify the reliability of pixels, which is beneficial for prediction of the depth range.
Unsupervised	SfMLearner (Zhou et al., 2017)	CVPR-2017	Single-view depth prediction and multi-view camera pose estimation from two independent networks.
	SSDC (Ma et al., 2019)	ICRA-2019	Propose learning a mapping from the sparse input to the dense prediction.
	lsf (Qu et al., 2020)	WACV-2020	Propose a least squares fitting model for the self-supervised learning benchmark network.
	VIO (Wong et al., 2020)	ICRA-2020	Use a piecewise planar scaffolding of a scene as supervision to achieve self-supervised learning.
	KBNet (Wong & Soatto, 2021)	ICCV-2021	Exploit the proposed calibrated backprojection layers to improve the baselines.
	Monitored Distillation (Liu et al., 2022c)	ECCV-2022	Introduce knowledge distillation into depth completion.
	CostDCNet (Kam et al., 2022)	ECCV-2022	Propose a two-branch network based on the cost volume-based depth estimation.
	Struct-MDC (Jeon et al., 2022)	RAL-2022	First perform mesh depth construction by leveraging point and line features.
	DDP (Yang et al., 2019)	CVPR-2019	Develop a conditional prior network to calculate a posterior probability over the depth of each pixel.
	AGG-CVCNet (Xu et al., 2022)	ACM MM-2022	Propose a novel adjacent geometry guided training loss.
	MSC (Zhang et al., 2022)	Remote Sensing-2022	Introduce multi-modal spatio-temporal consistency constraints to train models.
	LeoVR (Li et al., 2022a)	MobiSys-2022	

In unsupervised learning, stereo pairs also play an important role. Among the self-supervised stereo methods, most use photometric consistency (Godard et al., 2017; Yang et al., 2019; Shivakumar et al., 2019) to infer geometry and obtain promising results. In contrast, Xu et al. (2022) propose a novel adjacent geometry guided training loss to confine depth maps of low-confident regions by high-confident labels. Zhang et al. (2022); Li et al. (2022a) introduce multi-modal spatiotemporal consistency constraints to train models, which enables the model to better adapt to real-world environments, such as dark and low-texture objects.

To conclude this section, Tables 4 and 5 collect the methods discussed so far, divided into six categories. For each method, the venue and year are reported, together with a short description of the key idea behind it.

5 Benchmark Datasets and Objective Functions

We introduce existing datasets relevant to our survey and the most common loss functions used to train the methods we surveyed before.

5.1 Datasets

This section introduces public datasets for RGB guided ToF imaging; they are divided into low-resolution and sparse depth according to the image characteristics collected by ToF cameras.

5.1.1 Low-Resolution Depth

Here, we mainly introduce datasets composed of low-resolution images collected by RGB guided ToF cameras, which can be employed for depth super-resolution and depth completion through certain sampling strategies. We also introduce datasets that are commonly used in this field, such as Middlebury (Scharstein & Szeliski, 2003; Hirschmuller & Scharstein, 2007; Scharstein & Pal, 2007; Scharstein et al., 2014), NYU v2 (Silberman et al., 2012), MPI Sintel Depth (Butler et al., 2012) and ToF-FT3D (Wang et al., 2022b), although being collected by other sensors or synthesized.

ToFMark (Ferstl et al., 2013): This dataset captures 3 real-world scenes, i.e., Books, Shark and Devil, by PMD Nano ToF camera.

LU (Lu et al., 2014): This dataset consists of 6 RGBD images captured by the ASUS Xtion Pro camera, which usually serves as validation or test set.

SUN RGBD (Song et al., 2015): SUN RGBD dataset contains 10335 indoor image pairs acquired by four different sensors. This dataset is often applied to scene understanding

tasks. Additionally, this dataset can be used to investigate cross-sensor biases.

DIML (Cho et al., 2021): DIML is a large-scale RGBD database containing 2M RGBD frames. In order to obtain high-precision depth maps, they use Kinect v2 to capture indoor scenes and ZED stereo camera to capture outdoor scenes. In outdoor scenes, confidence maps of disparity are also provided.

RGBDD (He et al., 2021): RGBDD establishes the first depth map SR dataset, which can reflect the correspondence between real LR and HR depth maps. This dataset consists of 4811 images capturing various scenes, e.g., human body, stuffed dolls, toys and plants. Notably, LR and HR depth maps are collected by Huawei P30 Pro and Helios ToF camera, respectively.

Middlebury (Scharstein & Szeliski, 2003; Hirschmuller & Scharstein, 2007; Scharstein & Pal, 2007; Scharstein et al., 2014): The Middlebury dataset is a widely used dataset in the field of GDSR, and consists of four sub-datasets from different years. Middlebury 2003 (Scharstein & Szeliski, 2003), Middlebury 2005 (Hirschmuller & Scharstein, 2007), Middlebury 2006 (Scharstein & Pal, 2007), and Middlebury 2012 (Scharstein et al., 2014) are acquired by the stereo camera under indoor condition. These datasets use stereo image pairs to generate depth maps.

NYUv2 (Silberman et al., 2012): This dataset uses Microsoft Kinect v1 camera to capture 1449 images. Most previous works take the first 1000 RGBD images as the training set, and the rest of the 449 images as the test set. De Lutio et al. (2022); Metzger et al. (2022) randomly selected 849 images for training, 300 for validation, and 300 for evaluation.

MPI Sintel Depth (Butler et al., 2012): This dataset originates from an animated short film for flow evaluation. It provides synthetic video sequences, including 35 naturalistic scenes, for which depth maps are provided.

ToF-FT3D (Wang et al., 2022b): ToF-FT3D, also known as ToF-FlyingThings3D, is a synthetic dataset which captures 6250 different views using Blender. Similar to the FlyingThings3D (Mayer et al., 2016) for optical flow estimation, it consists of objects that fly along randomized 3D trajectories. Besides, this dataset provides ToF amplitude, ToF depth and RGB images with a resolution of 640 × 480.

5.1.2 Sparse Depth

In practical applications, sparse depth maps are usually produced by LiDAR and only contain about 4% valid pixels (Uhrig et al., 2017). Under some extreme conditions, even less than 1% of the pixels are effectively measured (Guizilini et al., 2020, 2021). Although sparse depth maps can be obtained from dense depth maps with sampling strategies, here we focus on datasets collected by LiDAR.

Table 6 RGB guided ToF Datasets, divided according to the depth map properties

	Dataset	References	Year	Sensor name	Capture condition	Modalities	Images
LR Depth	ToFMark	Ferstl et al. (2013)	2013	PMD Nano	Indoor	Color, Depth	3 Images
	Lu	Lu et al. (2014)	2014	ASUS Xiton Pro	Indoor	Color, Depth	6
	SUN RGBD	Song et al. (2015)	2015	Kinect v2	Indoor	Color, Depth	2860
	DIML	Cho et al. (2021)	2021	Kinect v2, ZED Stereo	Indoor, Outdoor	Color, Depth	>200 Scenes
	RGBDD	He et al. (2021)	2021	Huawei P30 Pro, Helios	Indoor, Outdoor	Color, Depth	4811 Images
	Middlebury	Scharstein and Szeliski (2003); Hirschmuller and Scharstein (2007); Scharstein et al. (2014)	2003 2007 2014	Structured Light Structured Light Stereo Camera	Indoor	Color,Depth	32 Images
	NYUv2	Silberman et al. (2012)	2012	Kinect V1	Indoor	Color, Depth	1449 Images
	MPI Sintel Depth	Butler et al. (2012)	2012	Blender	Indoor, Outdoor	Color, Depth Optical Flow	1064 Images
	ToF-FT3D	Wang et al. (2022b)	2022	synthetic	–	Color, Depth	6250 Views
Sparse depth	KITTI	Uhrig et al. (2017)	2017	Velodyne, Stereo Camera	Outdoor	Color, Depth	94k Frames
	DenseLivox	Yu et al. (2021)	2021	Livox LiDAR, RealSense d435i	Indoor, Outdoor	Color, Depth	6 Scenes
	KITTI-360	Liao et al. (2022)	2022	Velodyne, Stereo Camera	Outdoor	Color, Depth, GPS, IMU	150 Scenes
	Gibson	Xia et al. (2018)	2018	NaVis, Matterport Camera, DotProduct	Outdoor	Color, Depth	572 Scenes
	Virtual KITTI	Gaidon et al. (2016)	2016	Synthetic	Outdoor	Color;Depth	50 Videos
	Leddar Pixset	Déziel et al. (2021)	2021	Leddar Pixell LiDAR	Outdoor	Color, Depth IMU, Radar	29k Frames
	Waymo Perception	Sun et al. (2020)	2020	LiDAR	Outdoor	Color, Depth	1150 Scenes
	DDAD	Guizilini et al. (2020)	2020	Luminar-H2 LiDAR	Outdoor	Color, Depth	150 Scenes
	Near-Collision Set	Manglik et al. (2019)	2019	LiDAR (N/A), Stereo Camera	Indoor	Color, Depth	13658 Sequences

KITTI (Uhrig et al., 2017): KITTI is a dataset initially gathered by the Karlsruhe Institute of Technology (KIT) and Toyota Technological Institute at Chicago (TTI-C). As one of the most important datasets in the field of autonomous driving, it captures more than 93K depth maps with corresponding raw LiDAR scans and RGB images. The LiDAR and two color cameras used for data acquisition are Velodyne HDL-64E and Point Grey Flea 2 (FL2-14S3C-C), respectively. Based on the KITTI raw dataset, Uhrig et al. (2017) remove noise and artifacts in the scenes and aggregates multiple LiDAR scans over time to obtain denser ground truth depths, making this dataset suitable for GDC. There are 86 000, 7 000, and 1 000 image pairs for training, validation, and evaluation, respectively.

Virtual KITTI (Gaidon et al., 2016): This dataset is a photo-realistic synthetic video dataset that contains 50 videos with a total of 21 260 frames. Scenes in the dataset are generated with the Unity game engine and a real-to-virtual cloning method. It can be applied to a variety of tasks, such as object detection, object tracking, instance semantic segmentation, and more.

KITTI-360 (Liao et al., 2022): KITTI-360 is a newly collected large-scale dataset with rich sensory information and full annotations. This dataset captures over 320k images and 100k laser scans, including annotated static and dynamic 3D scene elements.

Gibson (Xia et al., 2018): Gibson contains 572 full buildings with 1447 floors covering a total area of 211k m^2. In addition to depth maps and the corresponding color maps, it provides normal maps and semantic object annotations.

