Zero-shot Depth Completion via Test-time Alignment with Affine-invariant Depth Prior

Lee Hyoseok[1], Kyeong Seon Kim[2], Kwon Byung-Ki[1], Tae-Hyun Oh[1,2,3]
[1]Grad.School of Artificial Intelligence, POSTECH
[2]Dept. of Electrical Engineering, POSTECH
[3]Institute for Convergence Research and Education in Advanced Technology, Yonsei University
hyos99@postech.ac.kr, ella94.ai@gmail.com, byungki.kwon@postech.ac.kr, taehyun@postech.ac.kr

Abstract

Depth completion, predicting dense depth maps from sparse depth measurements, is an ill-posed problem requiring prior knowledge. Recent methods adopt learning-based approaches to implicitly capture priors, but the priors primarily fit in-domain data and do not generalize well to out-of-domain scenarios. To address this, we propose a zero-shot depth completion method composed of an affine-invariant depth diffusion model and test-time alignment. We use pre-trained depth diffusion models as depth prior knowledge, which implicitly understand how to fill in depth for scenes. Our approach aligns the affine-invariant depth prior with metric-scale sparse measurements, enforcing them as hard constraints via an optimization loop at test-time. Our zero-shot depth completion method demonstrates generalization across various domain datasets, achieving up to a 21% average performance improvement over the previous state-of-the-art methods while enhancing spatial understanding by sharpening scene details. We demonstrate that aligning a monocular affine-invariant depth prior with sparse metric measurements is a sufficient strategy to achieve domain-generalizable depth completion without relying on extensive training datasets.

1 Introduction

Metric-scale dense depth provides precise spatial structure of a scene, crucial for physically accurate applications such as 3D scene understanding (Ji-Yeon et al. 2024), 3D reconstruction (Choe et al. 2021), and robotic grasping (Viereck et al. 2017). This depth information is essential for achieving reliable and robust performance across real-world perception and interaction, where failures can lead to significant risks. However, acquiring dense metric depth map in practical settings is challenging, as depth measurements captured by depth sensing approaches – long-range sensors (*e.g.*, LiDAR) (Ma and Karaman 2018) and SLAM/VIO systems (Wong and Soatto 2021) – are sparse potentially leading to safety risks. To complement this limitation, depth completion has been studied, which aims to complete the dense metric depth map from sparse measurements.

However, depth completion is an ill-posed problem requiring prior knowledge and additional cues, *e.g.*, RGB images as guidance (Ma and Karaman 2018; Hu et al. 2021; Tang et al. 2020; Qiu et al. 2019). Previous studies (Park et al. 2020; Zhang et al. 2023; Wong and Soatto 2021; Wang et al. 2023b) have focused on learning how to propagate sparse metric depth into a dense map according to the color or texture proximity. They are trained with paired dense depth maps and corresponding RGB images to learn depth affinity as prior knowledge, where the depth affinity represents the relationship between depth values in a scene based on spatial and structural features. Since previous methods (Zhang et al. 2023; Wong and Soatto 2021) focused on learning depth affinity within in-domain settings, they exhibit poor depth affinity in out-of-domain scenarios (see Fig. 1). To address this, Park, Gupta, and Wong (2024) proposed a test-time adaptation method that fine-tunes part of a pre-trained depth completion model using sparse depth. Nevertheless, This approach is less effective in out-of-domain scenarios due to the limited generalizability of the base depth completion model.

With the emergence of foundation models (Caron et al. 2021; Rombach et al. 2022), which learn comprehensive knowledge from large image data (referred to as image prior), these models have been frequently utilized as powerful prior to improve generalizability, enabling them to be applicable across diverse tasks and domains (Lee et al. 2024; Yang et al. 2023; Liu et al. 2023). We bring this versatile capability to the depth completion problem. In this regime, we propose zero-shot depth completion via a test-time alignment, which is generalizable to any domain by leveraging the rich semantic and structural understanding of the foundation model.

Specifically, we use pre-trained monocular depth diffusion models (Ke et al. 2024; Gui et al. 2024) as depth prior, demonstrating generalizability and facilitating high-quality depth estimation. Most monocular depth estimation models (Ranftl et al. 2022; Ke et al. 2024; Yang et al. 2024; Gui et al. 2024) operate in the affine-invariant depth space, where depth values are consistent up to offset and scale. While this approach enables training on large-scale dataset with diverse scene contents and varying camera intrinsics (Ke et al. 2024), it inherently introduces scale ambiguity, making fully accurate monocular metric depth estimation to be considered infeasible (Yin et al. 2023). Meanwhile, depth completion is free from scale ambiguity thank to sparse measurements of metric depths, but lacks generalizability and depth quality (Park, Gupta, and Wong 2024). Motivated by these trade-offs, we align the affine-invariant depth prior with sparse measure-

Figure 1: **3D-lifted depth completion results in out-of-domain cases.** Regardless of supervised (Zhang et al. 2023) or unsupervised methods (Wong and Soatto 2021), most depth completion models perform poorly on out-of-domain data. In contrast, our zero-shot depth completion method, which employs test-time alignment, consistently achieves robust results. In this example, the other models are trained on the KITTI Depth Completion dataset (Uhrig et al. 2017), while our zero-shot approach is not trained on any specific depth completion dataset. Both are tested on the nuScenes dataset (Caesar et al. 2020).

ments in the metric depth space, achieving generalizable and well-structured depth completion. By performing this alignment at test time, we can complete the metric depth map from any pair of RGB and synchronized sparse depth data, *i.e.*, zero-shot. Figure 1 illustrates the robustness of our method in the out-of-domain scenarios.

To this end, we propose a test-time alignment method that guides the reverse sampling process of the diffusion model by incorporating optimization loops to enforce the given sparse depth as hard constraints. We also introduce a prior-based outlier filtering method to ensure reliable measurements and a new loss function to maintain the structural prior inherent in the depth prior. Our method demonstrates superior generalization ability across various domain datasets (Silberman et al. 2012; McCormac et al. 2017; Sun et al. 2020; Caesar et al. 2020), including both indoor and outdoor environments. Our contribution points are as follows:

- We propose a novel zero-shot depth completion method that leverages foundation model prior to enhance domain generalization while capturing detailed scene structure.
- We introduce a test-time alignment that uses sparse measurements as hard constraint to guide the diffusion sampling process, aligning with an affine-invariant depth prior.
- We present a prior-based outlier filtering algorithm to improve the reliability of sparse measurements, enhancing the robustness of our method using sparse depth guidance.

