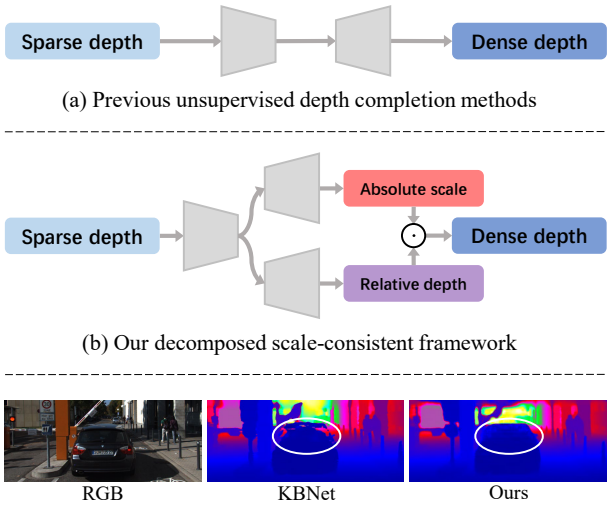
DesNet: Decomposed Scale-Consistent Network for Unsupervised Depth Completion

Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li*, Jian Yang*

PCA Lab, Nanjing University of Science and Technology, China
{Yanzq,kunwang,xiang.li.implus,junli,csjyang}@njust.edu.cn, zhangjesse@foxmail.com

Abstract

Unsupervised depth completion aims to recover dense depth from the sparse one without using the ground-truth annotation. Although depth measurement obtained from LiDAR is usually sparse, it contains valid and real distance information, *i.e.*, scale-consistent absolute depth values. Meanwhile, scale-agnostic counterparts seek to estimate relative depth and have achieved impressive performance. To leverage both the inherent characteristics, we thus suggest to model scale-consistent depth upon unsupervised scale-agnostic frameworks. Specifically, we propose the *decomposed scale-consistent learning* (DSCL) strategy, which disintegrates the absolute depth into relative depth prediction and global scale estimation, contributing to individual learning benefits. But unfortunately, most existing unsupervised scale-agnostic frameworks heavily suffer from depth holes due to the extremely sparse depth input and weak supervisory signal. To tackle this issue, we introduce the *global depth guidance* (GDG) module, which attentively propagates dense depth reference into the sparse target via novel dense-to-sparse attention. Extensive experiments show the superiority of our method on outdoor KITTI benchmark, ranking 1st and outperforming the best KBNet more than 12% in RMSE. In addition, our approach achieves state-of-the-art performance on indoor NYUv2 dataset.



(a) Previous unsupervised depth completion methods

(b) Our decomposed scale-consistent framework

(c) Visual comparison of depth results

Figure 1: (a) Previous methods directly predict the absolute dense depth, whilst (b) our approach decomposes it into relative depth prediction and absolute scale estimation, contributing to (c) not only higher accuracy but denser depth than the excellent KBNet (Wong and Soatto 2021).

Introduction

Depth completion, converting sparse depth to the dense one with or without the help of the corresponding image, is an indispensable part of many computer vision applications, *e.g.*, autonomous driving (Godard et al. 2019; Yan et al. 2022c), augmented reality (Zhong et al. 2016; Yan et al. 2022b), and 3D scene reconstruction (Zhang et al. 2019; Yan et al. 2022a). In these scenarios, dense ground-truth depth annotations are usually expensive and hard to obtain for supervised depth completion while sparse depth maps can be easily measured by depth sensors (*e.g.*, LiDAR). Hence, plenty of unsupervised approaches (Yang, Wong, and Soatto 2020; Wong and Soatto 2021) have been proposed to reduce the high cost since S2D (Ma, Cavalheiro, and Karaman 2019) establishes the first unsupervised framework for depth completion. In general, *for one thing*, these methods are supervised by the sparse depth input and photometric reconstruction, which directly output scale-consistent absolute depth upon the real scale information in sparse depth. *For another thing*, unsupervised depth estimation counterparts (only with color images as input) often suffer from scale ambiguity issues (Bian et al. 2019), which could only predict relative depth. However, in recent months, the counterparts have shown promising prospect that contributes to high depth accuracy (Petrovai and Nedevschi 2022; Mu et al. 2022). *These analyses motivate us to leverage the scale information in sparse depth and the high accuracy of scale-agnostic counterparts for unsupervised depth completion.*

Consequently, in this paper we attempt to explore a new solution to the unsupervised depth completion task, *i.e.*, the *decomposed scale-consistent learning* (DSCL) strategy. As illustrated in Fig. 1(a) and (b), totally different from previous approaches that directly estimates scale-consistent depth, our DSCL first learns scale-agnostic relative depth & real

*Corresponding authors

global scale factor and then outputs the absolute target. Theoretically, we prove that such individual learning is more effective and conduces to better depth results. However, as shown in Fig. 1(c), since the depth input is extremely sparse (about 5% valid pixels) and the supervised signal is very weak, the mainstream unsupervised depth completion methods, *e.g.*, the excellent Oral KBNet (Wong and Soatto 2021) in *ICCV*, heavily suffers from depth holes which thus lead to large inaccuracy as well as unreasonable visual effect.

To tackle this problem, we present the *global depth guidance* (GDG) module. It first produces coarse but dense depth reference by morphological dilation technology (Jackway and Deriche 1996). Then a novel dense-to-sparse attention is designed to effectively propagate the dense depth reference into the sparse target. Concretely, this attention urges to learn the residual of sparse target by capturing non-local correlations between the sparse-modal and dense-modal features, contributing to satisfactory compensation for depth holes. In addition, a fast version of the dense-to-sparse attention is further proposed to realize high efficiency, which largely reduces the complexity from quadratic to linear.

In summary, our main contributions are listed as follows:

- We introduce a new solution to the unsupervised depth completion task, *i.e.*, the decomposed scale-consistent learning framework that disintegrates the absolute depth into relative depth prediction and global scale estimation.
- A global depth guidance module is proposed to deal with the issue of depth holes, including a dense-to-sparse attention that learns long-range correlations between sparse-modal and dense-modal features.
- Extensive experiments verify the effectiveness of our approach, which achieves the state-of-the-art performance on both outdoor KITTI and indoor NYUv2 benchmarks.