DenseLivox (Yu et al., 2021): This dataset collects the images with a Livox Horizon LiDAR and Intel RalSense D435i camera. It contains dense, accurate depth as ground truth, with up to 88.3% valid pixel coverage.

Leddar Pixset (Déziel et al., 2021): Leddar Pixset is the first full-waveform flash LiDAR dataset for autonomous driving. It contains 29 000 frames in 97 sequences of various environments, weather conditions, and periods, annotated with more than 1.3 million 3D bounding boxes.

Waymo Perception (Sun et al., 2020): This dataset is collected by five LiDAR sensors and five high-resolution pinhole cameras, all synchronized and calibrated. It captures 1150 scenes, including urban and suburban areas, each lasting 20 s. Typically, 1 000 scenes are used for training and 150 for testing.

DDAD (Guizilini et al., 2020): DDAD is a new benchmark dataset for autonomous driving from Toyota Research Institute, which consists of monocular videos and accurate ground-truth depth. It has a long measurable distance, which can reach up to 250 m. There are 150 scenes with 12 650 individual image pairs in the training set, 50 scenes with 3 950 image pairs in the validation set, and 3 080 images in the test set.

Near-Collision Set (Manglik et al., 2019): Near-Collision Set is a large-scale, real-world dataset for near-collision prediction, collected with a stereo camera and LiDAR sensor. In this dataset, color images, accurate depth, and human pose annotations are provided. It contains 13 658 egocentric video snippets of humans navigating in indoor hallways.

To conclude, Tab. 6 lists the main datasets we discussed so far, categorizing them according to the input depth map being LR or sparse, as well as reporting the year of collection, the sensor used, the capture environment, the sensed modalities and the number of images.

5.2 Objective Functions

Notably, the performance of a trained model is directly related to the choice of loss functions, as they estimate the corrections to be applied to weights in the network during training. In this section, we discuss general loss functions for RGB guided ToF imaging, along with loss functions that are frequently employed for the two sub-tasks of GDSR and GDC, respectively.

5.2.1 General Losses

In RGB guided ToF imaging, the most frequently utilized losses are mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). MAE is used in many works (Jaritz et al., 2018; Shivakumar et al., 2019; Tang et al., 2021b, a), also referred to as l_1 loss

$$l_{MAE} = \frac{1}{N} \sum_{p \in N} \parallel Z_{hq}^p - \hat{Z}_{hq}^p \parallel_1 \qquad (25)$$

where $\parallel \cdot \parallel_1$ means the l_1 norm and N denotes the total number of valid pixels in depth maps. Some methods (Ma & Karaman, 21018; Zhao et al., 2022) adopted MSE (l_2 loss) which is defined as

$$l_{MSE} = \frac{1}{N} \sum_{p \in N} \parallel Z_{hq}^p - \hat{Z}_{hq}^p \parallel_2 \qquad (26)$$

where $\parallel \cdot \parallel_2$ represents the l_2 norm. Another common quantity is RMSE, i.e., the square root of MSE:

$$l_{RMSE} = \sqrt{\text{MSE}} \qquad (27)$$

As a perceptual metric for image quality assessment, the Structural Similarity Index (SSIM) is based on the visible structures and is defined as shown in Eq. 19.

$$l_{SSIM} = 1 - \text{SSIM}(p, q) \qquad (28)$$

5.2.2 Losses for GDSR

To achieve more robust training, a few approaches (Liu et al., 2021e) use the Charbonnier loss (Charbonnier et al., 1994):

$$l_{cha} = \sqrt{\| Z_{hq}^p - \hat{Z}_{hq}^p \|^2 + \epsilon^2} \tag{29}$$

where ϵ denotes a constant that ensures the penalty is non-zero. Sometimes, Peak Signal-to-Noise Ratio (PNSR) is also utilized for GDSR. PSNR can be formulated as follows:

$$l_{PSNR} = -10 log_{10}(\frac{Z_{max}^2}{\text{MSE}}) \tag{30}$$

where Z_{max} is the maximum depth value. The smaller the pixel value difference between the two depth maps, the higher the PSNR.

5.2.3 Losses for GDC

Depth loss. During training, since MAE treats all errors equally and MSE emphasizes the outliers, Huber loss combines the advantages of both:

$$l_{huber} = \begin{cases} \frac{1}{N}\sum_{p \in N} \frac{1}{2}(Z_{hq}^p - \hat{Z}_{hq}^p)^2, & |Z_{hq}^p - \hat{Z}_{hq}^p| \le 1 \\ \frac{1}{N}\sum_{p \in N}(|Z_{hq}^p - \hat{Z}_{hq}^p| - \frac{1}{2}), & |Z_{hq}^p - \hat{Z}_{hq}^p| > 1 \end{cases} \tag{31}$$

where $|\cdot|$ is the operator for absolute value. In Van Gansbeke et al. (2019), another loss named Focal-MSE, inspired by Lin et al. (2017), is proposed and proved to be better than MSE, which is formulated as:

$$l_{focal} = \frac{1}{N} \sum_{p \in N}(1 + \frac{\text{epoch}}{20}|Z_{hq}^p - \hat{Z}_{hq}^p|) \cdot (Z_{hq}^p - \hat{Z}_{hq}^p)^2 \tag{32}$$

Uncertainty-driven loss. Considering the uneven distribution in the captured depth maps from LiDAR, the uncertainty-driven loss is introduced (Xu et al., 2019; Eldesokey et al., 2020; Zhu et al., 2022) to focus on more reliable pixels. In their works, GDC is defined as maximizing the posterior probability. The final optimized loss with Jeffrey's prior (Figueiredo, 2001) can be written as

$$l_{ud} = \frac{1}{N} \sum_{p \in N}(e^{-s_n}(Z_{hq}^p - \hat{Z}_{hq}^p)^2 + 2s_p) \tag{33}$$

where s_n stands for prediction uncertainty at the pth pixel.

Adversarial loss. For method adopting a generative adversarial learning strategy (Khan et al., 2021; Wang et al., 2022a; Nguyen & Yoo, 2022) to perform depth completion, the adversarial loss is needed to discriminate between real and fake images, which can be written as:

$$l_{adv} = \min_G \max_D \mathbb{E}[\log D(Z_{hq})] + \mathbb{E}[\log(1 - D(G(Z_{lq}, I)))] \tag{34}$$

where $G(\cdot)$ and $D(\cdot)$ denote the generator and discriminator, respectively.

Normal constraint. To further improve depth accuracy, surface normal constraints have been exploited in several works (Zhang & Funkhouser, 2018; Lee et al., 2019; Xu et al., 2019). Zhang and Funkhouser (2018); Lee et al. (2019) compute the depth-normal consistency loss between the predicted normal $R(p)$ at pixel p and the estimated normal $v(p, q)$ from the depth map, as follows:

$$l_{n1} = \sum_{p,q \in N} \| < v(p, q), R(p) > \|^2 \tag{35}$$

where q is a neighbour of pixel p, and $< \cdot >$ denotes the inner product. To enable models to be trained in an end-to-end fashion, Xu et al. (2019) propose a negative cosine loss inspired by Eigen and Fergus (2015), which is defined as:

$$l_{n2} = -\frac{1}{N} \sum_{p \in N} R(p) \cdot \hat{R}(p) \tag{36}$$

where $\hat{R}(\cdot)$ is the computed normal vector.

Smoothness term. Based on the assumption of local smoothness, the smoothness term is introduced in many works (Ma et al., 2019; Shivakumar et al., 2019) to avoid local optimal solutions, penalizing gradients along $x-$ and $y-$ directions. In order to maintain the discontinuities in the final depth maps, some methods (Wong et al., 2020; Wong & Soatto, 2021; Choi et al., 2021; Song et al., 2021; Ryu et al., 2021; Wong et al., 2021) weight the gradients maps according to the local properties in RGB images:

$$l_{sm} = \frac{1}{N} \sum_{p \in N} \lambda_x(p)|\partial_x \hat{Z}_{hq}^p| + \lambda_y(p)|\partial_y \hat{Z}_{hq}^p| \tag{37}$$

where $\lambda_x(p) = e^{-|\partial_x I^p|}$ and $\lambda_y(p) = e^{-|\partial_y I^p|}$.

Photometric consistency. In unsupervised methods for GDC, photometric consistency based on the assumption of local smoothness is usually used to refine depth estimation. Given a reference image I_t and its neighboring frame I_τ along the temporal dimension where $\tau \in T = \{t-1, t+1\}$, the goal is to estimate the difference between I_t and its reconstructed version from I_τ as follows:

$$\hat{I}_\tau(x, \hat{Z}_{hq}) = I_\tau(\pi g_{\tau t} K^{-1} \overline{x} \hat{Z}_{hq}(x)) \tag{38}$$

where $\overline{x} = [x^T 1]^T$ are the homogeneous coordinates $x \subset \Omega$, π denotes the perspective projection, $g_{\tau t}$ is the relative pose of the camera from time t to τ, K represents the camera intrinsic matrix, and $\hat{Z}_{hq}(x)$ denotes the predicted depth of x. The relative pose can be either known or estimated by a network.

The photometric consistency is constructed by the combination of $L1$ loss and SSIM on the photometric reprojection as:

$$l_{ph} = \frac{1}{|\Omega|} \sum_{\tau \in T} \sum_{x \in \Omega} \omega_{co} |I_t(x) - \hat{I}_\tau(x)| +$$
$$\omega_{st}(1 - \text{SSIM}(I_t(x), \hat{I}_\tau(x))) \tag{39}$$

where ω_{co} and ω_{st} are weights to balance the two terms.

Pose consistency. Given a pair of images I_t, I_τ as input, a pose network predicts the relative pose $g_{\tau t} \in SE(3)$. When the input of the image sequence is reversed, we expect to obtain $g_t \tau$. The pose consistency loss is formulated as:

$$l_{pc} = \| \log(g_{\tau t} \cdot g_{t\tau}) \|_2^2 \tag{40}$$

where $\log : SE(3) \rightarrow se(3)$ is the logarithmic map.

Others. Finally, we review other commonly used loss functions in Table 7, including depth loss, structural loss, and others.

6 Evaluation

In this section, we compare the performance of the state-of-the-art GDSR and GDC approaches on widely used benchmarks. We first present the quantitative results for the two categories of RGB guided ToF imaging. Then, the pros and cons of the methods are analyzed. Notably, we focus more on the most recent methods for RGB guided ToF imaging.