2 Related Work

Depth completion. Depth completion is an ill-posed problem that aims to reconstruct unknown dense depth from observed sparse depth measurements, with missing areas typically covering less than 5% of an image for outdoor driving scenarios and 1% for indoor scenarios (Wong et al. 2020). Since the success of deep learning, the problem has been addressed by data-driven approaches that learn how to propagate sparse depth measurements guided by the RGB images (Wong and Soatto 2021; Park et al. 2020). Prior studies (Park et al. 2020; Lin et al. 2022; Zhang et al. 2023) use affinity-based spatial propagation methods (Liu et al. 2017) to learn the relationship between dense depth and RGB pairs. They learn how to propagate depth while preserving scene structure and boundaries. This learning process requires large pairs of RGB images and dense depth maps, but acquiring these dense maps in real-world scenarios is costly due to dedicated sensor systems and requires careful data processing and curation. (Uhrig et al. 2017; Wong et al. 2020). Depending on how to process data, domain discrepancies are introduced in each dataset, which makes depth completion models hard to generalize.

To mitigate these challenges arising from the lack of real data and domain gaps, unsupervised learning or domain adaptation methods have been proposed. Unsupervised methods (Wong and Soatto 2021; Ma, Cavalheiro, and Karaman 2019; Wong et al. 2020) train a model with pairs of a RGB image and synchronized sparse depth without a dense depth map. These methods exploit multi-view photometric consistency with multiple views to compensate for the lack of direct 3D supervision. As an alternative direction to mitigate lack of data and domain gaps, some works (Wong, Cicek, and Soatto 2021; Lopez-Rodriguez, Busam, and Mikolajczyk 2020) is initially trained in a synthetic domain with supervised learning, followed by unsupervised training on real datasets as a way of domain adaptation. Different from these research, we tackle the limitations by exploiting learned prior embeded in a foundation model. We use a pre-trained generative diffusion model that understands depth affinity, spatial detail, and scene context. This strong prior from the foundation model further enables zero-shot generalization to any domain.

Test-time Adaptation (TTA). Applying a model trained on a source domain to unseen test domains is crucial for generalization, especially in depth completion, where domain gaps arise from sensor variations, environmental conditions (*e.g.*, weather changes), scene variety (*e.g.*, driving locations), and depth ranges (*e.g.*, indoor vs. outdoor). TTA methods (Wang et al. 2021, 2022; Park, Gupta, and Wong 2024) address this by adapting models to unseen data. However, they still suffer from domain gaps due to reliance on the source dataset, and often require additional training and continual adaptation, which may not be feasible in zero-shot scenarios.

With the emergence of foundation models, there has been

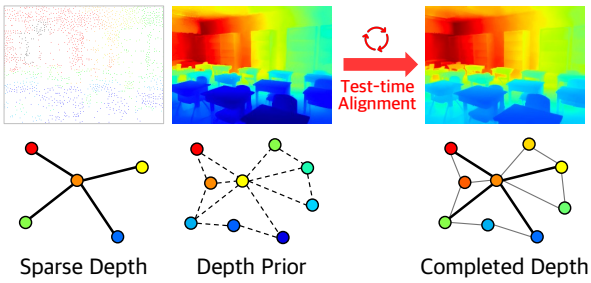

Figure 2: **Illustration of our approach.** At test time, we align the depth affinity from the prior (dashed lines) with the sparse depth measurements as a hard constraint (bold lines). This alignment propagates measurements across the scene to complete unobservable depth values.

a shift towards leveraging their prior knowledge for generalization across diverse tasks and domains (Jia et al. 2024; Liu et al. 2023). As a generative foundation model, diffusion models are similarily employed as a generalizable priors. To address the domain gaps in depth completion, we utilize a diffusion model that comprehends depth prior (Ke et al. 2024; Gui et al. 2024) by aligning it with sparse depth measurement using the proposed test time alignment method. This approach effectively mitigates issues caused by domain gaps and enables depth completion in a zero-shot manner.

3 Method

In this section, we introduce our zero-shot depth completion method, which leverages the depth prior (Ke et al. 2024; Gui et al. 2024) derived from the foundation model (Rombach et al. 2022). This enables our method to be generalizable across any domain. The core concept of our approach is to align the affine-invariant depth prior with sparse measurements on an absolute scale to complete the dense and well-structured depth map, as illustrated in Fig. 2.

3.1 Preliminary

Diffusion model and guided sampling. Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021) aim to model data distribution $p(\mathbf{x})$ through iterative perturbation and restoration, known as forward and reverse processes. This is represented by the score-based generative model (Song et al. 2021), learning the score function $\mathbf{s}_\theta$ parameterized by θ the gradient of the log probability density function with respect to the data, *i.e.*, $\mathbf{s}_\theta(\mathbf{x}) = \nabla_\mathbf{x} \log p(\mathbf{x}; \theta)$. Score-based diffusion models estimate the score $\mathbf{s}_\theta(\mathbf{x}_t)$ at intermediate state $\mathbf{x}_t$ for timestep t which defines a process.

For image generation and editing, diffusion models leverage the guidance function during the sampling process to adjust the output to the specific condition (Ho and Salimans 2022; Dhariwal and Nichol 2021). The guidance can be defined by any differentiable mapping output to guidance modality, as follows (Bansal et al. 2024):

$$\hat{\mathbf{s}}_\theta(\mathbf{x}_t, t, \mathbf{y}) = \mathbf{s}_\theta(\mathbf{x}_t, t) + w \nabla_{\mathbf{x}_t} \mathcal{L}\left(f\left(\mathbf{x}_0\left(\mathbf{x}_t\right)\right), \mathbf{y}\right), \quad (1)$$

where w and $\mathbf{y}$ represent weight and guidance, respectively. The function $f(\cdot)$ can be any differentiable function whose output can compute a loss $\mathcal{L}$ with guidance condition $\mathbf{y}$, and $\mathbf{x}_0\left(\mathbf{x}_t\right)$ is obtained by using Tweedie's formula (Efron 2011), which provides an approximation of the posterior mean. This guided sampling approach extends unconditional diffusion models to conditional ones without separate model training.