Related Work

Depth Completion. The basic task of depth completion has attracted much public attention since the work (Uhrig et al. 2017) first proposes sparsity invariant CNNs to fill missing depth values. In general, depth completion can be broadly categorized into supervised and unsupervised learning. *For supervised learning*, existing methods mainly take as input a single sparse depth or multiple sensor information (Yan et al. 2022c), which has greatly promoted the development of the depth completion task. For example, (Ma, Cavalheiro, and Karaman 2019) utilize an hourglass network to recover dense depth based on a single sparse depth. (Lu et al. 2020) employ sparse depth as the only input and further use the corresponding color image as an auxiliary supervisory signal to provide semantic information. CSPN (Cheng, Wang, and Yang 2018), NLSPN (Park et al. 2020), and DySPN (Lin et al. 2022) refine coarse depth by learning affinity matrix with spatial propagation network based on RGB-D pair. GuideNet (Tang et al. 2020) and RigNet (Yan et al. 2022c) present image-guided methods to benefit depth completion. DeepLiDAR (Qiu et al. 2019) jointly uses color image, surface normal, and sparse depth for more precise depth recovery. To robustly predict dense depth, uncertainty estimation (Van Gansbeke et al. 2019; Zhu et al. 2022) is introduced to tackle outlier and obscure. *For unsupervised learning*, there are lots of works (Yang, Wong, and Soatto 2020; Wong and Soatto 2021) focusing on simultaneously completing sparse depth and reducing expensive ground-truth costs. For example, (Ma, Cavalheiro, and Karaman 2019) build a solid framework to concurrently deal with supervised, self-supervised, and unsupervised depth completion. (Wong, Cicek, and Soatto 2021) utilize synthetic data for further improvement. Recently, KBNet (Wong and Soatto 2021) proposes calibrated backprojection that significantly facilitates the unsupervised depth completion task. However, most of these methods directly predict absolute depth and often suffer from depth holes near key elements for self-driving, *e.g.*, cars. Different from them, we present a new unsupervised solution that decomposes the absolute depth into relative depth prediction and global scale estimation.

Depth Estimation. The tasks of depth completion and depth estimation are closely relevant. The major difference between them is that the former has additional sparse depth information as input while the latter does not. Research on depth estimation can trace further back to the early method (Saxena et al. 2005). Since then, many *supervised approaches* (Roy and Todorovic 2016; Lee et al. 2019; Zhang et al. 2019) have been proposed which greatly promote the development of this domain. Furthermore, (Zhou et al. 2017) propose the first *unsupervised depth estimation* system, which takes view synthesis as the supervisory signal during end-to-end training. This work has laid a good foundation for subsequent research. After that, various unsupervised works (Godard et al. 2019; Bian et al. 2019; Shu et al. 2020; Zhao et al. 2020; Mu et al. 2022) are burgeoning over the past three years. Although these methods suffer from the scale ambiguity issues all along, they have achieved promising performance with high depth accuracy. Therefore, it becomes possible to explore a new solution to unsupervised depth completion task based on these scale-agnostic frameworks and real scale information in sparse depth input.

Scale Decomposition in Depth. There are some depth estimation works related to scale decomposition, including scale-agnostic and scale-consistent categories. *For scale-agnostic methods*, (Eigen, Puhrsch, and Fergus 2014) present to learn relative depth that is normalized to (0, 1) to tackle the ambiguous scale issue. Meanwhile, (Xian et al. 2020) propose pair-wise ranking loss guided by structure to improve the quality of depth prediction. Further, (Ranftl et al. 2020) ameliorates the generalization capability on multiple datasets with different scales. (Wang et al. 2020) build a new framework by depth and scale decomposition in a supervised manner. *For scale-consistent approaches*, many works (Chen, Schmid, and Sminchisescu 2019; Wang et al. 2021) utilize geometric consistency in 2D and 3D spaces to model consistent scales. In addition, (Guizilini et al. 2020) takes camera velocity as extra supervised signal to mitigate the scale-agnostic issue. Moreover, (Tiwari et al. 2020) introduce bundle-adjusted 3D scene structures to benefit their depth prediction network. Different from them, to seek a new solution to the unsupervised depth completion task, we predict scale-consistent depth via scale-agnostic basis with the help of real scale information in sparse depth input.

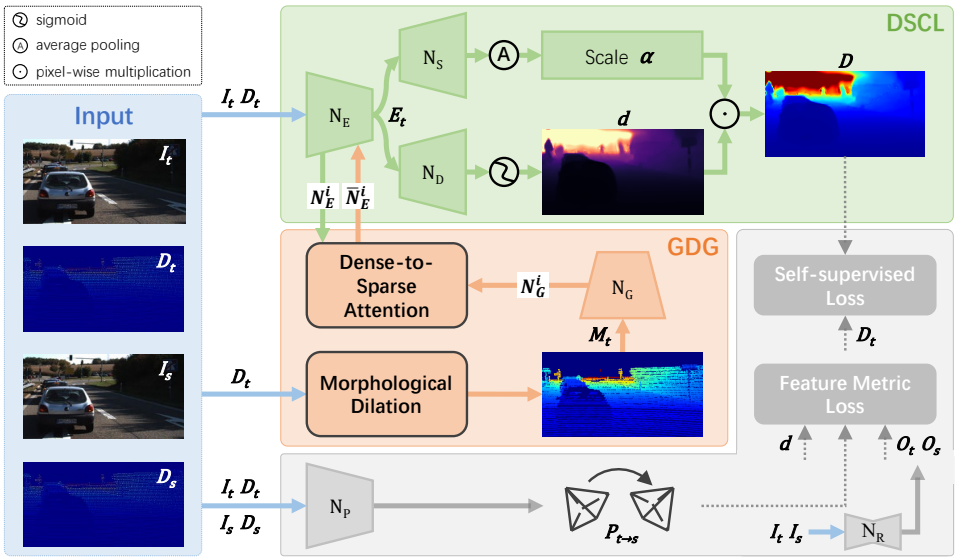


Figure 2: Overview of our unsupervised depth completion framework, where the *decomposed scale-consistent learning* (DSCL) is designed to disentangle the absolute depth into relative depth prediction and absolute scale estimation. Meanwhile, the *global depth guidance* (GDG) is introduced to provide the depth network in DSCL with dense depth reference.