6.1 Comparison of GDSR Methods

We select representative GDSR methods for the quantitative comparison on Middlebury, NYUv2, RGBDD, and DIML datasets with scaling factors of $4\times$, $8\times$, and $16\times$. For all the datasets, we follow the recent literature (Tang et al., 2021a; Zhao et al., 2022; Yuan et al., 2023a; Zhong et al., 2023a) to assess the depth quality according to the RMSE metric.

Table 8 shows the comparison of the state-of-the-art GDSR methods. In this experiment, bicubic interpolation is set as the baseline method. From this table, we can see that although the early deep learning methods (Hui et al., 2016; Li et al., 2016) are designed with simple structures, they show more powerful performance than traditional methods. Using an encoder-decoder architecture, DJFR (Li et al.,

2019), DepthSR (Guo et al., 2019) and PAC (Su et al., 2019) design two sub-networks that extract RGB and depth features respectively, and attain better performance compared to models built by several conventional CNN layers. In order to improve the representation capability of neural networks, PMBANet (Ye et al., 2020) design a complex structure and further boost the GDSR performance. Using a lightweight structure, FDSR (He et al., 2021) can still produce accurate results. As an alternative solution, the kernel prediction network (KPN) employed by DKN (Kim et al., 2021) and FDKN also achieves promising results. Further, Zhong et al. (2023a) employ the KPN to handle inconsistent structures between RGB and depth, which improves generalization to real scenes. CUNet (Deng & Dragotti, 2020) develops an optimization-based method and yields results comparable to KPNs. Inspired by these optimization-based methods, DCT-Net (Zhao et al., 2022) and LGR (De Lutio et al., 2022) attain better performance. LGR (De Lutio et al., 2022), in particular, achieves the state-of-the-art on Middlebury and DIML datasets. Unlike most methods that can only perform upsampling at fixed integer scales, GeoDSR (Wang et al., 2022c) learns a continuous representation and achieves strong performance in arbitrary scale depth super-resolution on the Middlebury dataset. With an effective multimodal feature fusion strategy, DSR-EI (Qiao et al., 2023b) attains the best results. However, the disadvantage is that it brings a large computational burden.

From another perspective, most state-of-the-art methods adopt multi-scale or coarse-to-fine strategies, such as SFG (Yuan et al., 2023b), JGF (Wang et al., 2023b), and RSAG (Yuan et al., 2023a). On the one hand, these strategies have achieved impressive results in other tasks, showing their effectiveness in learning feature representations. On the other hand, some methods, such as DCTNet (Zhao et al., 2022), LGR (De Lutio et al., 2022), DADA (Metzger et al., 2022) and SSDNet (Zhao et al., 2023), combine classic optimization algorithms with modern CNNs and demonstrate strong performance. Nonetheless, these models cannot be directly applied to many real scenarios, especially on mobile devices. Hence, efficient and effective approaches need to be further investigated.

6.2 Comparison of GDC Methods

This section gives the quantitative results achieved by several representative methods on the two benchmark datasets, i.e., KITTI driving scenes and NYUv2 indoor scenes datasets, as shown in Table 9. For the KITTI dataset, we use RMSE, MAE, iRMSE, and iMAE metrics to evaluate model performance. The results on NYUv2 are measured using RMSE and REL metrics.

From the table, we can see that unsupervised models have been making continuous progress. Even so, their per-

Table 7 Main loss terms deployed for GDSR and GDC

Types	Notation	Purpose
Depth loss	l_{berhu}	Berhu loss is a reversion of Huber loss l_{huber}
	l_{ce}	Cross-entropy is to measure the validaty of pixels
	l_{cyc}	Two l_1 losses for cycle consistency
Structural loss	l_{grad}	Gradient between the prediction and GT without weights
Others	l_{cpn}	RMSE between the prediction and its reconstruction from CPN
	l_{cos}	Cosine similarity measures the similarity between the prediction and GT
	l_{tp}	MAE between the initial depth and the prediction

Table 8 **GDSR – Results on Middlebury, NYUv2, RGBDD and DIML datasets.** The lower the RMSE, the better

Dataset	Middlebury			NYUv2			RGBDD			DIML		
Methods	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×
Bicubic	2.28	3.96	6.37	4.28	7.14	11.58	2.75	4.47	6.98	1.92	3.20	5.14
GF (He et al., 2010)	2.49	3.98	6.08	5.05	6.97	11.1	2.72	4.02	6.68	2.72	3.40	5.13
DMSG (Hui et al., 2016)	2.11	3.74	6.03	3.02	5.38	9.17	1.80	3.04	5.10	1.39	2.34	4.02
DJF (Li et al., 2016)	1.68	3.24	5.62	2.80	5.33	9.46	1.72	2.96	5.26	1.39	2.49	4.27
DJFR (Li et al., 2019)	1.32	3.19	5.57	2.38	4.94	9.18	1.50	2.72	5.05	1.27	2.34	4.13
DepthSR (Guo et al., 2019)	2.08	3.26	5.78	3.49	5.70	9.76	1.82	2.85	4.60	1.40	2.23	3.75
PAC (Su et al., 2019)	1.32	2.62	4.58	1.89	3.33	6.78	1.25	1.98	3.49	1.27	2.03	3.45
CUNet (Deng & Dragotti, 2020)	1.10	2.17	4.33	1.92	3.70	6.78	1.18	1.95	3.45	1.18	1.88	3.25
DKN (Kim et al., 2021)	1.23	2.12	4.24	1.62	3.26	6.51	1.30	1.96	3.42	1.27	1.86	3.22
FDKN (Kim et al., 2021)	1.08	2.17	4.50	1.86	3.58	6.96	1.18	1.91	3.41	1.13	1.84	3.29
PMBANet (Ye et al., 2020)	1.11	2.18	3.25	**1.06**	**2.28**	4.98	1.21	1.90	3.33	1.10	1.72	3.11
FDSR (He et al., 2021)	1.13	2.08	4.39	1.61	3.18	5.86	1.16	1.82	3.06	1.10	1.71	2.87
JIIF (Tang et al., 2021a)	1.09	1.82	3.31	1.37	2.76	5.27	1.15	1.77	<u>2.79</u>	1.17	1.79	2.86
SVLRM (Dong et al., 2021)	1.11	2.13	4.34	1.51	3.21	6.98	1.22	1.88	3.55	1.19	1.93	3.49
AHMF (Zhong et al., 2021)	1.07	1.63	3.14	1.40	2.89	5.64	1.10	1.73	3.04	1.10	1.70	2.83
DCTNet (Zhao et al., 2022)	1.10	2.05	4.19	1.59	3.16	5.84	1.08	1.74	3.05	1.07	1.74	3.09
LGR (De Lutio et al., 2022)	1.11	2.12	4.43	1.79	3.17	6.02	1.30	1.83	3.12	1.25	1.79	3.03
RSAG (Yuan et al., 2023a)	–	–	–	1.23	2.51	5.27	1.14	1.75	2.96	–	–	–
SFG (Yuan et al., 2023b)	–	–	–	1.45	2.84	5.56	–	–	–	–	–	–
DAGF (Zhong et al., 2023a)	1.15	1.80	3.70	1.36	2.87	6.06	1.17	1.75	3.10	1.15	1.76	3.16
JGF (Wang et al., 2023b)	–	–	–	1.57	2.48	**4.11**	1.25	1.85	3.03	–	–	–
DADA (Metzger et al., 2022)	<u>1.02</u>	1.72	3.16	1.55	2.88	5.34	1.12	1.70	2.89	<u>1.06</u>	<u>1.62</u>	<u>2.64</u>
GeoDSR (Wang et al., 2022c)	1.04	<u>1.68</u>	<u>3.10</u>	1.42	2.62	<u>4.86</u>	1.10	<u>1.69</u>	2.84	1.07	<u>1.62</u>	2.70
SSDNet (Zhao et al., 2023)	<u>1.02</u>	1.91	4.02	1.60	3.14	5.86	<u>1.04</u>	1.72	2.92	–	–	–
DSR-EI (Qiao et al., 2023b)	**0.97**	**1.53**	**2.32**	<u>1.21</u>	<u>2.46</u>	4.95	**0.91**	**1.37**	**2.10**	**0.69**	**1.19**	**1.96**

Best results are given in bold and second-best are underlined

formance still lags significantly behind their supervised counterparts. By using uncertainty as auxiliary information for depth, MS-Net (Eldesokey et al., 2019), pNCNN (Eldesokey et al., 2020), MJPM (Zhu et al., 2022) and PADNet (Peng et al., 2022) can filter out noisy input and invalid pixels, thus achieving better performance than versions without using uncertainty. As another auxiliary-based method, multi-task learning has been adopted by various methods to improve generalization performance. The results

produced by Multitask GANs (Zhang et al., 2021a) and SIUNet (Ramesh et al., 2023) demonstrate the effectiveness of this kind of approach. Networks using dual-branch structure (Tang et al., 2020; Liu et al., 2021c; Zhao et al., 2021c; Gu et al., 2021; Chen et al., 2023; Yu et al., 2023) also attain good performance. These methods fully leverage multi-scale features to recover the details of depth maps progressively. Based on the framework, RigNet (Yan et al., 2022) obtains the best results on NYUv2.

We also analyze the performance of the affinity-based methods. The affinity matrix produced by these methods, such as NLSPN (Park et al., 2020) and GraphCSPN (Liu et al., 2022d), are shown to improve depth accuracy through spatial propagation consistently. In addition to spatial propagation, DySPN (Lin et al., 2022) employs the attention mechanism to achieve excellent results. Furthermore, CompletionFormer (Zhang et al., 2023) combines CNN with a recently popular and dominant technique, Vision Transformer, to achieve better performance in depth completion. In the most recent work, LRRU (Wang et al., 2023c) outperform other state-of-the-art approaches by learning spatially-variant kernels and updating iteratively.

7 Discussion and Future Trends

Although some reviewed methods have achieved promising results on RGB guided ToF imaging, some issues hinder their practical application. Here, we mainly describe the challenges and future trends from the following four aspects.

7.1 Unsupervised Methods with Lightweight Structure

As described before, most methods perform GDSR and GDC after training with a supervised paradigm. However, labeled data is usually difficult to collect. In addition, the performance of mass-produced sensors generally fluctuates to a certain extent, and it is unrealistic to fine-tune the model on data collected from any existing sensor. Therefore, unsupervised learning methods need to be developed without ground truth depth. In particular, state-of-the-art unsupervised learning methods cannot achieve the performance of supervised learning methods, so there is still much room for improvement in this direction.