Inverse problem. The goal of an inverse problem is to determine an unknown variable from known measurement, often formulated as $\mathcal{A}(\mathbf{x}) = \mathbf{y}$, where $\mathcal{A}: \mathbb{R}^m \to \mathbb{R}^n$ represents the known forward measurement operator, $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{x} \in \mathbb{R}^m$, the measurement and the unknown variable, respectively. When $m > n$, it becomes an ill-posed problem, requiring a prior to find solve a Maximum A Posterior (MAP) estimation:

$$\arg\max p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x}) p(\mathbf{y}|\mathbf{x}), \quad (2)$$

where $p(\mathbf{x})$ represents our prior of the signal $\mathbf{x}$ and $p(\mathbf{y}|\mathbf{x})$ is likelihood measuring $\mathcal{A}(\mathbf{x}) \approx \mathbf{y}$, *e.g.*, $\|\mathbf{y} - \mathcal{A}(\mathbf{x})\|_2^2$. By taking $-\log(\cdot)$ to Eq. (2), it can be formulated as an optimization problem that regularizes the solution, ensuring that $\mathbf{x}$ follows the characteristics of the prior:

$$\arg\min_\mathbf{x} \|\mathbf{y} - \mathcal{A}\left(\mathbf{x}\right)\|_2^2 - \log p(\mathbf{x}). \quad (3)$$

Also, given the gradient of $\log p(\mathbf{x}|\mathbf{y})$ in Eq. (2) as

$$\nabla_\mathbf{x} \log p(\mathbf{x}|\mathbf{y}) = \nabla_\mathbf{x} \log p(\mathbf{x}) + \nabla_\mathbf{x} \log p(\mathbf{y}|\mathbf{x}), \quad (4)$$

the prior term $\nabla_\mathbf{x} \log p(\mathbf{x})$ corresponds to the score $\mathbf{s}_\theta(\mathbf{x})$, which can be obtained by diffusion models. Therefore, by simply adding the gradient of the likelihood term to the reverse sampling process, the inverse problem can be effectively solved while leveraging the diffusion prior (Chung et al. 2023) as follows:

$$\hat{\mathbf{s}}_\theta(\mathbf{x}_t, t, \mathbf{y}) = \mathbf{s}_\theta(\mathbf{x}_t, t) + w \nabla_{\mathbf{x}_t} \|\mathbf{y} - \mathcal{A}\left(\mathbf{x}_0\left(\mathbf{x}_t\right)\right)\|_2^2. \quad (5)$$

This has an analogous form with Eq. (1); thus, the inverse problem can be effectively tackled with the guided sampling.

With pre-trained image diffusion models, *e.g.*, Rombach et al. (2022), as the score function $\mathbf{s}_\theta(\mathbf{x})$ and a prior, it provides powerful image prior across various tasks by its comprehensive semantic understanding and structural knowledge learned from a lot of images (Wang et al. 2023a; Namekata et al. 2024). Ke et al. (2024) leverage this rich visual knowledge to achieve generalizable monocular depth estimation, resulting in high-quality outputs within an affine-invariant depth space. In our work, we exploit this depth diffusion model for computing the score as a depth prior.

Problem formulation. To leverage the prior knowledge, we formulate a depth completion as an inverse problem that estimates unknown dense depth from observed sparse measurements. $\mathbf{y}$ represents the observed sparse depth, $\mathbf{x}$ is the unknown dense depth, and $\mathcal{A}: \mathbb{R}^m \to \mathbb{R}^n$ is a binary measurement matrix of which entry $[\mathcal{A}]_{ij}$ is 1 if the entities $[\mathbf{y}]_i$ is measured from $[\mathbf{x}]_j$, 0 otherwise. We follow Eq. (5), where sparse depth serves as guidance. We use the depth diffusion models (Ke et al. 2024; Gui et al. 2024) extended from the latent diffusion model (LDM) (Rombach et al. 2022) as prior, where $\mathbf{x}$ is decomposed with the decoder $\mathcal{D}: \mathbf{z} \to \mathbf{x}$ as:

$$\hat{\mathbf{s}}_\theta = \mathbf{s}_\theta(\mathbf{z}_t, t) + w \nabla_{\mathbf{z}_t} \|\mathbf{y} - \mathcal{A}\left(\mathcal{D}\left(\mathbf{z}_0\left(\mathbf{z}_t\right)\right)\right)\|_2^2, \quad (6)$$

where $\mathbf{z} \in \mathbb{R}^{4 \times H \times W}$ represents the latent of LDM but the decoder output $\mathbf{x}$ is treated as a flatten vector for convenience.

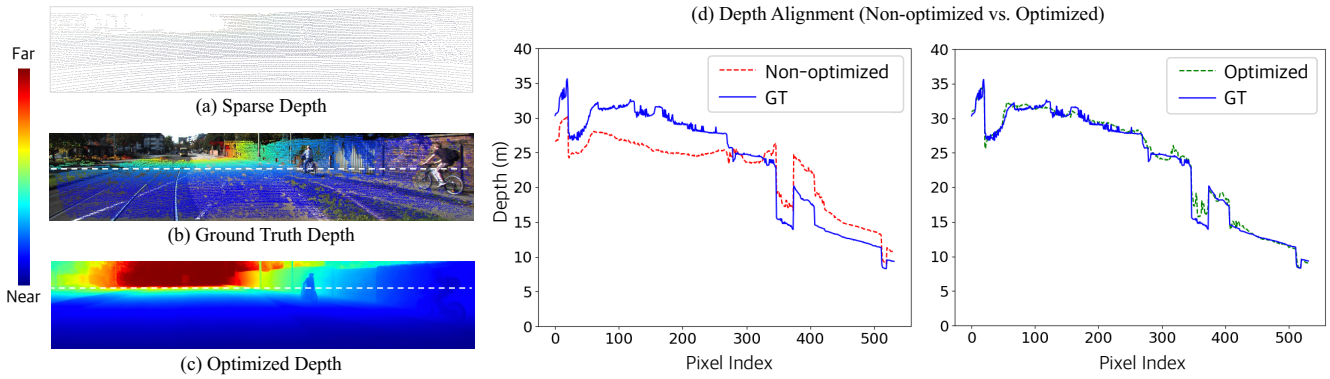

Figure 3: **Alignment with metric depth.** We evaluate our method's effectiveness against ground truth (GT), accumulated semi-densely. We use only sparse depth (a) to align with actual metric depth values in complex scenes, ensuring a desirable solution. The white lines in (b), (c), and the x-axis of (d) represent pixel indices with valid depth points in a row of the GT.

