This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Marigold-DC: Zero-Shot Monocular Depth Completion with Guided Diffusion

Massimiliano Viola Kevin Qu Nando Metzger Bingxin Ke Alexander Becker
Konrad Schindler Anton Obukhov

ETH Zürich

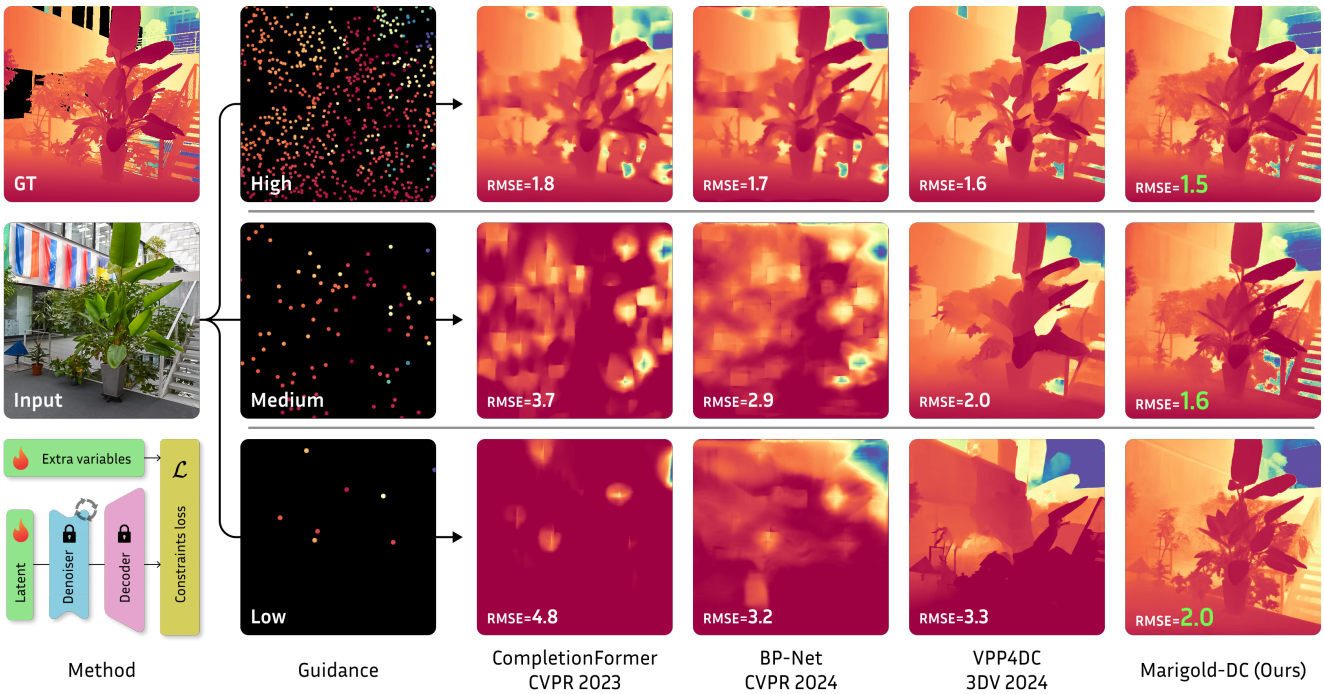

Figure 1. **Marigold-DC is a zero-shot, generative method for depth completion.** It leverages a pretrained, diffusion-based monocular depth estimator as scene understanding prior and integrates sparse depth guidance into the denoising process. No fine-tuning is required, as the method utilizes test-time optimization to update the latent representation. Compared to existing methods, Marigold-DC recovers plausible depth maps even from very sparse depth observations and excels at zero-shot generalization across a broad range of scenes.

Abstract

*Depth completion upgrades sparse depth measurements into dense depth maps, guided by a conventional image. Existing methods for this highly ill-posed task operate in tightly constrained settings, and tend to struggle when applied to images outside the training domain, as well as when the available depth measurements are sparse, irregularly distributed, or of varying density. Inspired by recent advances in monocular depth estimation, we reframe depth completion as image-conditional depth map generation, guided by a sparse set of measurements. Our method, **Marigold-DC**, builds on a pretrained latent diffusion model (LDM) for depth estimation and injects the depth observations as test-time guidance, via an optimization scheme that runs in tandem with the iterative inference of denoising diffusion. The method exhibits excellent zero-shot generalization across a diverse range of environments and handles even extremely sparse guidance effectively. Our results suggest that contemporary monodepth priors greatly robustify depth completion: it may be better to view the task as recovering dense depth from (dense) image pixels, guided by sparse depth; rather than as inpainting (sparse) depth, guided by an image. Project website: https://MarigoldDepthCompletion.github.io/*