On the other hand, these methods often require many parameters and much computation to achieve better results, which makes the models unable to run on many devices at high speed, especially widely used mobile devices. While transferring the information to the cloud for processing is one solution to this problem, it raises thorny privacy issues. Thus, efficiently recovering high-quality depth is necessary and remains an open challenge. One recent example is SeaFormer (Wan et al., 2023), which proposes a versatile mobile-friendly backbone based on the vision transformer. Developing a simple yet efficient unsupervised model for RGB guided ToF imaging is challenging but highly desirable.

7.2 Cross-Domain Generalization

Another issue concerns the ability of networks to generalize to new domains. This topic has been mostly overlooked in the literature despite its relevance for the widespread diffusion of depth estimation in practical applications. Purposely, Bartolomei et al. (2024) proposed a method to improve cross-domain generalization for depth completion by exploiting the virtual projection pattern paradigm proposed in Bartolomei et al. (2023) and casting depth completion as a correspondence problem using a virtual stereo setup. Processing properly hallucinated fictitious stereo pairs with a robust stereo network that possibly exploits the RGB image context, such as RAFT Lipson et al. (2021a), leads to more robust cross-domain generalization than state-of-the-art depth completion networks.

7.3 ToF with Varifocal Lens Color Camera

RGB cameras in many devices, such as mobile phones and drones, have variable focal lengths, while ToF cameras usually employ fixed focal length lenses. Different focal lengths correspond to different depths of field and angles of view. When a ToF camera captures a distant object, boundaries are usually blurred. If the color image autofocuses on the object, it can provide rich structural details and transfer to the depth map. Nonetheless, it also brings a problem that the scales of the images from the two modalities may differ. In this case, the RGB image and the depth map cannot be directly aligned by the registration technique, i.e., a new calibration is required. Hence, further investigation needs to be conducted on how to align two scale-inconsistent images, i.e., the RGB image and the depth map, and simultaneously transfer the scene geometry in the RGB image to the depth map.

7.4 Spot-ToF

The main disadvantage of ToF cameras using floodlight projection, currently the most commonly used technique in mobile devices, is their close operating distance. The scheme proposed by Achar et al. (2017) can significantly extend the measurable distance of the ToF camera, but the long acquisition time of each image limits the practical application. On the other hand, direct ToF (d-ToF) can capture distant objects, whereas the resolution of the sensors is much lower than that of indirect ToF (i-ToF) with tens of thousands of pixels, e.g., 8×8 or 24×24. LiDAR also faces the same problem, and its manufacturing cost is too expensive, making it currently unable to be used as a consumer-grade camera. Ge et al. (2021) provide an alternative to alleviate the issue. Specifically, they replaced the optical diffuser in the ToF camera projector with a diffraction optical element (DOE), which is used to evenly distribute the incident light into thousands of laser beams. In this way, each laser beam can have a mea-

Table 9 GDC – Results on KITTI and NYU-v2 datasets. **Un** and **Su** indicate unsupervised and supervised models. The lower the RMSE, the better

Method	Loss function	Learning	KITTI Dataset				NYUv2 Dataset	
			RMSE	MAE	iRMSE	iMAE	RMSE	REL
DDP (Yang et al., 2019)	$l_1 + l_{cpn} + l_{SSIM}$	Un	1263.19	343.46	3.58	1.32	–	–
VIO (Wong et al., 2020)	$l_{ph} + l_1 + l_{pc} + l_{sm}$	Un	1169.97	299.41	3.56	1.20	–	–
lsf (Qu et al., 2020)	$l_1 + l_{SSIM}$	Un	885.00	225.20	3.40	–	0.134	–
KBNet (Wong & Soatto, 2021)	$l_{ph} + l_1 + l_{sm}$	Un	1069.47	256.76	2.95	1.02	–	–
CSPN (Cheng et al., 2018)	l_2	Su	1019.64	279.46	2.93	1.15	0.117	0.016
MS-Net (Eldesokey et al., 2019)	l_1	Su	859.22	207.77	2.52	0.92	0.117	0.016
Deeplidar (Qiu et al., 2019b)	$l_2 + l_{n2}$	Su	758.38	226.50	2.56	1.15	–	–
pNCNN (Eldesokey et al., 2020)	l_2	Su	988.57	228.53	2.71	1.00	-	-
NLSPN (Park et al., 2020)	l_1 or l_2	Su	741.68	199.59	1.99	0.84	0.092	<u>0.012</u>
GuideNet (Tang et al., 2020)	l_2	Su	736.24	218.83	2.25	0.99	0.101	0.015
KernelNet (Liu et al., 2021c)	$l_1 + l_{ce} + l_{grad}$	Su	785.06	218.60	2.11	0.92	0.111	0.015
ACMNet (Zhao et al., 2021c)	$l_2 + l_{sm}$	Su	744.91	206.09	2.08	0.90	0.105	0.015
DenseLiDAR (Gu et al., 2021)	$l_2 + l_{grad} + l_{SSIM}$	Su	755.41	214.13	2.25	0.96	–	–
RigNet (Yan et al., 2022)	l_2	Su	712.66	203.35	2.08	0.90	<u>0.090</u>	0.013
Ssgp (Schuster et al., 2021)	l_2	Su	833.00	204.00	–	–	–	–
DenseLConv (Lee et al., 2022)	l_2	Su	729.88	210.06	2.10	0.93	0.099	0.015
GraphCSPN (Liu et al., 2022d)	l_1	Su	738.41	199.31	1.96	0.84	<u>0.090</u>	<u>0.012</u>
SpAgNet (Conti et al., 2023)	l_1	Su	844.79	218.39	-	-	0.114	0.015
Multitask GANs (Zhang et al., 2021a)	$l_{adv} + l_2 + l_{cyc}$	Su	746.96	267.71	2.24	1.10	-	-
SIUNet (Ramesh et al., 2023)	l_1	Su	1026.61	227.28	2.73	0.96	0.138	0.015
FCFR-Net (Liu et al., 2021d)	$l_1 + l_2$	Su	735.81	217.15	2.20	0.98	0.106	0.015
MFF-Net (Liu et al., 2023a)	$l_1 + l_2$	Su	719.85	208.11	2.21	0.94	0.100	0.015
MJPM (Zhu et al., 2022)	$l_1 + l_2$	Su	795.43	<u>190.88</u>	1.98	0.83	-	-
PADNet (Peng et al., 2022)	$l_1 + l_{focal}$	Su	746.19	197.99	1.96	0.85	0.094	<u>0.012</u>
DySPN (Lin et al., 2022)	$l_1 + l_2$	Su	709.12	192.71	<u>1.88</u>	<u>0.82</u>	<u>0.090</u>	<u>0.012</u>
CompletionFormer (Zhang et al., 2023)	$l_1 + l_2$	Su	<u>708.87</u>	203.45	2.01	0.84	<u>0.090</u>	<u>0.012</u>
AGGNet (Chen et al., 2023)	$l_2 + l_{huber}$	Su	–	–	–	–	0.092	0.014
PointDC (Yu et al., 2023)	$l_1 + l + 2 + l_{grad}$	Su	736.07	201.87	1.97	0.87	**0.089**	<u>0.012</u>
LRRU (Wang et al., 2023c)	$l_1 + l_2$	Su	**696.51**	**189.96**	**1.87**	**0.81**	0.091	**0.011**

Best results are given in bold and second-best are underlined

surable distance greater than that of floodlight illumination. However, this approach faces two challenges. First, a laser beam forms a spot in the sensor, corresponding to several pixels. The uneven intensity distribution of the spot may lead to inconsistencies in the measurement, so it is necessary to develop new depth correction methods to obtain an accurate depth map. Second, the captured images are sparse (similar to LiDAR but at a lower cost), so effective and efficient algorithms need to be developed to estimate dense depth under this setting.

7.5 Under-display RGBD Camera

In recent years, in pursuit of a better visual experience, full-screen devices have attracted the attention of industry and academia. This fact also brings a major challenge: images captured by under-display cameras are severely degraded, requiring effective strategies for restoration. Most of the existing work is to study the image restoration of under-display RGB cameras (UDC), while few works focus on under-display ToF (UD-ToF) depth restoration.

Under-display camera. Recently, smartphones with full-screen displays, eliminating the need for bezels, have become a new product trend and motivated manufacturers to design a brand-new imaging system, the Under-Display Camera. Placing a display in front of the camera lens can increase the screen-to-body ratio for a better user experience. However, the broad commercial manufacturing of UDC is prevented by poor imaging quality due to the inevitable degradations, such as blurs, noise, diffraction artifacts, color shifts, etc.

UDC image restoration has been studied in a few literary works. As the pioneered of this work, Zhou et al. (2021) devise a Monitor Camera Imaging System (MCIS) to collect paired data and offer two practical solutions for UDC image restoration, including a deconvolution-based Wierner Filter pipeline and the data-driven method. Yang and Sankaranarayanan (2021) redesign the display layout to improve the quality of recovered images. To address spatially variant blur and noise, Kwon et al. (2021) develop a controllable image restoration algorithm, which performs well on both a monitor-based aligned dataset (Zhou et al., 2020) and a real-world dataset. Feng et al. (2021) propose a domain-knowledge-based network to restore the UDC images. However, this work requires a point-spread function (PSF) as a prior, thus failing to work well when the PSF is unavailable. To remedy this, Koh et al. (2022) design an effective two-branched network for UDC image restoration from the perspective of low-frequency and high-frequency information learning. The UDC image restoration methods need paired image datasets that perform poor generalization capability for real degradations. The reason behind this phenomenon is the domain discrepancy between synthesized data and real-captured data. To alleviate this problem, Feng et al. (2023) employ collected non-aligned data for UDC image restoration and improve the robustness of the restoration network in real-world scenarios.

Under-display ToF. Currently, only one work (Qiao et al., 2023a) has investigated image restoration with a ToF camera placed under a display. In this paper, they propose a cascaded network to restore the depth in a coarse-to-fine manner. Differently from previous methods, in the first stage, they design a complex-valued neural network to recover the ToF raw measurements. In the second stage, they refine the depth based on the proposed multi-scale depth enhancement block. Moreover, to enable the training, they introduce real and synthetic datasets for UD-ToF imaging, respectively. Specifically, the large-scale synthetic dataset is created by analyzing noise in magnitude and phase.

If a color camera is located under a display with a ToF camera, there is greater potential for obtaining high-quality depth maps. However, the design of models to jointly compensate for multi-modal image degradation is still challenging.