3.2 Test-time Alignment with Hard Constraints

Depth measurements obtained in practice are often sparse, unevenly distributed, and noisy. When the sparse measurements are used as guidance, the ill-posed nature of the problem, combined with the stochastic behavior of diffusion models, can lead to scores that produce undesirable solutions (Kim et al. 2024) and does not even guarantee that the estimation corresponds to the known sparse measurements.To deal with this, we propose a test-time alignment that incorporates the correction step to enforce the sparse measurement as harder constraints than encouraging guidance in a soft manner by Eq. (6). This involves an optimization loop at regular intervals to enforce measurement constraints as a correction step.

Additionally, we adopt $\mathbf{z}_0(\mathbf{z}_t)$ as optimizable variable. Pre-trained diffusion models take input $\mathbf{z}_t$ aligend with the noise level at each timestep t. However, directly optimizing $\mathbf{z}_t$ without considering input characteristics may lead to suboptimal results (Chung et al. 2022, 2023; Chung, Lee, and Ye 2024). To address this, inspired by Song et al. (2024), we use $\mathbf{z}_0(\mathbf{z}_t)$ estimated from $\mathbf{z}_t$. The optimization loop is formulated as:

$$\hat{\mathbf{z}}_0(\mathbf{z}_t) = \underset{\mathbf{z}_0(\mathbf{z}_t)}{\arg\min} \|\mathbf{y} - \mathcal{A}\left(\mathcal{D}\left(\mathbf{z}_0\left(\mathbf{z}_t\right)\right)\right)\|_2^2 . \tag{7}$$

Then, to ensure adherence to the correct noise level, the measurement-consistent $\hat{\mathbf{z}}_0(\mathbf{z}_t)$ is remapped to an intermediate latent $\hat{\mathbf{z}}_t$ by adding time-scheduled Gaussian noise, as expressed below:

$$p\left(\hat{\mathbf{z}}_t|\hat{\mathbf{z}}_0(\mathbf{z}_t)\right) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\,\hat{\mathbf{z}}_0(\mathbf{z}_t), (1-\bar{\alpha}_t)I), \tag{8}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and α_t is variance schedule at time t.

Since the score $\hat{\mathbf{s}}_\theta(\mathbf{z}_t, t)$ is directly added to the latent $\mathbf{z}_t$ at each step, Eq. (6) can be rewritten in terms of $\mathbf{z}_0(\mathbf{z}_t)$ with a modulated weight factor ζ, as follows:

$$\hat{\mathbf{z}}_t = \mathbf{z}_t + \zeta \nabla_{\mathbf{z}_t} \|\mathbf{y} - \mathcal{A}\left(\mathcal{D}\left(\mathbf{z}_0(\mathbf{z}_t)\right)\right)\|_2^2 . \tag{9}$$

Here, Eq. (9) is replaced by the two-step process of Eq. (7) and Eq. (8), allowing our test-time alignment process to effectively achieve measurement-consistent desirable solutions. Figure 4 illustrates the our test-time alignment process. Figure 3 demonstrates how effectively our test-time alignment

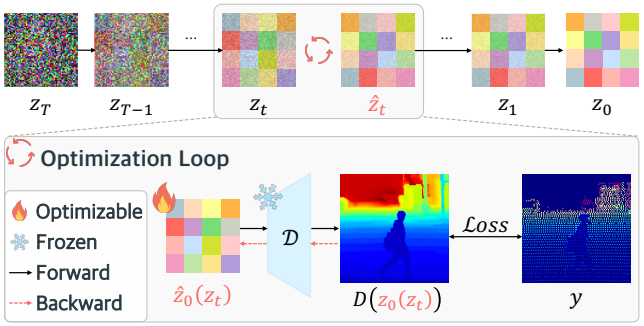

Figure 4: **Test-time alignment process.** We incorporate an optimization loop into the reverse sampling process, optimizing $\mathbf{z}_0(\mathbf{z}_t)$ at regular intervals. The latent is then decoded into depth, where the loss is measured against sparse depth. For visibility, the sparse depth points are enlarged.

method estimates unseen depth areas by aligning sparse measurements with an affine-invariant depth prior. This result highlights the need for correction.

Until now, in solving Eq. (6), we use an affine-invariant depth model for completing metric depths without special care. However, a natural question arises: "*Is the affine-invariant depth model compatible with estimating metric depths in our framework?*" The following analysis shows that it may be sufficient.

Can we use an affine-invariant depth model for completing metric depths? Depth estimation models are often trained to estimate affine-invariant depth with scale and shift invariant loss to achieve generalizable performance (Ranftl et al. 2022; Ke et al. 2024; Eigen, Puhrsch, and Fergus 2014). Thus, depth prior operates in the affine-invariant depth space, which does not directly correspond to the metric depth used in measurements. Even though the given sparse metric depth is normalized between 0 and 1, their statistics including mean and variance can differ, and the relationship between real metric depth and estimated affine-invariant depth is often non-linear (see the left of Fig. 3 (d)). Therefore, to determine if Eq. (6) can be used to solve this problem, we need to verify

whether the normalized metric depth space lies within the data distribution generated by the diffusion model.