Decomposed Scale-Consistent Network

In this section, we first introduce some prior knowledge in Sec. and the overall network architecture in Sec. . Then we elaborate on the two key designs of our method in Secs. and . For simplicity, the proposed **De**composed **S**cale-Consistent **Net**work is termed as **DesNet**.

Prior Knowledge

Camera model. The optical camera projects a 3D point $Q = (X, Y, Z)$ to a 2D pixel $q = (u, v)$ by

$$\pi\left(Q\right) = \left(h_x \frac{X}{Z} + c_x, h_y \frac{Y}{Z} + c_y\right), \tag{1}$$

where π represents camera operator, (h_x, h_y, c_x, c_y) are the optical camera intrinsic parameters. Given depth $\boldsymbol{d}$ and its pixel d_q, the corresponding backprojection process is

$$\pi^{-1}\left(q, d_q\right) = d_q \left(\frac{x - c_x}{h_x}, \frac{y - c_y}{h_y}, 1\right)^T. \tag{2}$$

Ego-motion. Ego-motion can be modeled by transformation G. Warping function ω maps a pixel q in one frame to another frame, obtaining the corresponding pixel $\widetilde{q}$. It can be described as

$$\widetilde{q} = \omega\left(q, d_q, G\right) = \pi\left(G \cdot \pi^{-1}\left(q, d_q\right)\right). \tag{3}$$

Overview of Network Architecture

The whole framework of our method is shown in Fig. 2. Without loss of generality, we define monocular RGB-D target frames $\boldsymbol{I}_t$ (color image), $\boldsymbol{D}_t$ (sparse depth), and source frames $\boldsymbol{I}_s$, $\boldsymbol{D}_s$ as input. $\boldsymbol{O}_t$ and $\boldsymbol{O}_s$ are the color space features of $\boldsymbol{I}_t$ and $\boldsymbol{I}_s$, which are encoded by a shared image reconstruction network N_R (Shu et al. 2020).

For decomposed scale-consistent learning (DSCL), we first predict the relative depth $\boldsymbol{d}$ by the depth network N_D with sigmoid mapping. Concurrently, the global scale factor α is estimated by the scale network N_S with average pooling based on $\boldsymbol{E}_t$ produced from the shared encoder N_E, where a dense-to-sparse attention is proposed to propagate dense depth reference. Finally, we multiply $\boldsymbol{d}$ by α to generate the scale-consistent absolute depth prediction $\boldsymbol{D}$.

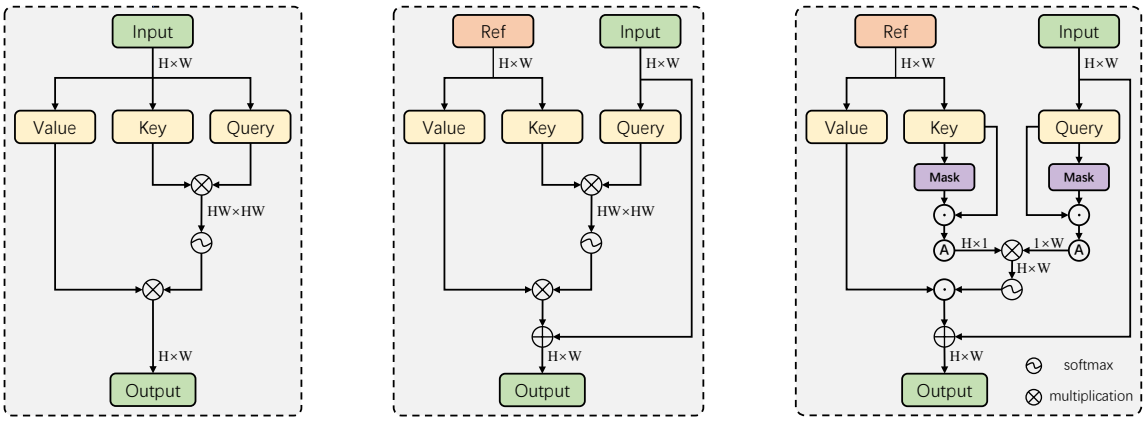
For global depth guidance (GDG), we first transform $\boldsymbol{D}_t$ to a denser depth $\boldsymbol{M}_t$ by morphological dilation technology. Then we employ the guidance network N_G to map $\boldsymbol{M}_t$ into feature space. The features in ith layer of N_G and N_E are $\boldsymbol{N}_G^i$ and $\boldsymbol{N}_E^i$ respectively, both of which are input into the dense-to-sparse attention module to update $\boldsymbol{N}_E^i$ to $\bar{\boldsymbol{N}}_E^i$.

For supervised signal, we first use $\boldsymbol{D}_t$ as the primary supervision of the absolute depth prediction $\boldsymbol{D}$. Then following (Shu et al. 2020), we employ the feature metric loss as the auxiliary supervision, which inputs $\boldsymbol{O}_t$, $\boldsymbol{O}_s$, and the pose $\boldsymbol{P}_{t\to s}$ that is predicted by the pose network N_P. Next, we warp $\boldsymbol{O}_t$ to $\boldsymbol{O}_{t\to s}$ with $\boldsymbol{d}$ and $\boldsymbol{P}_{t\to s}$ used. A cross-view reconstruction loss is thus applied between $\boldsymbol{O}_s$ and $\boldsymbol{O}_{t\to s}$.

It is worth noting that, when testing, our model only needs $\boldsymbol{I}_t$ and $\boldsymbol{D}_t$ to generate the final depth prediction $\boldsymbol{D}$.

Decomposed Scale-consistent Learning

Sparse depth maps possess actual and precise depth values, which contain real scale information (Uhrig et al. 2017) that can provide significant guidance for the unsupervised setting. On the other hand, existing scale-agnostic unsupervised depth estimation methods (Godard et al. 2019) have achieved impressive performance, especially the high accuracy. Accordingly, we attempt to leverage the real scale information in sparse depth and the high accuracy of scale-agnostic counterparts for the unsupervised depth completion



(a) Conventional attention (b) Our dense-to-sparse attention (c) Our fast dense-to-sparse attention

Figure 3: Comparison of conventional attention (Dosovitskiy et al. 2020) and our proposed dense-to-sparse attention.

task. Hence, as shown in the green box of Fig. 2, we propose the decomposed scale-consistent learning strategy.