7.6 Hybrid Models

Typically, RGBD-based imaging can be addressed by either conventional or deep-learning methods, each of which possesses advantages and disadvantages. A major advantage of conventional methods is that, through well-interpretable hand-crafted models, the predicted depth can be guaranteed to be loyal to the source. However, designing a model with excellent feature representation capabilities is usually extremely challenging. In contrast, deep-learning methods

can perform very well on a specific task with sufficient and representative data for training. On the other hand, when there is insufficient data or a biased distribution, these trained models may not achieve the desired results during testing. Therefore, combining the advantages of both to design hybrid models is one of the future trends in RGB guided ToF camera imaging. In particular, the hand-crafted prior, complemented by the powerful feature representation capability of neural networks, can make it easier to achieve a lightweight approach.

7.7 Online Alignment of RGB and Depth

In previous work, RGB and depth maps were usually assumed to have been well-aligned in spatial by calibration operations. However, in practice, the intrinsic c_x, c_y and extrinsic parameters t_x, t_y change when the camera is deployed on a device or subjected to vibration, which requires online weakly calibration for the RGB guided ToF camera. On the other hand, cross-modal online alignment in the time dimension is also necessary for practical applications, as different sensors may not achieve completely accurate synchronization in sensing the environment. Especially when the acquisition is out of sync during motion, the imaging quality of the RGBD camera will be seriously degraded. Although some previous efforts (Qiu et al., 2019a) have been devoted to cross-modal correspondence matching, the algorithms consume a long inference time, so temporally and spatially efficient alignment methods remain to be further explored.

7.8 Lightweight ToF

To promote the application of ToF cameras in various consumer electronic devices, lightweight ToF cameras Li et al. (2022c); Liu et al. (2023b) were launched on the market and have received the favor of many manufacturers for their low power consumption, low cost, and compact structure. Unlike common ToFs, lightweight ToF cameras generate measurements as depth distributions rather than specific values. Specifically, this type of camera can be used in many fields, such as AR/VR and obstacle avoidance. Nonetheless, the lightweight design inevitably brings drawbacks such as extremely low resolution (e.g., 8×8), which leads to a degradation in the depth quality and cannot support applications requiring high-quality depth. Therefore, using other modal information, such as RGB or IR, to improve the resolution of lightweight ToF depth becomes one of the future trends.

7.9 Multi-frame Processing

Currently, most algorithms deal with RGB guided ToF imaging in a single-frame fashion, using one degraded depth map

as input to infer the corresponding high-quality output. However, the format we usually capture with cameras is a data stream (video) rather than a single image. Therefore, it is crucial to consider the temporal correlation of data and output stable depth maps. There have been some works Wronski et al. (2019); Patil et al. (2022) employing multi-frame processing technology for image denoising, enhancement, and super-resolution, but only a few works Sun et al. (2023) have applied it to ToF depth improvement. Based on the above analysis, fusing ToF depth and RGB images with multi-frame processing technology is a promising direction.

8 Conclusion

In this paper, we have presented a survey of RGB guided ToF imaging methods based on deep learning. Our review covers preliminaries, evaluation metrics, network design, learning protocols, benchmark datasets, and objective functions. According to the measurable distance and the purpose of the ToF camera, we roughly divide the problem faced into two categories: guided depth super-resolution and guided depth completion. Moreover, we collect a quantitative comparison of the surveyed methods on widely used benchmarks and analyze their performance and respective characteristics. Finally, we summarize the current challenges in practical applications as well as promising future trends.

References

Achar, S., Bartels, J. R., Whittaker, W. L., Kutulakos, K. N., & Narasimhan, S. G. (2017). Epipolar time-of-flight imaging. *ACM Transactions on Graphics (ToG), 36*(4), 1–8.

Atapour-Abarghouei, A. ,& Breckon, T.P. (2019). To complete or to estimate, that is the question: A multi-task approach to depth completion and monocular depth estimation. In *2019 international conference on 3d vision (3dv)* (pp. 183–193).

Bartolomei, L. , Poggi, M. , Conti, A. , Tosi, F. , & Mattoccia, S. (2024). Revisiting depth completion from a stereo matching perspective for cross-domain generalization. In *International conference on 3d vision* 2024 (3dv 2024).

Bartolomei, L. , Poggi, M. , Tosi, F. , Conti, A. , & Mattoccia, S. (2023). Active stereo without pattern projector. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)* (p. 18470–18482).

Butler, D.J. , Wulff, J. , Stanley, G.B. , & Black, M.J. (2012). A naturalistic open source movie for optical flow evaluation. In *Computer vision–eccv 2012: 12th european conference on computer vision, florence, italy*, October 7–13, 2012, proceedings, part vi 12 (pp. 611–625).

Carranza-García, M., Galán-Sales, F. J., Luna-Romera, J. M., & Riquelme, J. C. (2022). Object detection using depth completion and camera-lidar fusion for autonomous driving. *Integrated Computer-Aided Engineering, 29*(3), 241–258.

Charbonnier, P. , Blanc-Feraud, L. , Aubert, G. , & Barlaud, M. (1994). Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st international conference on image processing* (Vol 2, pp. 168–172).

Chen, D. , Huang, T. , Song, Z. , Deng, S. , &Jia, T. (2023). Agg-net: Attention guided gated-convolutional network for depth image completion. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 8853–8862).

Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., & Liu, Z. (2020). Dynamic convolution: Attention over convolution kernels. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11030–11039).

Cheng, X., Wang, P., Guan, C., & Yang, R. (2020). Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *Proceedings of the aaai conference on artificial intelligence* (Vol 34, pp. 10615–10622).

Cheng, X., Wang, P., & Yang, R. (2018). Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the european conference on computer vision (eccv)* (pp. 103–119).

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

Chodosh, N., Wang, C., & Lucey, S. (2019). Deep convolutional compressed sensing for lidar depth completion. In *Computer vision–accv 2018: 14th asian conference on computer vision, perth, Australia*, December 2–6, 2018, Revised selected papers, part i 14 (pp. 499–513).

Choi, J., Jung, D., Lee, Y., Kim, D., Manocha, D., & Lee, D. (2021). Self-deco: Self-supervised monocular depth completion in challenging indoor environments. In *2021 ieee international conference on robotics and automation (icra)* (pp. 467–474).

Chugunov, I., Baek, S.-H., Fu, Q., Heidrich, W., & Heide, F. (2021). Mask-tof: Learning microlens masks for flying pixel correction in time-of-flight imaging. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9116–9126).

Cong, R., Lei, J., Fu, H., Huang, Q., Cao, X., & Ling, N. (2018). Hscs: Hierarchical sparsity based co-saliency detection for rgbd images. *IEEE Transactions on Multimedia, 21*(7), 1660–1671.

Conti, A., Poggi, M., & Mattoccia, S. (2023). Sparsity agnostic depth completion. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 5871–5880).

De Lutio, R., Becker, A., D'Aronco, S., Russo, S., Wegner, J.D., & Schindler, K. (2022). Learning graph regularisation for guided super-resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 1979–1988).

Deng, P., Ge, C., Qiao, X., & Wei, H. (2022). Multi-stream face anti-spoofing system using 3d information. In *2022 ieee international conference on consumer electronics (icce)* (pp. 1–6).

Deng, X.,& Dragotti, P.L. (2019). Coupled ista network for multi-modal image super-resolution. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1862–1866).

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*(10), 3333–3348.

Déziel, J.-L., Merriaux, P., Tremblay, F., Lessard, D., Plourde, D., Stanguennec, J., ... & Olivier, P. (2021). Pixset: An opportunity for 3d computer vision to go beyond point clouds with a full-waveform lidar dataset. In *2021 ieee international intelligent transportation systems conference (itsc)* (pp. 2987–2993).

Diebel, J., & Thrun, S. (2005). An application of markov random fields to range sensing. *Advances in Neural Information Processing Systems, 18*, 291–298.

Dong, J., Pan, J., Ren, J. S., Lin, L., Tang, J., & Yang, M.-H. (2021). Learning spatially variant linear representation models for joint filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(11), 8355–8370.

Dong, X., Yokoya, N., Wang, L., & Uezato, T. (2022). Learning mutual modulation for self-supervised cross-modal super-resolution. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX Computer vision–eccv 2022: 17th european conference, tel aviv, israel, october 23–27, 2022, proceedings, part xix* (pp. 1–18).

Eigen, D.,& Fergus, R. (2015). Proceedings of the ieee international conference on computer vision (pp. 2650–2658).

Eldesokey, A., Felsberg, M., Holmquist, K., & Persson, M. (2020). Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12014–12023).

Eldesokey, A., Felsberg, M., & Khan, F.S. (2018). Propagating confidences through cnns for sparse data regression. arXiv preprint arXiv:1805.11913,

Eldesokey, A., Felsberg, M., & Khan, F. S. (2019). Confidence propagation through cnns for guided sparse depth regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(10), 2423–2436.

Fan, R., Li, Z., Poggi, M.,& Mattoccia, S. (2022). A cascade dense connection fusion network for depth completion. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. 33rd british machine vision conference 2022, BMVC 2022, london, uk, november 21-24, 2022*.

Feng, R., Li, C., Chen, H., Li, S., Gu, J., & Loy, C.C. (2023). Generating aligned pseudo-supervision from non-aligned data for image restoration in under-display camera. arXiv preprint arXiv:2304.06019,

Feng, R., Li, C., Chen, H., Li, S., Loy, C.C., & Gu, J. (2021). Removing diffraction image artifacts in under-display camera via dynamic skip connection network. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 662–671).

Ferstl, D., Reinbacher, C., Ranftl, R., Rüther, M., & Bischof, H. (2013). Image guided depth upsampling using anisotropic total generalized variation. In *Proceedings of the ieee international conference on computer vision* (pp. 993–1000).

Figueiredo, M. (2001). Adaptive sparseness using jeffreys prior. *Advances in Neural Information Processing Systems, 14*, 722.

Gaidon, A., Wang, Q., Cabon, Y., & Vig, E. (2016). Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4340–4349).

Ge, C., Qiao, X., Huimin, Y., Zhou, Y., & Deng, P. (2021). Tof depth sensor based on laser speckle projection and distance measuring method thereof. Google Patents. US Patent 11,143,880.

Godard, C., Mac Aodha, O., & Brostow, G.J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 270–279).

Godard, C., Mac Aodha, O., Firman, M., & Brostow, G.J. (2019). Digging into self-supervised monocular depth estimation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 3828–3838).