To confirm this, we conduct an empirical investigation through the following procedure: given $\tilde{\mathbf{x}}_0$, dense depth map estimated from the pre-trained depth completion model, we perform its reconstruction using an affine-invariant depth diffusion model. This process involves sequentially encoding $\tilde{\mathbf{x}}_0$ to $\tilde{\mathbf{z}}_0$, then doing inversion by adding noise (Song, Meng, and Ermon 2021), which results in $\tilde{\mathbf{z}}_t$. Next, we perform reverse sampling, $\nabla_{\mathbf{z}_t} \log p(\tilde{\mathbf{z}}_t)$ with only the affine-invariant depth diffusion prior. The reconstructed result achieves similar performance compared to the original one, $\tilde{\mathbf{x}}_0$, excluding encoding-decoding information loss. This result suggests that the affine-invariant depth prior is sufficiently capable of handling the metric depth space, which corresponds to:

$$\nabla_{\mathbf{z}_t} \log p(\tilde{\mathbf{z}}_t) \approx \nabla_{\tilde{\mathbf{z}}_t} \log p(\tilde{\mathbf{z}}_t). \tag{10}$$

Thus, we just need to align this prior with metric depth cue validating using Eq. (6) to solve ill-posed depth completion.

3.3 Prior-based Outlier Filtering

Practical depth sensing methods often produce outliers, such as unsynchronized depth with RGB or see-through points (Conti et al. 2022)), making sparse depth measurements unreliable. This degrades the performance of methods relying on sparse depth supervision (Wong and Soatto 2021; Wong et al. 2020). We also use sparse depth measurement as supervision during test-time alignment, this makes the alignment process prone to divergence or slow convergence. To address this, we utilize data-driven depth prior (Ke et al. 2024; Gui et al. 2024), which benefits from the more precise synchronization with RGB images and depth affinity. To obtain outlier-free sparse points $\mathbf{y}^*$, we adopt a divide-and-conquer approach. We define local segments based on depth affinity, grouping regions where relative depth values are similar within a spatially local area. Within these segments, the depth distribution can be easily categorized into inliers and outliers, enabling us to effectively identify outliers.

Affine-invariant depth map D_r is divided into local segments S_i, which are regions with a high probability of having similar depths with considering location. For this clustering we leverage the superpixel algorithm (Achanta et al. 2012; Li and Chen 2015). In each region, we perform linear least-square fitting to map affine-invariant depth to metric depth using sparse metric depth measurements $\mathbf{y}_i$. However, since these sparse measurements are influenced by outliers, we use RANSAC (Fischler and Bolles 1981) to perform outlier-robust linear least-square fitting on points where noisy $\mathbf{y}$ intersects S_i *i.e.*, $\mathbf{y}_i \leftarrow S_i \cap \mathbf{y}$. This allows us to estimate outlier-robust metric depth values $\hat{\mathbf{y}}_i$ in local regions S_i. Then, points with significant deviations exceeding τ are identified as outliers and filtered out. Our proposed filtering algorithm, based on monocular depth prior, is detailed in Algorithm 1.

3.4 Losses

Our objective for optimization includes sparse depth consistency loss and regularization terms: a local smoothness loss to preserve depth prior and a new relative structure similarity loss to maintain structural prior inherent in depth prior. Sparse depth consistency loss $\mathcal{L}_{depth}$ is an L1 loss that enforces consistency with metric depth using sparse depth. Local smoothness loss $\mathcal{L}_{smooth}$ regularizes to preserve knowledge of depth, ensuring the locally smooth property in the depth diffusion model is not lost. However, using only these loss functions may dilute the structural prior in the pre-trained depth diffusion model, which is key for detail sharpness.

To address this, we design a new structure regularization term that transfers structure from the depth estimated by an off-the-shelf model to regularize overly smooth structures. Inspired by the SSIM (Wang et al. 2004), we propose the Relative SSIM (R-SSIM) loss, designed to transfer structure across domains. This loss is derived from SSIM by dropping the luminance term, which relies on absolute values:

$$\mathcal{L}_{r-ssim}(d_1, d_2) = 1 - \frac{2\sigma_{d_1 d_2} + C}{\sigma_{d_1}^2 + \sigma_{d_2}^2 + C}, \tag{11}$$

where d_1 and d_2 represent spatial information in different domains, C is a constant, and σ denotes the normalized standard deviation of pixel values. Here, d_1 is the relative depth map, and d_2 is the estimated complete depth map (or vice versa). The key point is that these domains may differ in pixel value ranges and statistics.

Algorithm 1: Prior-based outlier filtering algorithm.

Parameters: Number of segments N, Filter threshold τ
Input: Estimated relative depth D_r, Sparse metric depth $\mathbf{y}$, Set of sparse point locations $\Omega(\mathbf{y})$.
Output: Set of reliable sparse point locations $\Omega(\mathbf{y}^*)$.
$\{\Omega(S_i)\}_{i=1\cdots N} \leftarrow$ SuperPixel (D_r, N)
for $i = 1$ **to** N **do**
    $\Omega(\mathbf{y}_i) \leftarrow \Omega(\mathbf{y}) \cap \Omega(S_i)$
    $\hat{\mathbf{y}}_i \leftarrow$ RANSAC Regressor$(\mathbb{1}_{\Omega(\mathbf{y}_i)} \odot D_r,\ \mathbf{y}_i)$
    $\Omega(\mathbf{y}_i^*) \leftarrow |\hat{\mathbf{y}}_i - \mathbf{y}_i| > \tau$
$\Omega(\mathbf{y}^*) \leftarrow \bigcup_{i=1}^{N} \Omega(\mathbf{y}_i^*)$

4 Experiments

In this section, we demonstrate the effectiveness of our prior-based depth completion method in indoor (NYUv2 (Silberman et al. 2012), SceneNet (McCormac et al. 2017), VOID (Wong et al. 2020)) and outdoor (Waymo (Sun et al. 2020), nuScenes (Caesar et al. 2020), KITTI DC (Uhrig et al. 2017)) scenarios, through both quantitative and qualitative evaluations. For evaluation, we use the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), both standard metrics in depth completion where lower values indicate better performance. The results are reported in meters.