Specifically, we decompose the absolute depth $\boldsymbol{D}$ into relative depth $\boldsymbol{d}$ prediction and global scale α estimation. The depth network N_D produces $\boldsymbol{d} \in (0, 1]$ with sigmoid function used, which is different from existing unsupervised depth completion methods (Ma, Cavalheiro, and Karaman 2019; Yang, Wong, and Soatto 2020; Wong and Soatto 2021) that directly generate $0{\sim}80m$ absolute depth. Meanwhile, we leverage the scale network N_S to estimate the real scale α with $\boldsymbol{E}_t$ as input, which is the final layer feature of the shared ResNet-18 encoder N_E. The DSCL is defined as

$$\begin{aligned} \boldsymbol{D} &= \alpha \cdot \boldsymbol{d}, \\ \alpha &= N_S\left(\boldsymbol{E}_t\right), \end{aligned} \tag{4}$$

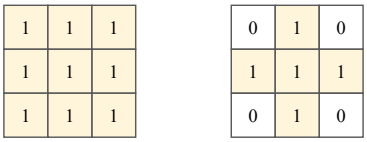
where $N_S(\cdot)$ refers to the corresponding function of depth network N_S. Other functions have the same definition next.

Here, we provide a theory to show that our DSCL can help the network to recover better depth using $\mathcal{L}_2$ loss.

Theorem 0.1. *Given a sparse depth $\boldsymbol{D}_t$ and a depth prediction $\boldsymbol{D}$ with network parameters w, if w does not satisfy $\boldsymbol{D} = 0$, then there exists a scale factor $\alpha \neq 0$ such that $\sum_\Omega(\boldsymbol{D}_t - \boldsymbol{D})^2 \geq \sum_\Omega(\boldsymbol{D}_t - \alpha\boldsymbol{D})^2$, where $\Omega \neq \emptyset$ is the index set of pixel location using the supervision.*

Proof. For simplicity, we consider only one pixel q in the proof process. For the traditional self-supervision, the loss is $\min_w(\boldsymbol{D}_t^q - \boldsymbol{D}^q)^2$. For DSCL, we introduce a scale factor α to the loss, and have a new loss $\min_w(\boldsymbol{D}_t^q - \alpha\boldsymbol{D}^q)^2$. It is easy to prove that if $\boldsymbol{D}^q \neq 0$, then there has a α such that $(\boldsymbol{D}_t^q - \boldsymbol{D}^q)^2 \geq (\boldsymbol{D}_t^q - \alpha\boldsymbol{D}^q)^2$. Furthermore, α has a closed form $\alpha = \boldsymbol{D}_t^q / \boldsymbol{D}^q$. When $\boldsymbol{D}^q = \boldsymbol{D}_t^q$, $\alpha = 1$. □

Theorem 0.1 shows that $\alpha\boldsymbol{D}$ is closer to the supervised signal $\boldsymbol{D}_t$ than the original $\boldsymbol{D}$. In fact, it is impossible for the network to predict $\boldsymbol{D}$ that equals to $\boldsymbol{D}_t$. Thus, $\alpha \neq 1$ drives that $\alpha\boldsymbol{D}$ has better approximation than $\boldsymbol{D}$. It reveals that our network has a strong practical significance to predict the depth. Furthermore, we employ a scale network to learn the scale factor from the data and map the absolute $\boldsymbol{D}$ to a relative $\boldsymbol{d}$, which is easier for networks to optimize.



(a) Square kernel (b) Cross kernel

Figure 4: Different kernels with square and cross shapes.

Global Depth Guidance

Existing unsupervised depth completion works (Ma, Cavalheiro, and Karaman 2019; Shivakumar et al. 2019; Yang, Wong, and Soatto 2020) have shown promising performance. However, most of them (Wong et al. 2020; Wong, Cicek, and Soatto 2021; Wong et al. 2021; Wong and Soatto 2021) suffer from depth holes in depth results. When we try to mitigate this issue, there are *two problems* impeding us.

The first problem is that, compared with related image processing works whose inputs are totally complete, the depth input $\boldsymbol{D}_t$ so sparse that it cannot provide dense information, which is the main cause of depth holes. To alleviate this issue, we take advantage of morphological dilation technology (Jackway and Deriche 1996) with different kernels (Fig. 4) to provide coarse but much denser depth $\boldsymbol{M}_t$.

The second problem is that, how to fuse the dense depth reference and the sparse depth input? Inspired by the conventional attention (Dosovitskiy et al. 2020), as illustrated in Fig. 3, we propose the dense-to-sparse attention. We first use N_E and N_G to map the RGB-D input ($\boldsymbol{I}_t$, $\boldsymbol{D}_t$) and the dense $\boldsymbol{M}_t$ into feature spaces, generating $\boldsymbol{N}_E^i$ and $\boldsymbol{N}_G^i$ in the ith network layer. Then, the dense-to-sparse attention propagates the dense depth reference $\boldsymbol{N}_G^i$ into the sparse depth feature $\boldsymbol{N}_E^i$, and thus obtaining the updated feature $\bar{\boldsymbol{N}}_E^i$. In addition, it is well-known that such attention has very high complexity even though its strong performance. Therefore, we design a fast version to deal with this issue.

The above process can be described as

$$\begin{aligned} \bar{\boldsymbol{N}}_E^i &= f_{att}\left(\boldsymbol{N}_E^i, \boldsymbol{N}_G^i\right), \\ \boldsymbol{N}_E^i &= N_E\left(\boldsymbol{I}_t, \boldsymbol{D}_t\right), \\ \boldsymbol{N}_G^i &= N_G\left(f_{dil}\left(\boldsymbol{D}_t\right)\right), \end{aligned} \tag{5}$$

where $f_{att}(\cdot)$ denotes the densn-to-sparse attention function and $f_{dil}(\cdot)$ is the morphological dilation function.