Gu, J., Xiang, Z., Ye, Y., & Wang, L. (2021). Denselidar: A real-time pseudo dense depth guided depth completion network. *IEEE Robotics and Automation Letters, 6*(2), 1808–1815.

Gu, S., Guo, S., Zuo, W., Chen, Y., Timofte, R., Van Gool, L., & Zhang, L. (2019). Learned dynamic guidance for depth image recon-

struction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*, 102437–2452.

Guizilini, V., Ambrus, R., Burgard, W., & Gaidon, A. (2021). Sparse auxiliary networks for unified monocular depth prediction and completion. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11078–11088).

Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., & Gaidon, A. (2020). 3d packing for self-supervised monocular depth estimation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2485–2494).

Guizilini, V., Li, J., Ambrus, R., Pillai, S., & Gaidon, A. (2020). Robust semi-supervised monocular depth estimation with reprojected distances. In *Conference on robot learning* (pp. 503–512).

Guo, C., Li, C., Guo, J., Cong, R., Fu, H., & Han, P. (2019). Hierarchical features driven residual learning for depth map super-resolution. *IEEE Transactions on Image Processing, 28*(5), 2545–2557. https://doi.org/10.1109/TIP.2018.2887029

Gupta, M., Nayar, S. K., Hullin, M. B., & Martin, J. (2015). Phasor imaging: A generalization of correlation-based time-of-flight imaging. *ACM Transactions on Graphics (ToG), 34*(5), 1–18.

Gutierrez-Barragan, F., Chen, H., Gupta, M., Velten, A., & Gu, J. (2021). itof2dtof: A robust and flexible representation for data-driven time-of-flight imaging. *IEEE Transactions on Computational Imaging, 7*, 1205–1214.

Häne, C., Heng, L., Lee, G. H., Fraundorfer, F., Furgale, P., Sattler, T., & Pollefeys, M. (2017). 3d visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *Image and Vision Computing, 68*, 14–27.

He, K., Sun, J., & Tang, X. (2010). Guided image filtering. In *European conference on computer vision* (pp. 1–14).

He, K., Sun, J., & Tang, X. (2012). Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(6), 1397–1409.

He, L., Zhu, H., Li, F., Bai, H., Cong, R., Zhang, C., & Zhao, Y. (2021). Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9229–9238).

Heide, F., Heidrich, W., Hullin, M., & Wetzstein, G. (2015). Doppler time-of-flight imaging. *ACM Transactions on Graphics (ToG), 34*(4), 1–11.

Hirata, A., Ishikawa, R., Roxas, M., & Oishi, T. (2019). Real-time dense depth estimation using semantically-guided lidar data propagation and motion stereo. *IEEE Robotics and Automation Letters, 4*(4), 3806–3811.

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).

Holynski, A., & Kopf, J. (2018). Fast depth densification for occlusion-aware augmented reality. *ACM Transactions on Graphics (ToG), 37*(6), 1–11.

Hu, J., Bao, C., Ozay, M., Fan, C., Gao, Q., Liu, H., & Lam, T. L. (2022). Deep depth completion from extremely sparse data: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2022.3229090

Hu, M., Wang, S., Li, B., Ning, S., Fan, L., & Gong, X. (2021). Penet: Towards precise and efficient image guided depth completion. In *2021 ieee international conference on robotics and automation (icra)* (pp. 13656–13662).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K.Q. (2017). Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4700–4708).

Hui, T.-W., Loy, C.C., & Tang, X. (2016). Depth map super-resolution by deep multi-scale guidance. In *Computer vision–eccv 2016: 14th*

european conference, amsterdam, the netherlands, october 11-14, 2016, proceedings, part iii 14 (pp. 353–369).

Hussmann, S., Knoll, F., & Edeler, T. (2013). Modulation method including noise model for minimizing the wiggling error of tof cameras. *IEEE Transactions on Instrumentation and Measurement, 63*(5), 1127–1136.

Jaritz, M., De Charette, R., Wirbel, E., Perrotton, X., & Nashashibi, F. (2018). Sparse and dense data with cnns: Depth completion and semantic segmentation. In *2018 international conference on 3d vision (3dv)* (pp. 52–60).

Jeon, J., Lim, H., Seo, D.-U., & Myung, H. (2022). Struct-mdc: Mesh-refined unsupervised depth completion leveraging structural regularities from visual slam. *IEEE Robotics and Automation Letters, 7*(3), 6391–6398.

Jeon, Y., Kim, H., & Seo, S.-W. (2021). Abcd: Attentive bilateral convolutional network for robust depth completion. *IEEE Robotics and Automation Letters, 7*(1), 81–87.

Jiang, X., Cambareri, V., Agresti, G., Ugwu, C.I., Simonetto, A., Cardinaux, F., & Zanuttigh, P. (2022). A low memory footprint quantized neural network for depth completion of very sparse time-of-flight depth maps. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2687–2696).

Kalia, M., Navab, N., & Salcudean, T. (2019). A real-time interactive augmented reality depth estimation technique for surgical robotics. In *2019 international conference on robotics and automation (icra)* (pp. 8291–8297).

Kam, J., Kim, J., Kim, S., Park, J., & Lee, S. (2022). Costdcnet: Cost volume based depth completion for a single rgb-d image. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II Computer vision–eccv 2022: 17th european conference, tel aviv, israel, october 23–27, 2022, proceedings, part ii* (pp. 257–274).

Ke, Y., Li, K., Yang, W., Xu, Z., Hao, D., Huang, L., & Wang, G. (2021). Mdanet: Multi-modal deep aggregation network for depth completion. In *2021 ieee international conference on robotics and automation (icra)* (pp. 4288–4294).

Khan, M.F.F., Troncoso Aldas, N.D., Kumar, A., Advani, S., & Narayanan, V. (2021). Sparse to dense depth completion using a generative adversarial network with intelligent sampling strategies. In *Proceedings of the 29th acm international conference on multimedia* (pp. 5528–5536).

Kim, B., Ponce, J., & Ham, B. (2021). Deformable kernel networks for joint image filtering. *International Journal of Computer Vision, 129*(2), 579–600.

Koh, J., Lee, J., & Yoon, S. (2022). Bnudc: A two-branched deep neural network for restoring images from under-display cameras. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 1950–1959).

Kwon, K., Kang, E., Lee, S., Lee, S.-J., Lee, H.-E., Yoo, B., & Han, J.-J. (2021). Controllable image restoration for under-display camera in smartphones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2073–2082).

Lee, B.-U., Jeon, H.-G., Im, S., & Kweon, I.S. (2019). Depth completion with deep geometry and context guidance. In *2019 international conference on robotics and automation (icra)* (pp. 3281–3287).

Lee, S., Lee, J., Kim, D., & Kim, J. (2020). Deep architecture with cross guidance between single image and sparse lidar data for depth completion. *IEEE Access, 8*, 79801–79810.

Lee, S., Yi, E., Lee, J., & Kim, J. (2022). Multi-scaled and densely connected locally convolutional layers for depth completion. In *2022 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 8360–8367).

Li, A., Yuan, Z., Ling, Y., Chi, W., & Zhang, C., et al.(2020). A multi-scale guided cascade hourglass network for depth completion. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 32–40).

Li, D., Xu, J., Yang, Z., Zhang, Q., Ma, Q., Zhang, L., & Chen, P. (2022). Motion inspires notion: self-supervised visual-lidar fusion for environment depth estimation. In *Proceedings of the 20th annual international conference on mobile systems, applications and services* (pp. 114–127).

Li, J., Yue, T., Zhao, S., & Hu, X. (2022). Fisher information guidance for learned time-of-flight imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 16334–16343).

Li, T., Lin, H., Dong, X., & Zhang, X. (2020). Depth image super-resolution using correlation-controlled color guidance and multi-scale symmetric network. *Pattern Recognition, 107*, 107513.

Li, Y., Huang, J.-B., Ahuja, N., & Yang, M.-H. (2016). Deep joint image filtering. In *European conference on computer vision* (pp. 154–169).

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*, 81909–81923.

Li, Y., Liu, X., Dong, W., Zhou, H., Bao, H., Zhang, G., & Cui, Z. (2022). Deltar: Depth estimation from a light-weight tof sensor and rgb image. In *European conference on computer vision* (pp. 619–636).

Liang, M., Yang, B., Chen, Y., Hu, R., & Urtasun, R. (2019). Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 7345–7353).

Liao, Y., Xie, J., & Geiger, A. (2022). Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(3), 3292–3310.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the ieee international conference on computer vision* (pp. 2980–2988).

Lin, Y., Cheng, T., Zhong, Q., Zhou, W., & Yang, H. (2022). Dynamic spatial propagation network for depth completion. In *Proceedings of the aaai conference on artificial intelligence* (Vol 36, pp. 1638–1646).

Lipson, L., Teed, Z., & Deng, J. (2021). Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International conference on 3d vision (3dv)*.

Liu, A., Tan, Z., Wan, J., Escalera, S., Guo, G., & Li, S.Z. (2021). Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 1179–1187).

Liu, J., Gong, X., & Liu, J. (2012). Guided inpainting and filtering for kinect depth maps. In *Proceedings of the 21st international conference on pattern recognition (icpr2012)* (pp. 2055–2058).

Liu, J., & Jung, C. (2022). Nnnet: New normal guided depth completion from sparse lidar data and single color image. *IEEE Access, 10*, 114252–114261.

Liu, L., Liao, Y., Wang, Y., Geiger, A., & Liu, Y. (2021). Learning steering kernels for guided depth completion. *IEEE Transactions on Image Processing, 30*, 2850–2861.

Liu, L., Song, X., Lyu, X., Diao, J., Wang, M., Liu, Y., & Zhang, L. (2021). Fcfr-net: Feature fusion based coarse-to-fine residual learning for depth completion. In *Proceedings of the aaai conference on artificial intelligence* (Vol 35, pp. 2136–2144).

Liu, L., Song, X., Sun, J., Lyu, X., Li, L., Liu, Y., & Zhang, L. (2023). Mff-net: Towards efficient monocular depth completion with multi-modal feature fusion. *IEEE Robotics and Automation Letters, 8*(2), 920–927.

Liu, P., Zhang, Z., Meng, Z., & Gao, N. (2021). Deformable enhancement and adaptive fusion for depth map super-resolution. *IEEE Signal Processing Letters, 29*, 204–208.