4.1 Domain Generalization

Table 1 summarizes the domain generalization performance of our method and previous test-time adaptation methods (Wang et al. 2021, 2022; Park, Gupta, and Wong 2024) on indoor (NYU, SceneNet) and outdoor (Waymo, nuScenes). Across various datasets, our prior-based approach consistently achieves the best or second-best performance. No-

Method	Indoor				Outdoor			
	NYUv2		SceneNet		Waymo		nuScenes	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Pre-trained	0.446	0.189	0.443	0.173	2.821	1.514	3.998	1.967
BNAdapt	0.410	0.189	0.446	0.176	2.194	1.122	1.801	0.828
CoTTA	0.376	0.147	0.405	0.136	2.652	1.227	2.668	1.222
ProxyTTA	0.203	0.095	0.357	0.125	2.178	0.971	1.755	0.799
Ours (+Marigold)	0.149	0.059	0.207	0.099	2.115	1.121	1.561	0.561
Ours (+DepthFM)	0.145	0.077	0.178	0.081	2.162	1.133	1.622	0.618

Table 1: **Quantitative comparison of generalizable performance.** We evaluate the generalizability of our method by comparing it with test-time adaptation methods across various domain datasets. In this table, the pre-trained depth completion model is CostDCNet (Kam et al. 2022), trained on KITTI DC for outdoor and VOID for indoor adaptation. It is used for each adaptation method—BNAdapt (Wang et al. 2021), CoTTA (Wang et al. 2022), and ProxyTTA (Park, Gupta, and Wong 2024)—excluding ours, for adapting to each domain. The first best is marked in red, the second in orange, and the third in yellow.

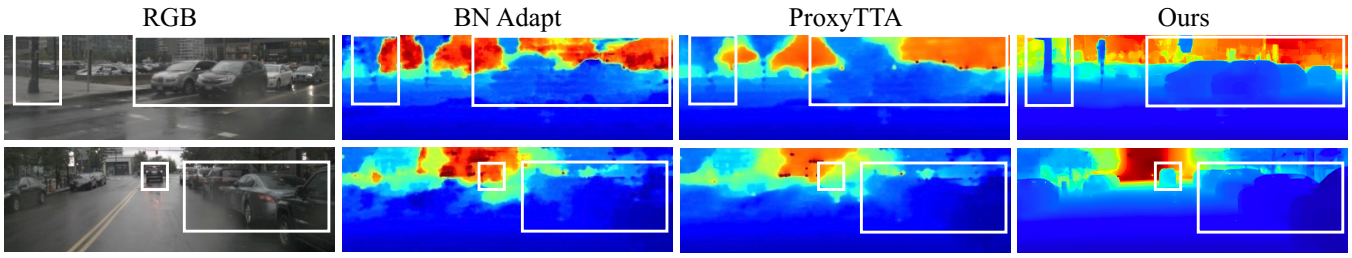

Figure 5: **Qualitative comparison on the nuScenes test set.** In outdoor scenarios, our test-time alignment method performs robustly even under extreme weather conditions, clearly identifying critical elements such as vehicles and signs.

Base Model	Inference time	RMSE	MAE
Marigold (**50** steps)	101s	1.413	0.397
DepthFM (**2** steps)	31s	1.499	0.377
DepthFM (**1** step)	16s	1.601	0.428

Table 2: **Efficiency evaluation on the KITTI validation set.** Inference time of our method is measured as base models (Ke et al. 2024; Gui et al. 2024) with varying sampling

tably, unlike test-time adaptation methods relying on pre-trained depth completion models in metric depth space, our method operates in affine-invariant depth space while achieving impressive performance. Additionally, we demonstrate the model generality of our method by applying it to two depth diffusion models, Marigold (Ke et al. 2024) and DepthFM (Gui et al. 2024), as shown in Table 1. Table 2 further presents the inference time of our method across base models and sampling steps, demonstrating its potential for improving efficiency with minimal performance. We also observe that our method captures details on the scene, reflecting true performance and demonstrating robust domain generalization as shown in Fig. 5 and 6.

In the outdoor datasets, the ground truth is obtained by accumulating LiDAR points after removing those corresponding to moving objects, which can lead to variations in the ground truth. For a more reliable benchmark, we use the ground truth provided by Park, Gupta, and Wong (2024) for the Waymo and by Huang et al. (2022) for the nuScenes.

4.2 Comparison with Unsupervised Methods

We compare our zero-shot depth completion method with unsupervised methods (Wong and Soatto 2021; Ma, Cavalheiro, and Karaman 2019; Wong, Cicek, and Soatto 2021) trained on the split training dataset of each benchmark, *i.e.*, in-domain training. As shown in Table 3, our method demonstrates favorable performance without dense depth data, multi-view, and in-domain training on KITTI DC and VOID. Additionally, our method achieves comparable performance when adopting manual filtering, that is, the outlier filtering method suggested by each benchmark. Figure 7 shows qualitative results of ours and unsupervised methods. Our method achieves higher-fidelity depth completion, preserving the depth affinity better than other unsupervised methods.

4.3 Ablation Studies

Table 4 shows ablation studies to assess the efficacy of the test-time alignment method, R-SSIM loss, and outlier filtering algorithm. The ablation studies are conducted on both indoor (VOID) and outdoor (KITTI DC) datasets. Compared to other sampling methods, *i.e.*, no guidance and the guided sampling (Bansal et al. 2024), the proposed test-time alignment method brings significant performance gain. The R-SSIM loss further enhances the performance and has a remarkable effect on preserving depth affinity. The prior-based outlier filtering is more effective on the outdoor dataset than on the indoor dataset, as the sparse depth in the indoor dataset consists of reliable points sampled from the ground truth.

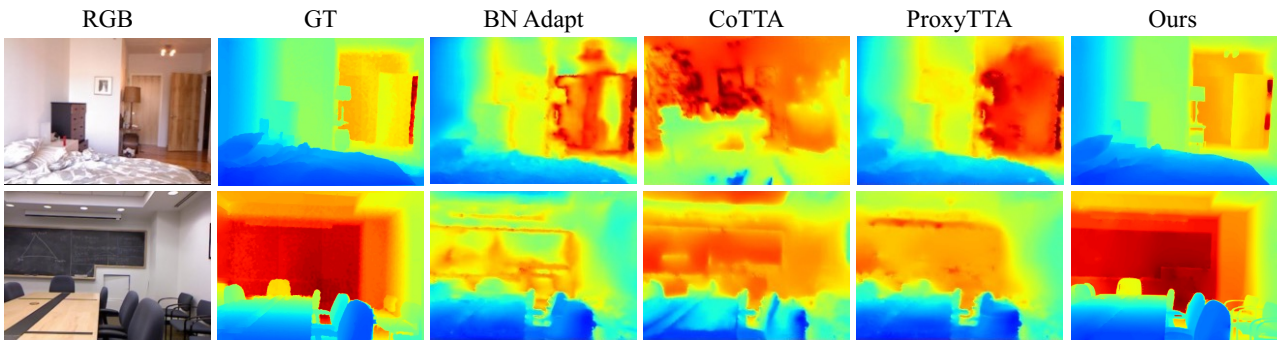

Figure 6: **Qualitative comparison on the NYU test set.** In indoor scenarios, our test-time alignment method accurately captures scene structures (*e.g.*, chairs) compared to the existing test-time adaptation methods.