Different from the conventional attention (Dosovitskiy et al. 2020) in Fig. 3(a) that takes single-modal input as value, key, and query, our attention in Fig. 3(b) employs sparse depth feature as query, and dense depth reference as value and key. Also unlike the multi-modal attention (Rho, Ha, and Kim 2022; Li et al. 2022) which use absolutely different-modal (RGB and LiDAR) data, our attention only leverages LiDAR data, *i.e.*, sparse-modal and dense-modal depth. Additionally, to reduce the high complexity of (a) and (b) that equals to $(HW)^2$ when calculating the correlation between key and query, we design a fast version of the dense-to-sparse attention in Fig. 3(c), where *two key steps* are conducted. *One key step* is the binary mask, aiming to reduce redundancy existed in the long-range correlation since not every pixel in key is always related to that in query. *Another key step* is the strip average pooling, compressing the HW key and HW query into H and W, respectively. Then we multiply the compressed key by the compressed query to obtain the correlation matrix whose complexity is only HW, much smaller than $(HW)^2$ of the conventional attention.

Loss Function

The total loss function contains a cross-view reconstruction loss $\mathcal{L}_{t\to s}$, a single-view reconstruction loss $\mathcal{L}_{si}$, and a self-supervised loss $\mathcal{L}_2$ to predict the final depth result $\boldsymbol{D}$.

Cross-view reconstruction loss. Based on the geometry model defined in Eq. 1, the source frame $\boldsymbol{O}_s$ can be rebuilt from target frame $\boldsymbol{O}_t$ via $\widetilde{\boldsymbol{O}}_{t\to s}(q) = \boldsymbol{O}_t(\widetilde{q})$. Then, the cross-view reconstruction loss is

$$\mathcal{L}_{t\to s} = \sum\nolimits_q \left|\boldsymbol{O}_t\left(\widetilde{q}\right) - \boldsymbol{O}_s\left(q\right)\right|. \tag{6}$$

Single-view reconstruction loss. Given color image $\boldsymbol{I}$, the shared reconstruction network N_R maps the feature representation $\boldsymbol{O}$. The single-view reconstruction loss is

$$\begin{aligned} \mathcal{L}_{si} = &\sum\nolimits_q \left|\boldsymbol{I}(q) - \boldsymbol{O}(q)\right| + \alpha \sum\nolimits_q \left|\nabla^2 \boldsymbol{O}(q)\right| \\ &+ \beta\left(-\sum\nolimits_q e^{-\left|\nabla^1 \boldsymbol{I}(q)\right|} \cdot \left|\nabla^1 \boldsymbol{O}(q)\right|\right), \end{aligned} \tag{7}$$

where $\alpha = \beta = 1e-3$, ∇^1 and ∇^2 denote the first-order derivative and the the second-order derivative, respectively.

Decomposed scale-consistent learning loss. The final depth $\boldsymbol{D}$ is supervised by $\mathcal{L}_2$ loss, which can be defined as

$$\mathcal{L}_2 = \sum\nolimits_{\mathrm{q}} \left|\boldsymbol{D}_{gt}\left(q\right) - \boldsymbol{D}\left(q\right)\right|^2. \tag{8}$$

Total loss. Finally, the total loss function is written as

$$\mathcal{L} = \mathcal{L}_{t\to s} + \mathcal{L}_{si} + \gamma \mathcal{L}_2, \tag{9}$$

where γ is set to 1 during training. Please refer to (Shu et al. 2020) for more details about $\mathcal{L}_{t\to s}$ and $\mathcal{L}_{si}$ loss functions.

Experiment

Here, we first introduce related datasets and implementation details. Then we conduct ablation studies to verify the effectiveness of our method. Finally, we compare our method against other state-of-the-art approaches. Following KITTI benchmark, RMSE (mm) is selected as the *primary metric*.

Datasets and Implementation Details

KITTI benchmark (Uhrig et al. 2017) consists of **86,898** RGB-D pairs for training, 7,000 for validating, and another 1,000 for testing. The official 1,000 validation images are used during training while the remaining images are ignored. Following GuideNet (Tang et al. 2020), RGB-D pairs are bottom center cropped from 1216×352 to 1216×256, as there are no valid LiDAR values near top 100 pixels.

NYUv2 dataset (Silberman et al. 2012) contains 464 RGB-D indoor scenes with 640×480 resolution. Following KB-Net (Wong and Soatto 2021), we train our model on 46K frames and test on the official test set with 654 images. The sparse depth input is artificially produced by sampling about 1500 valid points from the ground-truth (GT) depth.

Implementation Details. We implement DesNet on Pytorch with 2 TITAN RTX GPUs. We train it for 25 epochs with Adam (Kingma and Ba 2014) optimizer. The learning rate is gradually warmed up to 10^{-4} in 3 steps, where each step increases learning rate by $10^{-4}/3$ in 500 iterations. After that, the learning rate 10^{-4} is used for the first 20 epochs and is reduced to half at the beginning of the 20th epoch.

Ablation Studies

This subsection verifies the effectiveness of DesNet, including the decomposed scale-consistent learning (DSCL) and global depth guidance (GDG), on **KITTI validation split**. *Gray background* in Tabs. 1- 4 refers to our *default setting*.

DesNet. As reported in Tab. 1, the baseline DesNet-i directly predicts absolute depth that is supervised by the total loss in Eq. 9 without using scale decomposition, *i.e.*, its depth network is an UNet which consists of our N_E and N_D without using sigmoid mapping. **(1)** When employing our DSCL strategy (DesNet-ii), we observe that all four evaluation metrics are consistently improved, *e.g.*, RMSE is reduced by $91.9mm$ and MAE by $24.1mm$. As shown in the 3rd and 4th columns of Fig. 5, DSCL notably contributes to sharper depth details and more complete object shapes. These numerical and visual results provide evidence that our DSCL, disintegrating the learning of absolute depth into *explicit* relative depth prediction and scale estimation, assuredly reduces the learning difficulty and brings individual learning benefits (Theorem 0.1). **(2)** When conducting our GDG module (DesNet-iii), the model performance is significantly improved. The RMSE, MAE, iRMSE, and iMAE outperform the baseline by $114.6mm$, $27.9mm$, $0.5km^{-1}$, and $0.3km^{-1}$, respectively. As illustrated in the 2nd and 3rd columns of Fig. 5, GDG remarkably corrects the wrong depth values near cars and can compensate depth holes well where even if the GT depth annotations have no valid pixels. These evidences indicate that GDG is able to provide valid

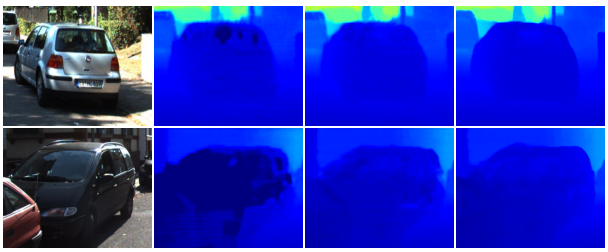


Figure 5: Visual comparison of ablation studies in Tab. 1.