Liu, P., Zhang, Z., Meng, Z., Gao, N., & Wang, C. (2022). Pdr-net: Progressive depth reconstruction network for color guided depth map super-resolution. *Neurocomputing, 479*, 75–88.

Liu, R., Zhong, G., Cao, J., Lin, Z., Shan, S., & Luo, Z. (2016). Learning to diffuse: A new perspective to design pdes for visual analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*, 122457–2471.

Liu, S., De Mello, S., Gu, J., Zhong, G., Yang, M.-H., & Kautz, J. (2017). Learning affinity via spatial propagation networks. *Advances in Neural Information Processing Systems, 30*, 1519–1529.

Liu, T.Y., Agrawal, P., Chen, A., Hong, B.-W., & Wong, A. (2022). Monitored distillation for positive congruent depth completion. In *Computer vision–eccv 2022: 17th european conference, tel aviv, israel, october 23–27, 2022, proceedings, part ii* (pp. 35–53).

Liu, X., Li, Y., Teng, Y., Bao, H., Zhang, G., Zhang, Y., & Cui, Z. (2023). Multi-modal neural radiance field for monocular dense slam with a light-weight tof sensor. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 1–11).

Liu, X., Shao, X., Wang, B., Li, Y., & Wang, S. (2022). Graphcspn: Geometry-aware depth completion via dynamic gcns. In *Computer vision–eccv 2022: 17th european conference, tel aviv, israel, october 23–27, 2022, proceedings, part xxxiii* (pp. 90–107).

Lopez-Rodriguez, A., Busam, B., & Mikolajczyk, K. (2020). Project to adapt: Domain adaptation for depth completion from noisy and sparse sensor data. In *Proceedings of the asian conference on computer vision*.

Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. *Advances in Neural Information Processing Systems* **29**

Lu, K., Barnes, N., Anwar, S., & Zheng, L. (2020). From depth what can you see? depth completion via auxiliary image reconstruction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11306–11315).

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).

Lutio, R.d., D'aronco, S., Wegner, J.D., & Schindler, K. (2019). Guided super-resolution as pixel-to-pixel transformation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 8829–8837).

Ma, F., Cavalheiro, G.V., & Karaman, S. (2018). Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *2019 International Conference on Robotics and Automation (ICRA)* (pp. 3288-3295).

Ma, F., Cavalheiro, G.V., & Karaman, S. (2019). Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *2019 international conference on robotics and automation (icra)* (pp. 3288–3295).

Ma, F., & Karaman, S. (2018). Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 ieee international conference on robotics and automation (icra)* (pp. 4796–4803).

Manglik, A., Weng, X., Ohn-Bar, E., & Kitanil, K.M. (2019). Forecasting time-to-collision from monocular video: Feasibility, dataset, and challenges. In *2019 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 8081–8088).

Marivani, I., Tsiligianni, E., Cornelis, B., & Deligiannis, N. (2020). Multimodal deep unfolding for guided image super-resolution. *IEEE Transactions on Image Processing, 29*, 8443–8456.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., & Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4040–4048).

Metzger, N., Daudt, R.C., & Schindler, K. (2022). Guided depth super-resolution by deep anisotropic diffusion. arXiv preprint arXiv:2211.11592,

Nazir, D., Pagani, A., Liwicki, M., Stricker, D., & Afzal, M. Z. (2022). Semattnet: Toward attention-based semantic aware guided depth completion. *IEEE Access, 10*, 120781–120791.

Nguyen, T., & Yoo, M. (2022). Patchgan-based depth completion in autonomous vehicle. In *2022 international conference on information networking (icoin)* (pp. 498–501).

Park, J., Joo, K., Hu, Z., Liu, C.-K., & So Kweon, I. (2020). Non-local spatial propagation network for depth completion. In *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part xiii 16* (pp. 120–136).

Patil, P. W., Dudhane, A., Chaudhary, S., & Murala, S. (2022). Multiframe based adversarial learning approach for video surveillance. *Pattern Recognition, 122*, 108350.

Peng, R., Zhang, T., Li, B., & Wang, Y. (2022). Pixelwise adaptive discretization with uncertainty sampling for depth completion. In *Proceedings of the 30th acm international conference on multimedia* (pp. 3926–3935).

Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., & Toyama, K. (2004). Digital photography with flash and no-flash image pairs. *ACM Transactions on Graphics (TOG), 23*(3), 664–672.

Qi, C.R., Su, H., Mo, K., & Guibas, L.J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 652–660).

Qiao, X., Ge, C., Deng, P., Wei, H., Poggi, M., & Mattoccia, S. (2023). Depth restoration in under-display time-of-flight imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(5), 5668–5683. https://doi.org/10.1109/TPAMI.2022.3209905

Qiao, X., Ge, C., Yao, H., Deng, P., & Zhou, Y. (2020). Valid depth data extraction and correction for time-of-flight camera. In *Twelfth international conference on machine vision (icmv 2019)* (Vol 11433, pp. 696–703).

Qiao, X., Ge, C., Zhang, Y., Zhou, Y., Tosi, F., Poggi, M., & Mattoccia, S. (2023). Depth super-resolution from explicit and implicit high-frequency features. *Computer Vision and Image Understanding, 237*, 103841.

Qiao, X., Ge, C., Zhao, C., Tosi, F., Poggi, M., & Mattoccia, S. (2023). Self-supervised depth super-resolution with contrastive multiview pre-training. *Neural Networks, 168*, 223–236.

Qiu, D., Pang, J., Sun, W., & Yang, C. (2019). Deep end-to-end alignment and refinement for time-of-flight rgb-d module. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9994–10003).

Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B., & Pollefeys, M. (2019). Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3313–3322).

Qu, C., Nguyen, T., & Taylor, C. (2020). Depth completion via deep basis fitting. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 71–80).

Ramesh, A.N., Giovanneschi, F., & González-Huici, M.A. (2023). Siunet: Sparsity invariant u-net for edge-aware depth completion. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 5818–5827).

Riegler, G., Ferstl, D., Rüther, M., & Horst, B. (2016). A deep primal-dual network for guided depth super-resolution. In *British machine vision conference*.

Riegler, G., Rüther, M., & Bischof, H. (2016). Atgv-net: Accurate depth super-resolution. In *Computer vision–eccv 2016: 14th european conference, amsterdam, the netherlands, october 11-14, 2016, proceedings, part iii* 14 (pp. 268–284).

Romera, E., Alvarez, J. M., Bergasa, L. M., & Arroyo, R. (2017). Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems, 19*(1), 263–272.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–miccai 2015: 18th international conference, munich, germany, october 5-9, 2015, proceedings, part iii 18* (pp. 234–241).

Ryu, K., Lee, K.-I., Cho, J., & Yoon, K.-J. (2021). Scanline resolution-invariant depth completion using a single image and sparse lidar point cloud. *IEEE Robotics and Automation Letters, 6*(4), 6961–6968.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., & Westling, P. (2014). High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern recognition: 36th german conference, gcpr 2014, münster, germany, september 2-5, 2014, proceedings 36* (pp. 31–42).

Scharstein, D., & Pal, C. (2007). Learning conditional random fields for stereo. In *2007 ieee conference on computer vision and pattern recognition* (pp. 1–8).

Scharstein, D., & Szeliski, R. (2003). High-accuracy stereo depth maps using structured light. In *2003 ieee computer society conference on computer vision and pattern recognition, 2003. proceedings.* (Vol 1, pp. I–I).

Schuster, R., Wasenmuller, O., Unger, C., & Stricker, D. (2021). Ssgp: Sparse spatial guided propagation for robust and generic interpolation. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 197–206).

Shacht, G., Danon, D., Fogel, S., & Cohen-Or, D. (2021). Single pair cross-modality super resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 6378–6387).

Shin, J., Shin, S., & Jeon, H.-G. (2023). Task-specific scene structure representations. arXiv preprint arXiv:2301.00555,

Shivakumar, S.S., Nguyen, T., Miller, I.D., Chen, S.W., Kumar, V., & Taylor, C.J. (2019). Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion. In *2019 ieee intelligent transportation systems conference (itsc)* (pp. 13–20).

Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *European conference on computer vision* (pp. 746–760).

Song, P., Deng, X., Mota, J. F., Deligiannis, N., Dragotti, P. L., & Rodrigues, M. R. (2019). Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries. *IEEE Transactions on Computational Imaging, 6*, 57–72.

Song, S., Lichtenberg, S.P., & Xiao, J. (2015). Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 567–576).

Song, X., Dai, Y., Zhou, D., Liu, L., Li, W., Li, H., & Yang, R. (2020). Channel attention based iterative residual learning for depth map super-resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5631–5640).

Song, Z., Lu, J., Yao, Y., & Zhang, J. (2021). Self-supervised depth completion from direct visual-lidar odometry in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems, 23*(8), 11654–11665.

Su, H., Jampani, V., Sun, D., Gallo, O., Learned-Miller, E., & Kautz, J. (2019). Pixel-adaptive convolutional neural networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11166–11175).

Su, S., Heide, F., Wetzstein, G., & Heidrich, W. (2018). Deep end-to-end time-of-flight imaging. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6383–6392).

Sun, B., Ye, X., Li, B., Li, H., Wang, Z., & Xu, R. (2021). Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 7792–7801).

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., & ...others (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2446–2454).

Sun, Z., Ye, W., Xiong, J., Choe, G., Wang, J., Su, S., & Ranjan, R. (2023). Consistent direct time-of-flight video depth super-resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5075–5085).

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114).

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).

Tang, J., Tian, F.-P., Feng, W., Li, J., & Tan, P. (2020). Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing, 30*, 1116–1129.

Tang, Q., Cong, R., Sheng, R., He, L., Zhang, D., Zhao, Y., & Kwong, S. (2021). Bridgenet: A joint learning network of depth map super-resolution and monocular depth estimation. In *Proceedings of the 29th acm international conference on multimedia* (pp. 2148–2157).

Tomasi, C., & Manduchi, R. (1998). Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271) Sixth international conference on computer vision (ieee cat. no. 98ch36271)* (pp. 839–846).

Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., & Geiger, A. (2017). Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV) 2017 international conference on 3d vision (3dv)* (pp. 11–20).

Van Gansbeke, W., & Neven, D., De Brabandere, B., & Van Gool, L. (2019). Sparse and noisy lidar completion with rgb guidance and uncertainty. In *2019 16th international conference on machine vision applications (mva)* (pp. 1–6).

Voynov, O., Artemov, A., Egiazarian, V., Notchenko, A., Bobrovskikh, G., Burnaev, E., & Zorin, D. (2019). Perceptual deep depth super-resolution. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 5653–5663).