1. Introduction

Depth completion aims to convert sparse depth measurements into a dense depth map, using an image – typically standard RGB or grayscale – as guidance (Fig. 1). It is

useful across a range of computer vision applications that combine conventional cameras with sparser range sensors, e.g., robotics, autonomous driving, and 3D city modeling. Traditional approaches predominantly rely on convolutional neural networks (CNNs) [6, 7, 47, 54] or transformer models [58, 85, 88] and achieve satisfactory results within their particular problem setting. However, they do not generalize well and fail, often catastrophically, when transferred to new domains (Fig. 1). This is all the more concerning because depth sensors are barely standardized and each system comes with its own sampling pattern, data gaps, and noise characteristics.

In contrast, depth estimation only from a single view, *without* depth guidance, has progressed to a point where it generalizes remarkably well and can handle a broad range of images "in the wild" [3, 57, 81, 83], which begs the question: why does depth completion not generalize at least as well?

The answer, we claim, lies in the much more powerful visual prior. Modern monodepth estimators build on top of foundation models like DINOv2 [53] or Stable Diffusion [60] and inherit their rich knowledge about the structure of the visual world. We therefore bridge the gap between depth completion and monocular depth and adapt Marigold [33], a latent diffusion model (LDM) for monodepth estimation, to the depth completion task. Marigold casts depth estimation from a single image as generating a depth map, conditioned on the input image. Marigold-DC uses this capability as a basis and adds sparse depth observations as a further guidance signal. Importantly, our proposed scheme is based on test-time optimization and does not alter the Marigold model; thus avoiding the risk of degrading the prior, as well as the effort of collecting or generating suitable training data.

Our proposed guidance scheme exploits the iterative nature of Denoising Diffusion Probabilistic Models (DDPMs) [28, 72]. This class of models has revolutionized image generation. They are capable of synthesizing photorealistic images from pure noise, and by appropriate training, the generative process can be conditioned on various inputs, including text [60, 65] but also images and depth maps [86]. What is more, also fully trained DDPMs can, at test time, be guided towards specific outputs by injecting suitable signals into the inference process [1, 15], which opens up the possibility to repurpose them for certain applications without having to retrain them.

In this paper, we leverage such test-time guidance for depth completion, by guiding the inference loop of a pretrained monodepth estimator with additional depth input. Our contributions are:

- We rethink depth completion from the perspective of monocular depth prediction, such that it can benefit from the comprehensive visual knowledge baked into state-of-the-art monodepth estimators, and the associated ability to generalize.
- We introduce a computational scheme that seamlessly integrates sparse depth cues into the diffusion process of a pretrained LDM, and thus achieves depth completion without any architectural modifications or retraining of that base model.
- We design a strategy to anchor affine-invariant predictions in latent space on sparse depth cues in metric space. By dynamically adjusting the corresponding scale and shift parameters during inference, our method effectively aligns model predictions with the available depth measurements.

In experiments on several datasets, we demonstrate that Marigold-DC sets a new state of the art for the depth completion task, especially in the desirable but challenging zero-shot setting. With our take on depth completion as a constrained form of monocular depth estimation, we hope to close the widening performance gap between those two closely related computer vision problems and inspire further research towards depth completion methods that generalize to unseen images, environments, and sensor setups.

2. Related Work

2.1. Depth Completion

Early depth completion methods rely solely on sparse depth inputs and employ classical interpolation techniques or sparsity-invariant convolutional neural networks (CNNs) [8, 36, 75]. These approaches often produce blurred predictions lacking fine structural details, especially around object boundaries. To address this, it has become common to incorporate an RGB image as guidance, enabling sharper transitions and improved extrapolation in regions without depth information. Other works [30, 47, 48, 73] leverage both sparse depth and RGB inputs using encoder-decoder multi-modal fusion networks with a ResNet [26] backbone.