DesNet	DSCL	GDG	RMSE	MAE	iRMSE	iMAE
i			1176.4	336.3	3.7	1.8
ii	✓		1084.5	312.2	3.3	1.6
iii		✓	1061.8	308.4	3.2	1.5
iv	✓	✓	**969.3**	**285.0**	**3.0**	**1.3**

Table 1: Ablation on DesNet with DSCL and GDG designs.

DSCL	RMSE	MAE	iRMSE	iMAE
num.=1	1084.5	312.2	3.3	1.6
num.=4	1076.8	310.5	3.3	1.5
num.=8	1070.6	308.7	3.2	1.5
num.=16	**1068.2**	**308.2**	**3.1**	**1.4**
num.=HW	1189.4	343.5	3.5	1.6

Table 2: Ablation on DSCL with different numbers (num.) of elements in the scale factor matrix.

reference for areas of missing depth, owing to the dilation and attention designs from which the model is urged to learn prior density. **(3)** Finally, to combine the best of both worlds, we simultaneously embed DSCL and GDG (DesNet-iv) into the baseline. Consequently, our model performs much better than DesNet-i, enormously surpassing it by $207.1mm$ in RMSE, $51.3mm$ in MAE, $0.7km^{-1}$ in iRMSE, and $0.5km^{-1}$ in iMAE. Besides, comparing the 4th column with the 2nd column of Fig. 5, it is noticeable that depth predictions of DesNet-iv clearly possess more reasonable visual effects than those of the baseline DesNet-i.

DSCL. The final layer of our scale network is the adaptive average pooling function. Therefore, the number of elements in the scale factor matrix can be arbitrary in theory. For example, num.=4 will lead to quartering relative depth. Then we multiply each of the quartering by the corresponding element in the scale factor matrix, finally obtaining the absolute depth prediction. Tab. 2 reports the cases of 1, 4, 8, 16, and HW. We can find that, **(1)** as the number increases (num.$\leq$ 16), the performance of the model gets better and better. It demonstrates that multiple region-aware scale elements is more accurate than single scale element for the full-resolution relative depth. **(2)** When num. reaches the maximum HW, the model performs even worse than the baseline DesNet-i in Tab. 1, which is mainly caused by the more difficult model learning. That said, two equally complex prediction targets, the full-resolution scale and relative depth, are harder for the network to learn than the single absolute depth target. Hence, the number of elements in our decomposed scale matrix is bounded in light of good performance.

GDG-dilation	size	RMSE	MAE	iRMSE	iMAE
bilinear	-	1162.7	335.1	3.6	1.8
nearest	-	1148.4	331.0	3.4	1.7
cross	3	1106.6	321.2	3.5	1.7
square	3	1088.3	315.7	3.3	1.6
square	5	**1061.8**	**308.4**	**3.2**	**1.5**
square	7	1112.9	324.3	3.5	1.6

Table 3: Ablation on GDG with different dense depth reference produced by bilinear/nearest interpolation and morphological dilation with cross/square dilation kernels.

GDG-attention	RMSE	MAE	Memory	Time
CA	1040.2	304.0	+13.64	+64.5
DSA	**1024.5**	**297.4**	+13.64	+64.6
FDSA w/o mask	1065.7	310.6	**+4.20**	**+12.3**
FDSA w/ mask	1061.8	308.4	+4.21	+12.5

Table 4: Ablation on GDG with different attentions, *i.e.*, the conventional attention (CA), our dense-to-sparse attention (DSA), and our fast dense-to-sparse attention (FDSA). GPU memory (G) and inference time (ms) are also considered.

GDG-dilation. Based on DesNet-i, Tab 3 displays the comparison of model performance using different manners to generate coarse but dense depth. On the whole, we observe that, **(1)** both interpolation and dilation methods can benefit our model since they can produce denser depth reference to compensate depth holes. Specifically, **(2)** the interpolation approach only slightly improves the baseline DesNet-i, while the dilation manner performs better. It shows that the dilation manages generating more precise dense depth than interpolation, of which can also find some evidence in (Ku, Harakeh, and Waslander 2018). **(3)** For one thing, dilation kernels with same size but different shapes have diverse performance. Square-3 is $18.3mm$ slightly lower than that of cross-3, owing to the higher-quality depth results with denser pixels and lower error generated by the denser square dilation kernel. For another thing, dilation kernels with same shape but different sizes still have slight distinctions. The last three rows of Table 3 verifies that square-5 achieves the lowest errors among square kernels with 3×3, 5×5, and 7×7 sizes, which can be viewed as size-accuracy trade-off.

GDG-attention. Based on square-5, Tab 4 validates different attention mechanisms. Overall, we discover that, **(1)** CA (Dosovitskiy et al. 2020), our DSA, and our FDSA consistently have positive impacts on error metrics since they effectively propagate valid dense depth reference into sparse targets. **(2)** Our DSA is superior to all three others in terms of RMSE and MAE. With similar complexity, DSA surpasses CA by $15.7mm$ in RMSE and $6.6mm$ in MAE, showing