Wan, Q., Huang, Z., Lu, J., Yu, G., & Zhang, L. (2023). Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation. In *International Conference on Learning Representations (ICLR). International conference on learning representations (iclr).*

Wang, H., Wang, M., Che, Z., Xu, Z., Qiao, X., Qi, M., & Tang, J. (2022). Rgb-depth fusion gan for indoor depth completion. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 6209–6218).

Wang, J., Liu, P., & Wen, F. (2022). Self-supervised learning for rgb-guided depth enhancement by exploiting the dependency between rgb and depth. *IEEE Transactions on Image Processing, 32*, 159–174.

Wang, J., Liu, P., & Wen, F. (2023). Self-supervised learning for rgb-guided depth enhancement by exploiting the dependency between rgb and depth. *IEEE Transactions on Image Processing*. https://doi.org/10.1109/TIP.2022.3226419

Wang, K., Zhao, L., Zhang, J., Zhang, J., Wang, A., & Bai, H. (2023). Joint depth map super-resolution method via deep hybrid-cross guidance filter. *Pattern Recognition, 136*, 109260.

Wang, X. , Chen, X. , Ni, B. , Tong, Z. , & Wang, H. (2022). Learning continuous depth representation via geometric spatial aggregator. arXiv preprint arXiv:2212.03499,

Wang, Y. , Chao, W.-L. , Garg, D. , Hariharan, B. , Campbell, M. , & Weinberger, K.Q. (2019). Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8445–8453).

Wang, Y. , Li, B. , Zhang, G. , Liu, Q. , Tao, G. , & Dai, Y. (2023). Lrru: Long-short range recurrent updating networks for depth completion. In *Proceedings of the ieee international conference on computer vision (iccv)* (pp. 9422–9432).

Wang, Y., Yang, J., & Yue, H. (2023). Depth map continuous super-resolution with local implicit guidance function. *Displays, 78*, 102418.

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.

Weickert, J. & ,et al. (1998). Anisotropic diffusion in image processing (Vol 1). Teubner Stuttgart.

Wen, Y., Sheng, B., Li, P., Lin, W., & Feng, D. D. (2019). Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution. *IEEE Transactions on Image Processing, 28*(2), 994–1006. https://doi.org/10.1109/TIP.2018.2874285

Wong, A., Cicek, S., & Soatto, S. (2021). Learning topology from synthetic data for unsupervised depth completion. *IEEE Robotics and Automation Letters, 6*(2), 1495–1502.

Wong, A., Fei, X., Tsuei, S., & Soatto, S. (2020). Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters, 5*(2), 1899–1906.

Wong, A. ,& Soatto, S. (2021). Unsupervised depth completion with calibrated backprojection layers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 12747–12756).

Wronski, B., Garcia-Dorado, I., Ernst, M., Kelly, D., Krainin, M., Liang, C.-K., & Milanfar, P. (2019). Handheld multi-frame super-resolution. *ACM Transactions on Graphics (ToG), 38*(4), 1–18.

Wu, X. , Peng, L. , Yang, H. , Xie, L. , Huang, C. , Deng, C. , & Cai, D. (2022). Sparse fuse dense: Towards high quality 3d detection with depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5418–5427).

Xia, F. , Zamir, A.R. , He, Z. , Sax, A. , Malik, J. , & Savarese, S. (2018). Gibson env: Real-world perception for embodied agents. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 9068–9079).

Xie, Z., Yu, X., Gao, X., Li, K., & Shen, S. (2022). Recent advances in conventional and deep learning-based depth completion: A survey. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2022.3201534

Xu, J. , Zhu, Y. , Wang, W. , & Liu, G. (2023). A real-time semi-dense depth-guided depth completion network. *The Visual Computer* pp. 1–11,

Xu, L. , Guan, T. , Wang, Y. , Luo, Y. , Chen, Z. , Liu, W. , & Yang, W. (2022). Self-supervised multi-view stereo via adjacent geometry guided volume completion. In *Proceedings of the 30th acm international conference on multimedia* (pp. 2202–2210).

Xu, Y. , Zhu, X. , Shi, J. , Zhang, G. , Bao, H. , & Li, H. (2019). Depth completion from sparse lidar data with depth-normal constraints. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 2811–2820).

Xu, Z. , Yin, H. , & Yao, J. (2020). Deformable spatial propagation networks for depth completion. In *2020 ieee international conference on image processing (icip)* (pp. 913–917).

Yan, L., Liu, K., & Belyaev, E. (2020). Revisiting sparsity invariant convolution: A network for image guided depth completion. *IEEE Access, 8*, 126323–126332.

Yan, Z. , Wang, K. , Li, X. , Zhang, Z. , Li, J. , & Yang, J. (2022). Rignet: Repetitive image guided network for depth completion. In *Computer vision–eccv 2022: 17th european conference, tel aviv, israel, october 23–27, 2022, proceedings, part xxvii* (pp. 214–230).

Yang, A., & Sankaranarayanan, A. C. (2021). Designing display pixel layouts for under-panel cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(7), 2245–2256.

Yang, Y. , Wong, A. , & Soatto, S. (2019). Dense depth posterior (ddp) from single image and sparse range. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3353–3362).

Ye, X., Sun, B., Wang, Z., Yang, J., Xu, R., Li, H., & Li, B. (2020). Pmbanet: Progressive multi-branch aggregation network for scene depth super-resolution. *IEEE Transactions on Image Processing, 29*, 7427–7442.

Yu, Q. , Chu, L. , Wu, Q. , & Pei, L. (2021). Grayscale and normal guided depth completion with a low-cost lidar. In *2021 ieee international conference on image processing (icip)* (pp. 979–983).

Yu, Z. , Sheng, Z. , Zhou, Z. , Luo, L. , Cao, S.-Y. , Gu, H. , & Shen, H.-L. (2023). Aggregating feature point cloud for depth completion. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 8732–8743).

Yuan, J. , Jiang, H. , Li, X. , Qian, J. , Li, J. , & Yang, J. (2023a). Recurrent structure attention guidance for depth super-resolution. arXiv preprint arXiv:2301.13419,

Yuan, J. , Jiang, H. , Li, X. , Qian, J. , Li, J. , & Yang, J. (2023b). Structure flow-guided network for real depth super-resolution. arXiv preprint arXiv:2301.13416,

Zhang, C., Tang, Y., Zhao, C., Sun, Q., Ye, Z., & Kurths, J. (2021). Multitask gans for semantic segmentation and depth completion with cycle consistency. *IEEE Transactions on Neural Networks and Learning Systems, 32*(12), 5404–5415.

Zhang, Q., Chen, X., Wang, X., Han, J., Zhang, Y., & Yue, J. (2022). Self-supervised depth completion based on multi-modal spatio-temporal consistency. *Remote Sensing, 15*(1), 135.

Zhang, Y. ,& Funkhouser, T. (2018). Deep depth completion of a single rgb-d image. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 175–185).

Zhang, Y. , Guo, X. , Poggi, M. , Zhu, Z. , Huang, G. , & Mattoccia, S. (2023). Completionformer: Depth completion with convolutions and vision transformers. Cvpr.

Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering, 34*(12), 5586–5609.

Zhao, C., Yen, G. G., Sun, Q., Zhang, C., & Tang, Y. (2020). Masked gan for unsupervised depth and pose prediction with scale consistency. *IEEE Transactions on Neural Networks and Learning Systems, 32*(12), 5392–5403.

Zhao, S., Gong, M., Fu, H., & Tao, D. (2021). Adaptive context-aware multi-modal network for depth completion. *IEEE Transactions on Image Processing, 30*, 5264–5276.

Zhao, Z. , Zhang, J. , Gu, X. , Tan, C. , Xu, S. , Zhang, Y. , & Van Gool, L. (2023). Spherical space feature decomposition for guided depth map super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 12547-12558).

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.

Zhong, Z. , Liu, X. , Jiang, J. , Zhao, D. , & Ji, X. (2023a). Deep attentional guided image filtering. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhong, Z. , Liu, X. , Jiang, J. , Zhao, D. , & Ji, X. (2023b). Guided depth map super-resolution: A survey. *ACM Computing Surveys*.

Zhou, M., Yan, K., Pan, J., Ren, W., Xie, Q., & Cao, X. (2023). Memory-augmented deep unfolding network for guided image super-resolution. *International Journal of Computer Vision, 131*(1), 215–242.

Zhou, T. , Brown, M. , Snavely, N. , & Lowe, D.G. (2017). Unsupervised learning of depth and ego-motion from video. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1851–1858).

Zhou, Y. , Kwan, M. , Tolentino, K. , Emerton, N. , Lim, S. , & Large, T. , ...others (2020). Udc 2020 challenge on image restoration of under-display camera: Methods and results. In *Computer vision–eccv 2020 workshops: Glasgow, uk, august 23–28, 2020, proceedings, part v 16* (pp. 337–351).

Zhou, Y. , Ren, D. , Emerton, N. , Lim, S. , & Large, T. (2021). Image restoration for under-display camera. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9179–9188).

Zhu, X. , Hu, H. , Lin, S. , & Dai, J. (2019). Deformable convnets v2: More deformable, better results. Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 9308–9316).

Zhu, Y. , Dong, W. , Li, L. , Wu, J. , Li, X. , & Shi, G. (2022). Robust depth completion with uncertainty-driven loss functions. In *Proceedings of the aaai conference on artificial intelligence* (Vol 36, pp. 3626–3634).

Zoph, B. , Vasudevan, V. , Shlens, J. , & Le, Q.V. (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 8697–8710).

Zou, N., Xiang, Z., Chen, Y., Chen, S., & Qiao, C. (2020). Simultaneous semantic segmentation and depth completion with constraint of boundary. *Sensors, 20*(3), 635.

Zuo, Y., Fang, Y., An, P., Shang, X., & Yang, J. (2020). Frequency-dependent depth map enhancement via iterative depth-guided affine transformation and intensity-guided refinement. *IEEE Transactions on Multimedia, 23*, 772–783.

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*(12), 4676–4687.

Zuo, Y., Wang, H., Fang, Y., Huang, X., Shang, X., & Wu, Q. (2021). Mig-net: Multi-scale network alternatively guided by intensity and gradient features for depth map super-resolution. *IEEE Transactions on Multimedia, 24*, 3506–3519.

Zuo, Y., Wu, Q., Fang, Y., An, P., Huang, L., & Chen, Z. (2019). 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), 297–306.