Method	Features			KITTI DC		VOID	
	Sparse Depth Supervision	Photometric Consistency Loss	In-domain Training	RMSE	MAE	RMSE	MAE
Self-S2D	✓	✓ (two-view)	✓	1.384	0.358	0.243	0.178
VOICED	✓	✓ (multi-view)	✓	1.230	0.308	0.169	0.085
ScaffNet	✓	✓ (multi-view)	✓	1.182	0.286	0.119	0.059
KBNet	✓	✓ (multi-view)	✓	1.126	0.260	0.095	0.039
SPTR	✓	✓ (multi-view)	✓	1.111	0.254	0.091	0.040
Ours w/ Our Filtering	✓	✗ (monocular)	✗	1.413	0.397	0.111	0.044
Ours w/ Manual Filtering				1.198	0.287	0.112	0.045

Table 3: **Quantitative comparison with unsupervised methods.** Despite weaker settings, our method performs comparably to unsupervised methods (Self-S2D (Ma, Cavalheiro, and Karaman 2019), VOICED (Wong et al. 2020), ScaffNet (Wong, Cicek, and Soatto 2021), KBNet (Wong and Soatto 2021), and SPTR (Zhao et al. 2024)) when sparse depth, *i.e.* the supervision signal, is reliable. To demonstrate this, we ablate two filtering methods: our prior-based filtering and manual filtering, which is the outlier filtering method suggested by each benchmark. In this table, our method uses Marigold (Ke et al. 2024) as the base model.

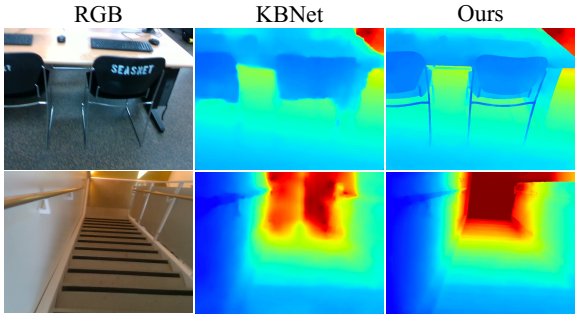

Figure 7: **Qualitative comparison on theVOID test set.** Compared to the state-of-the-art unsupervised method KBNet (Wong and Soatto 2021), which uses multi-view photometric consistency, our prior-based approach better preserves scene structures and details using only monocular input.

Sampling Method	R-SSIM Loss	Outlier Filtering	KITTI DC		VOID	
			RMSE	MAE	RMSE	MAE
Naïve			3.514	1.942	0.199	0.130
Guided			2.113	0.801	0.210	0.138
Ours			1.610	0.406	0.125	0.046
Ours	✓		1.502	0.409	0.111	0.044
Ours	✓	✓	1.413	0.397	0.112	0.045

Table 4: **Ablation studies.** We ablate our proposed methods including test-time alignment, R-SSIM loss, and prior-based outlier filtering, to demonstrate their effectiveness.

5 Conclusion

We propose a novel prior-based zero-shot depth completion method, the first study demonstrating the importance of monocular depth prior knowledge in addressing the challenge of domain shifts. Our test-time alignment approach ensures that the completed depth map remains consistent with sparse measurements while incorporating structural depth affinity of the scene derived from the depth prior. This prior-based approach enhances the performance of depth completion across various domains, capturing the context of the scene.

Limitation. Our zero-shot depth completion is the first work to use monocular depth foundation model priors for generalizable depth completion, but it adopts the standard guided sampling approach in latent diffusion models, which may be slow to process. As a next step, accelerating this process through consistency models and developing effective priors using consistency models could be promising directions.

Acknowledgements

This work was supported by the RideFlux and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (Development of Artificial Intelligence Technology for Self-Improving Competency Aware Learning Capabilities; No.RS-2022-II220124, Artificial Intelligence Innovation Hub; No. 2019-0-01906, Artificial Intelligence Graduate School Program(POSTECH)).

Code — https://github.com/postech-ami/Zero-Shot-Depth-Completion

Project page — https://hyoseok1223.github.io/zero-shot-depth-completion/

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11): 2274–2282.

Bansal, A.; Chu, H.-M.; Schwarzschild, A.; Sengupta, S.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2024. Universal Guidance for Diffusion Models. In *Int. Conf. Learn. Represent.*

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; ; and Beijbom, O. 2020. nuscenes: A multi- modal dataset for autonomous driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Int. Conf. Comput. Vis.*

Choe, J.; Im, S.; Rameau, F.; Kang, M.; and Kweon, I. S. 2021. Volumefusion: Deep depth fusion for 3d scene reconstruction. In *Int. Conf. Comput. Vis.*, 16086–16095.

Chung, H.; Kim, J.; Mccann, M. T.; Klasky, M. L.; and Ye, J. C. 2023. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *Int. Conf. Learn. Represent.*

Chung, H.; Lee, S.; and Ye, J. C. 2024. Decomposed Diffusion Sampler for Accelerating Large-Scale Inverse Problems. In *Int. Conf. Learn. Represent.*

Chung, H.; Sim, B.; Ryu, D.; and Ye, J. C. 2022. Improving Diffusion Models for Inverse Problems using Manifold Constraints. In *Adv. Neural Inform. Process. Syst.*

Conti, A.; Poggi, M.; Aleotti, F.; and Mattoccia, S. 2022. Unsupervised confidence for LiDAR depth maps and applications. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Adv. Neural Inform. Process. Syst.*

Efron, B. 2011. Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association*, 106(496): 1602–1614. PMID: 22505788.

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

Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6): 381–395.