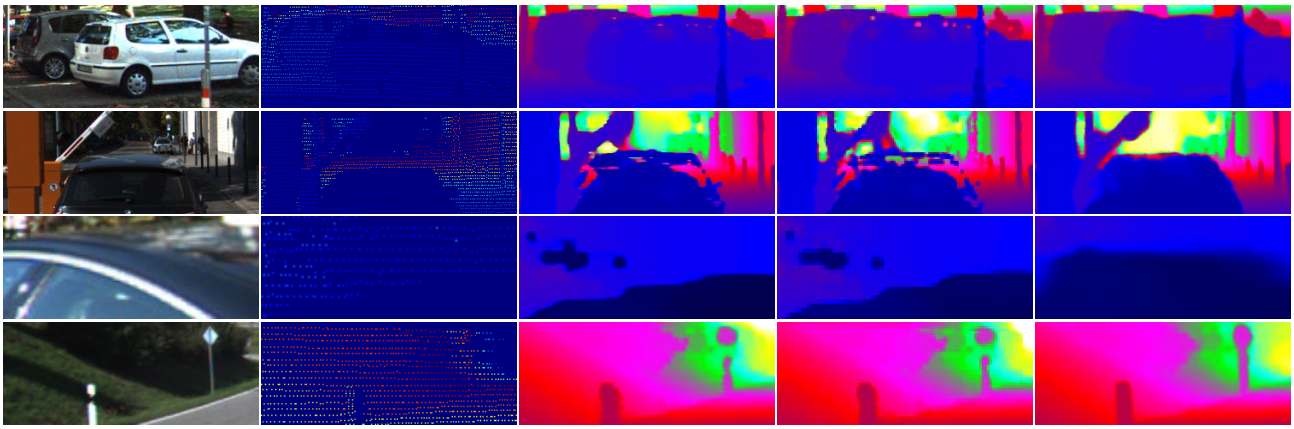


Figure 6: Visual comparison on KITTI online depth completion benchmark, where warmer color refers to longer distance.

Method	RMSE	MAE	iRMSE	iMAE	#P
S2D	1299.85	350.32	4.07	1.57	27.8
IP-Basic	1288.46	302.60	3.78	1.29	0.0
DFuseNet	1206.66	429.93	3.62	1.79	n/a
DDP*	1263.19	343.46	3.58	1.32	18.8
VOICED	1169.97	299.41	3.56	1.20	9.7
AdaFrame	1125.67	291.62	3.32	1.16	6.4
ScaffNet*	1121.93	280.76	3.30	1.15	7.8
SynthProj*	1095.26	280.42	3.53	1.19	2.6
KBNet	1068.07	**258.36**	3.01	**1.03**	6.9
DesNet	**938.45**	266.24	**2.95**	1.13	13.4

Table 5: Quantitative results on KITTI test set. $*$ denotes extra synthetic data and #P refers to model parameters (M).

that the dense-modal and sparse-modal fusion in DSA is more reasonable than the simple addition in CA. **(3)** Compared with CA, With acceptable performance degradation, *i.e.*, averagely $23.55mm$ in RMSE and $5.5mm$ in MAE, our FDSA without mask largely reduces GPU memory by 9.44G, and accelerate the inference speed from $64.6ms$ to $12.3ms$. Therefore, our FDSA design is GPU-friendly. Besides, FDSA with mask slightly outperforms FDSA without mask, demonstrating the robustness of our mask strategy.

Comparison with SoTA Methods

Here, we compare our DesNet with existing state-of-the-art (SoTA) methods on KITTI and NYUv2 datasets, including S2D, IP-Basic, DFuseNet, and Alex Wong's series of works.

On outdoor KITTI, as shown in Tab. 5, by combining DSCL and GDG designs, DesNet achieves the lowest RMSE & iRMSE and competitive MAE & iMAE among all mentioned approaches. Especially in RMSE, our DesNet surpasses the best KBNet by a large margin $129.62mm$, while KBNet outperforms the second best SynthProj* only by $27.19mm$. Also, DesNet obtains competitive results in MAE and iMAE, ranking second. As we know that, RMSE is very sensitive to large depth value while MAE is more sensitive to small one. Thus, **(i)** the lowest RMSE denotes that DesNet can predict more accurate depth in long-range region. **(ii)** Worse MAE indicates that DesNet is not very good enough at recovering precise depth in close-range region. **(iii)** However, as shown in Figs. 1 and 6, cars in close-range region predicted by DesNet are *much denser* than others. Our recovery has more reasonable visual effect. Besides, the semi-dense (about 30%) GT depth also lacks many valid points, where the pixels are ignored when computing errors, obscuring the merits of DesNet in terms of evaluation metrics. To further validate the generalization of our method, we conduct comparative experiment **on indoor NYUv2**. Tab. 6 demonstrates that our DesNet achieves outstanding performance as well, which is, *e.g.*, $9.51mm$ superior to the best KBNet in RMSE. In short, these evidences confirm that our DesNet actually possesses strong and robust performance.

Method	RMSE	MAE	iRMSE	iMAE
SynthProj	235.64	134.62	57.13	29.84
VOICED	228.38	127.61	54.70	28.89
ScaffNet	199.31	117.49	44.06	24.89
KBNet	197.77	105.76	42.74	**21.37**
DesNet	**188.26**	**103.42**	**38.57**	21.44

Table 6: Quantitative results on NYUv2 official test split.

Conclusion

In this paper, we proposed DesNet to utilize both the real scale information in sparse depth and the high accuracy of scale-agnostic counterpart for unsupervised depth completion, which decomposed the learning of absolute depth into relative depth prediction and global scale estimation. Such explicit learning does bring benefit. Further, to tackle the issue of depth holes, we introduced the global depth guidance to produce denser depth reference and attentively propagate it into the sparse target, severally using morphological dilation and dense-to-sparse attention. Owing to these designs, DesNet is remarkably superior to existing SoTA approaches.

Acknowledgements

The authors would like to thank all reviewers for their instructive comments. This work was supported by the National Science Fund of China under Grant Nos. U1713208 and 62072242. Note that the PCA Lab is associated with, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology.

References

Bian, J.; Li, Z.; Wang, N.; Zhan, H.; Shen, C.; Cheng, M.-M.; and Reid, I. 2019. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *NeurIPS*.

Chen, Y.; Schmid, C.; and Sminchisescu, C. 2019. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *ICCV*, 7063–7072.

Cheng, X.; Wang, P.; and Yang, R. 2018. Learning Depth with Convolutional Spatial Propagation Network. In *ECCV*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*.

Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. 2019. Digging into self-supervised monocular depth estimation. In *ICCV*.

Guizilini, V.; Ambrus, R.; Pillai, S.; Raventos, A.; and Gaidon, A. 2020. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2485–2494.