Advancements include multi-scale prediction objectives [31, 37], intermediate representations such as surface normals [55, 79, 87] or coarse depth estimates [42], and graph-based approaches for modeling neighborhood relations [78, 89]. Post-processing refinement methods, mostly following the spatial propagation network (SPN) mechanism [43], have been proposed to enhance output quality. Notable examples include CSPN [6] and its successor CSPN++ [7], which introduce convolutions with fixed and adaptive kernels, respectively, improving efficiency. Further improvements involve non-local neighborhoods in NLSPN [54], adaptive affinity matrices in DySPN [40], and varying kernel scopes in LRRU [76]. This paradigm has been extended to a three-stage method in BP-Net [74].

While most architectures process features in 2D, some methods [5, 69, 80, 85] leverage 3D geometry information, though this requires knowledge of camera intrinsics, limiting generalization. The vision transformer (ViT) [16],

widely successful in computer vision, has also been explored for depth completion in works like GuideFormer [58], PointDC [85], and CompletionFormer [88]. To handle sparse and irregular patterns, SpAgNet [12] proposes a sparsity-agnostic framework and obtains reasonable predictions even with $<$10 guidance points. For robust generalization, VPP4DC [2] revisits depth completion from a fictitious stereo-matching perspective, OGNI-DC [92] iteratively optimizes a depth gradient field, and Prompt Depth Anything [39] fine-tunes a feedforward depth foundation model [82] for sparse depth prompting.

Generative approaches have also been explored: DepthFM [24] employs flow matching [41, 44] conditioned on the RGB image and the sparse depth, densified with distance functions, to implement refinement. Similarly, DepthLab [46] embeds the interpolated sparse depth in latent space to condition a diffusion model alongside the RGB. Such densification before or during processing is a viable strategy, as also seen in concurrent work [29] leveraging depth priors and [23], which encodes a blend of sparse guidance and intermediate predictions back into the latent space. The drawback of intermediate densification is that it introduces high-frequency variations around sparse guidance points, resulting in corruption of the latent codes. This requires spatial smoothing to manage artifacts, adding further design choices and complexity to the guidance framework. In contrast, our guidance approach integrates observations via test-time optimization, penalizes decoded predictions in the pixel space, and can accommodate varying sparsity levels, *cf*. Fig. 1.

2.2. Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) [28] generate high-quality samples by reversing a Gaussian noise diffusion process. Their enhanced sample quality [15] and computational efficiency [51] have been well documented. Denoising Diffusion Implicit Models (DDIMs) [70] offer speedups with non-Markovian inference. Conditional diffusion models allow controlled generation by incorporating inputs like text [65], images [64], and semantic maps [86]. Stable Diffusion (SD) [60] exemplifies text-based image generation, using an LDM trained on the large-scale LAION-5B dataset [68]. By performing denoising in compressed latent space via a U-Net [61], it reduces complexity, making the model more scalable and easier to fine-tune. A separately trained variational autoencoder (VAE) maps images to and from latent space, thus embedding extensive image knowledge into the model weights, which has been utilized for various downstream tasks.

2.3. Diffusion Guidance

To allow fine-grained control over the output, guidance-based diffusion [14] incorporates external supervision alongside the original conditioning, using a guidance function that measures whether certain criteria are met. In guided image generation, classifier guidance [15] enables class-conditional outputs from a pretrained, unconditional diffusion model, via gradients from a classifier trained on ImageNet [63] images at different noise scales. Similarly, gradients from a CLIP model [56] trained on noisy images can guide generation toward a user-defined text caption [52]. An alternative, classifier-free guidance [27, 52], achieves similar control without training a separate classifier, by parameterizing both conditional and unconditional diffusion models within the same network. The approach is further extended to handle general nonlinear inverse problems [9, 10], using gradients calculated on the expected denoised images. Alternatively, one can optimize the constraints on the clean signal [91] and then reintroduce noise. Guidance is commonly framed from a score-based perspective on denoising diffusion [71, 72], where an unconditional model approximates the time-dependent score function of the log data distribution. Finally, a variety of universal constraints, such as segmentation masks, image style, and object location, have been applied with SD under a single framework [1], fully exploiting the flexibility and control of diffusion-based image generation.