Gui, M.; Fischer, J. S.; Prestel, U.; Ma, P.; Kotovenko, D.; Grebenkova, O.; Baumann, S. A.; Hu, V. T.; and Ommer, B. 2024. DepthFM: Fast Monocular Depth Estimation with Flow Matching. arXiv:2403.13788.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Adv. Neural Inform. Process. Syst.*

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Hu, M.; Wang, S.; Li, B.; Ning, S.; Fan, L.; and Gong, X. 2021. PENet: Towards Precise and Efficient Image Guided Depth Completion. In *IEEE International Conference on Robotics and Automation*.

Huang, S.; Gojcic, Z.; Huang, J.; and Andreas Wieser, K. S. 2022. Dynamic 3D Scene Analysis by Point Cloud Accumulation. In *Eur. Conf. Comput. Vis.*

Ji-Yeon, K.; Hyun-Bin, O.; Byung-Ki, K.; Kim, D.; Kwon, Y.; and Oh, T.-H. 2024. Uni-DVPS: Unified Model for Depth-Aware Video Panoptic Segmentation. *IEEE Robotics and Automation Letters*, 9(7): 6186–6193.

Jia, Y.; Hoyer, L.; Huang, S.; Wang, T.; Gool, L. V.; Schindler, K.; and Obukhov, A. 2024. DGInStyle: Domain-Generalizable Semantic Segmentation with Image Diffusion Models and Stylized Semantic Control. In *European Conference on Computer Vision, ECCV*.

Kam, J.; Kim, J.; Kim, S.; Park, J.; and Lee, S. 2022. CostDCNet: Cost Volume Based Depth Completion for a Single RGB-D Image. In *Eur. Conf. Comput. Vis.*, 257–274. Springer.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Kim, J.; Park, G. Y.; Chung, H.; and Ye, J. C. 2024. Regularization by Texts for Latent Diffusion Inverse Solvers. arXiv:2311.15658.

Lee, H.-Y.; Tseng, H.-Y.; Lee, H.-Y.; and Yang, M.-H. 2024. Exploiting Diffusion Prior for Generalizable Dense Prediction. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Li, Z.; and Chen, J. 2015. Superpixel segmentation using Linear Spectral Clustering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 1356–1363.

Lin, Y.; Cheng, T.; Zhong, Q.; Zhou, W.; and Yang, H. 2022. Dynamic Spatial Propagation Network for Depth Completion. In *AAAI*.

Liu, S.; Mello, S. D.; Gu, J.; Zhong, G.; Yang, M.-H.; and Kautz, J. 2017. Learning Affinity via Spatial Propagation Networks. In *Adv. Neural Inform. Process. Syst.*

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv:2303.05499.

Lopez-Rodriguez, A.; Busam, B.; and Mikolajczyk, K. 2020. Project to Adapt: Domain Adaptation for Depth Completion from Noisy and Sparse Sensor Data. In *Asian Conf. Comput. Vis.*

Ma, F.; Cavalheiro, G. V.; and Karaman, S. 2019. Self-supervised Sparse-to-Dense: Self-supervised Depth Completion from LiDAR and Monocular Camera. In *IEEE International Conference on Robotics and Automation*.

Ma, F.; and Karaman, S. 2018. Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image. In *IEEE International Conference on Robotics and Automation*.

McCormac, J.; Handa, A.; Leutenegger, S.; and Davison, A. J. 2017. Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. In *Int. Conf. Comput. Vis.*

Namekata, K.; Sabour, A.; Fidler, S.; and Kim, S. W. 2024. EmerDiff: Emerging Pixel-level Semantic Knowledge in Diffusion Models. In *Int. Conf. Learn. Represent.*

Park, H.; Gupta, A.; and Wong, A. 2024. Test-Time Adaptation for Depth Completion. In *IEEE Conf. Comput. Vis. Pattern Recog.*

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

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 *IEEE Conf. Comput. Vis. Pattern Recog.*

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2022. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3).

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *Eur. Conf. Comput. Vis.*, 746–760. Springer.

Song, B.; Kwon, S. M.; Zhang, Z.; Hu, X.; Qu, Q.; and Shen, L. 2024. Solving Inverse Problems with Latent Diffusion Models via Hard Data Consistency. In *Int. Conf. Learn. Represent.*

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *Int. Conf. Learn. Represent.*

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *Int. Conf. Learn. Represent.*

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; Vasudevan, V.; Han, W.; Ngiam, J.; Zhao, H.; Timofeev, A.; Ettinger, S.; Krivokon, M.; Gao, A.; Joshi, A.; Zhang, Y.; Shlens, J.; Chen, Z.; and Anguelov, D. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*

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.

Uhrig, J.; Schneider, N.; Schneider, L.; Franke, U.; Brox, T.; and Geiger, A. 2017. Sparsity Invariant CNNs. In *International Conference on 3D Vision (3DV)*.

Viereck, U.; Pas, A.; Saenko, K.; and Platt, R. 2017. Learning a visuomotor controller for real world robotic grasping using simulated depth images. In *Conference on robot learning*, 291–300. PMLR.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *Int. Conf. Learn. Represent.*

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C. K.; and Loy, C. C. 2023a. Exploiting Diffusion Prior for Real-World Image Super-Resolution. arXiv:2305.07015.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual Test-Time Domain Adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Wang, Y.; Li, B.; Zhang, G.; Liu, Q.; Gao, T.; and Dai, Y. 2023b. LRRU: Long-short Range Recurrent Updating Networks for Depth Completion. In *Int. Conf. Comput. Vis.*

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

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.; 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 *Int. Conf. Comput. Vis.*

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; of Methods, C. M. A. C. S.; Applications, B.; and Yang, M.-H. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39.

Yin, W.; Zhang, C.; Chen, H.; Cai, Z.; Yu, G.; Wang, K.; Chen, X.; and Shen, C. 2023. Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image. In *Int. Conf. Comput. Vis.*

Zhang, Y.; Guo, X.; Poggi, M.; Zhu, Z.; Huang, G.; and Mattoccia, S. 2023. Completionformer: Depth completion with convolutions and vision transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhao, L.; Zheng, W.; Duan, Y.; Zhou, J.; and Lu, J. 2024. SPTR: Structure-Preserving Transformer for Unsupervised Indoor Depth Completion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4): 2439–2452.