Jackway, P. T.; and Deriche, M. 1996. Scale-space properties of the multiscale morphological dilation-erosion. *IEEE transactions on pattern analysis and machine intelligence*, 18(1): 38–51.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *Computer Science*.

Ku, J.; Harakeh, A.; and Waslander, S. L. 2018. In defense of classical image processing: Fast depth completion on the cpu. In *CRV*.

Lee, J. H.; Han, M.-K.; Ko, D. W.; and Suh, I. H. 2019. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*.

Li, Y.; Yu, A. W.; Meng, T.; Caine, B.; Ngiam, J.; Peng, D.; Shen, J.; Lu, Y.; Zhou, D.; Le, Q. V.; et al. 2022. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *CVPR*, 17182–17191.

Lin, Y.; Cheng, T.; Zhong, Q.; Zhou, W.; and Yang, H. 2022. Dynamic spatial propagation network for depth completion. In *AAAI*.

Lu, K.; Barnes, N.; Anwar, S.; and Zheng, L. 2020. From Depth What Can You See? Depth Completion via Auxiliary Image Reconstruction. In *CVPR*.

Ma, F.; Cavalheiro, G. V.; and Karaman, S. 2019. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *ICRA*.

Mu, H.; Le, H.; Yikai, B.; Jian, R.; Jin, X.; and Jian, Y. 2022. RA-Depth: Resolution Adaptive Self-Supervised Monocular Depth Estimation. In *ECCV*.

Park, J.; Joo, K.; Hu, Z.; Liu, C.-K.; and Kweon, I. S. 2020. Non-Local Spatial Propagation Network for Depth Completion. In *ECCV*.

Petrovai, A.; and Nedevschi, S. 2022. Exploiting Pseudo Labels in a Self-Supervised Learning Framework for Improved Monocular Depth Estimation. In *CVPR*, 1578–1588.

Qiu, J.; Cui, Z.; Zhang, Y.; Zhang, X.; Liu, S.; Zeng, B.; and Pollefeys, M. 2019. DeepLiDAR: Deep Surface Normal Guided Depth Prediction for Outdoor Scene From Sparse LiDAR Data and Single Color Image. In *CVPR*.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*.

Rho, K.; Ha, J.; and Kim, Y. 2022. GuideFormer: Transformers for Image Guided Depth Completion. In *CVPR*, 6250–6259.

Roy, A.; and Todorovic, S. 2016. Monocular depth estimation using neural regression forest. In *CVPR*.

Saxena, A.; Chung, S. H.; Ng, A. Y.; et al. 2005. Learning depth from single monocular images. In *NeurIPS*.

Shivakumar, S. S.; Nguyen, T.; Miller, I. D.; Chen, S. W.; Kumar, V.; and Taylor, C. J. 2019. Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion. In *ITSC*.

Shu, C.; Yu, K.; Duan, Z.; and Yang, K. 2020. Feature-metric Loss for Self-supervised Learning of Depth and Egomotion. In *ECCV*.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*.

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

Tiwari, L.; Ji, P.; Tran, Q.-H.; Zhuang, B.; Anand, S.; and Chandraker, M. 2020. Pseudo rgb-d for self-improving monocular slam and depth prediction. In *ECCV*, 437–455.

Uhrig, J.; Schneider, N.; Schneider, L.; Franke, U.; Brox, T.; and Geiger, A. 2017. Sparsity Invariant CNNs. In *3DV*.

Van Gansbeke, W.; Neven, D.; De Brabandere, B.; and Van Gool, L. 2019. Sparse and noisy lidar completion with rgb guidance and uncertainty. In *MVA*, 1–6.

Wang, L.; Wang, Y.; Wang, L.; Zhan, Y.; Wang, Y.; and Lu, H. 2021. Can Scale-Consistent Monocular Depth Be

Learned in a Self-Supervised Scale-Invariant Manner? In *ICCV*, 12727–12736.

Wang, L.; Zhang, J.; Wang, O.; Lin, Z.; and Lu, H. 2020. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *CVPR*, 541–550.

Wong, A.; Cicek, S.; and 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.; Hong, B.-W.; and Soatto, S. 2021. An Adaptive Framework for Learning Unsupervised Depth Completion. *IEEE Robotics and Automation Letters*, 6(2): 3120–3127.

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

Wong, A.; and Soatto, S. 2021. Unsupervised Depth Completion with Calibrated Backprojection Layers. In *ICCV*.

Xian, K.; Zhang, J.; Wang, O.; Mai, L.; Lin, Z.; and Cao, Z. 2020. Structure-guided ranking loss for single image depth prediction. In *CVPR*, 611–620.

Yan, Z.; Li, X.; Wang, K.; Zhang, Z.; Li, J.; and Yang, J. 2022a. Multi-modal masked pre-training for monocular panoramic depth completion. *arXiv preprint arXiv:2203.09855*.

Yan, Z.; Wang, K.; Li, X.; Zhang, Z.; Li, G.; Li, J.; and Yang, J. 2022b. Learning Complementary Correlations for Depth Super-Resolution With Incomplete Data in Real World. *IEEE Transactions on Neural Networks and Learning Systems*.

Yan, Z.; Wang, K.; Li, X.; Zhang, Z.; Li, J.; and Yang, J. 2022c. RigNet: Repetitive image guided network for depth completion. In *ECCV*.

Yang, Y.; Wong, A.; and Soatto, S. 2020. Dense Depth Posterior (DDP) From Single Image and Sparse Range. In *CVPR*.

Zhang, Z.; Cui, Z.; Xu, C.; Yan, Y.; Sebe, N.; and Yang, J. 2019. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*.

Zhao, W.; Liu, S.; Shu, Y.; and Liu, Y.-J. 2020. Towards better generalization: Joint depth-pose learning without posenet. In *CVPR*.

Zhong, Y.; Wu, C.-Y.; You, S.; and Neumann, U. 2016. Deep rgb-d canonical correlation analysis for sparse depth completion. In *NeurIPS*, volume 32.

Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised learning of depth and ego-motion from video. In *CVPR*.

Zhu, Y.; Dong, W.; Li, L.; Wu, J.; Li, X.; and Shi, G. 2022. Robust Depth Completion with Uncertainty-Driven Loss Functions. In *AAAI*.