2.4. Diffusion for Monocular Depth Estimation

Monocular depth estimation predicts per-pixel depth from a single RGB image, a challenging and ill-posed problem due to the absence of definitive depth information. Deep learning approaches leverage features learned from large datasets to tackle this. More recently, several methods have employed DDPMs for generative depth estimation. DDP [32] conditions diffusion on image features for dense visual prediction. DiffusionDepth [17] performs latent space diffusion conditioned on features from a Swin Transformer [45]. DepthGen [67] and its successor DDVM [66] extend multi-task diffusion models for depth estimation, addressing noisy ground truth and emphasizing pretraining on synthetic and real data for improved quality. VPD [90] utilizes a pretrained SD with text input as its image feature extractor for various visual perception tasks, highlighting the semantic knowledge embedded in these models.

Marigold [33] repurposes Stable Diffusion to denoise depth maps conditioned on an input image, achieving impressive zero-shot monocular depth estimation with fine details across diverse datasets. Trained purely on synthetic data and relying on the foundational knowledge of Stable Diffusion, this affine-invariant model outputs depth predictions in a fixed $[0, 1]$ range. Motivated by our view that depth completion is essentially monocular depth estimation anchored at sparse points, we develop a plug-and-play optimization framework around Marigold by lifting its output to metric space, enabling effective depth completion without fine-tuning or architectural changes.

3. Method

3.1. Guided Diffusion Formulation

We formulate depth completion as a guided monocular depth estimation task and use Marigold, the pretrained, affine-invariant, diffusion-based model, as our prior. At inference time, we dynamically refine its predictions by incorporating a penalizing loss $\mathcal{L}$ at sparse point measurements in metric space. Marigold generates a linearly normalized depth map $\hat{\mathbf{d}} \in \mathbb{R}^{W \times H}$ within the range $[0, 1]$ by sampling from the conditional distribution $D(\mathbf{d} \mid \mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^{W \times H \times 3}$ is an RGB image, and $\mathbf{d}$ is a pixel-wise corresponding depth map. Let $\mathbf{c} \in \mathbb{R}^{W \times H}$ denote a sparse metric depth map in the same image space, where only a limited subset of pixels contains valid depth values. Let further $\hat{a}$ and $\hat{b}$ represent the scale and shift coefficients for linear prediction scaling, maintained as additional parameterized variables throughout the process and initialized as in Sec. 3.2.

The proposed modified inference pipeline for depth completion is presented in Fig. 2. We start by encoding the input image $\mathbf{x}$ into latent space, using the SD encoder $\mathcal{E}$ to obtain the latent code $\mathbf{z}^{(\mathbf{x})} := \mathcal{E}(\mathbf{x})$. We also sample a random noise tensor $\mathbf{z}_T^{(\mathbf{d})} \sim \mathcal{N}(0, I)$ as the initial depth latent, which we also treat as an optimizable parameter. We employ the DDIM scheduler [70] for accelerated inference with $T = 50$ steps, adopting the fix for trailing timesteps [20]. Then, at every denoising iteration t, we feed the image latent concatenated with the depth latent $\mathbf{z}_t^{(\mathbf{d})}$ into the U-Net to obtain a noise estimate $\hat{\boldsymbol{\epsilon}}_t := \boldsymbol{\epsilon}_\theta(\mathbf{z}_t^{(\mathbf{d})}, \mathbf{z}^{(\mathbf{x})}, t)$. Instead of immediately completing the current iteration and continuing with timestep $t - 1$, we "preview" the final, denoised depth latent $\mathbf{z}_{0|t}^{(\mathbf{d})}$ by computing a posterior mean estimate using Tweedie's formula [18]:

$$\mathbf{z}_{0|t}^{(\mathbf{d})} = \frac{\mathbf{z}_t^{(\mathbf{d})} - \sqrt{1 - \bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_t}{\sqrt{\bar{\alpha}_t}}, \tag{1}$$

where $\bar{\alpha}_t$ is defined by the noise schedule. After obtaining $\mathbf{z}_{0|t}^{(\mathbf{d})}$, we decode it through the SD decoder $\mathcal{D}$ to produce a predicted clean depth sample $\hat{\mathbf{d}}_t$ in pixel space. This affine-invariant depth is then scaled and shifted using our parameterized coefficients to render it in metric units as $\hat{\mathbf{d}}_t^{(\mathrm{m})} := \hat{\mathbf{d}}_t \cdot \hat{a} + \hat{b}$. We then compute the loss function $\mathcal{L}$ between the sparse depth $\mathbf{c}$ and the metric estimate $\hat{\mathbf{d}}_\mathrm{m}$ at the pixels where $\mathbf{c}$ is valid, resizing the prediction as needed if processing at an intermediate resolution (which is possible in Marigold). For $\mathcal{L}$, we use an equally weighted sum of mean absolute error and mean squared error. Intuitively, this combination penalizes large geometric errors while also encouraging fine-scale adjustments. Given the loss, we compute gradients for both scale and shift coefficients, as well as the latent depth variable $\mathbf{z}_t^{(\mathbf{d})}$, by backpropagating through the decoder $\mathcal{D}$, the Eq. (1), and the U-Net prediction.

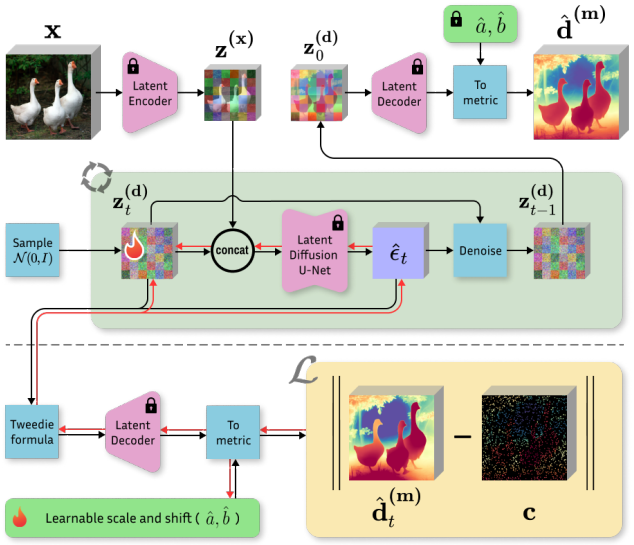

Figure 2. **Overview of Marigold-DC inference scheme for depth completion.** Our method extends the existing Marigold architecture (above the dashed line) by incorporating task-specific guidance (below the line). Starting from the current depth latent variable $\mathbf{z}_t^{(\mathbf{d})}$, our method calculates a "preview" of the final denoised depth map via the Tweedie formula. This preview is then decoded and scaled using the learnable scale parameter $\hat{a}$ and shift $\hat{b}$. We backpropagate the loss (red arrows) between the "preview" and sparse depth and adjust the latent simultaneously with the scale and shift. Finally, we execute a scheduler step to proceed to the next denoising iteration.

We found that scaling the gradient w.r.t. the depth latent proves effective in practice, such that its L_2 norm is proportional to that of the predicted score gradient [84]. We update $\hat{a}$, $\hat{b}$, and $\mathbf{z}_t^{(\mathbf{d})}$ based on their respective gradients using the Adam [34] optimizer, with a learning rate of 0.005 for the affine parameters and 0.05 for the depth latent. The former two always remain positive thanks to the parametrization described in Sec. 3.2.

After this, we perform a scheduler step with the previously computed noise estimate $\hat{\boldsymbol{\epsilon}}$ and proceed to the next denoising iteration. Once the process is completed and the final denoising iteration has been reached, we decode the optimized affine-invariant depth map, apply the scale and shift coefficients, and return a depth map in metric units.

The optimization loop is repeated at each timestep for a total of T iterations, similarly to [9, 19, 22, 38], but with additional variables included in the loop. This proposed update encourages affine-invariant predictions to align with both the image conditioning and sparse depth measurements, as a standard depth completion architecture would, but with the added benefit of a comprehensive representation of the visual world inherited from Stable Diffusion. This knowledge prevents overfitting to noisy depth measurements because the image prior serves as a strong regularizer.

Our end-to-end optimization approach is a variant of guid-

ance methods that use Bayes' rule to adjust the score function with a conditional term [1, 15, 84]. This often involves the diffusion posterior sampling (DPS) [9] approximation of the likelihood, calculated on the expected clean samples. Our choice of a simpler yet effective variant is motivated by three factors: (i) For latent inverse problems, a straightforward extension of DPS lacks the theoretical guarantees of its pixel-space counterpart [62] due to extra nonlinearity introduced by $\mathcal{D}$ and the absence of a one-to-one mapping from latent to pixel space [14]; (ii) We introduce additional variables that require optimization, as the mapping from the base model output to sparse depth is linear but unknown, unlike traditional inverse approaches, which keep the mapping function between prediction and the condition fixed. (iii) Modifying the score gradient during our experiments led to less stable optimization and reduced overall performance. In this way, our streamlined guidance approach is able to handle varying levels of sparsity in the depth guidance.

3.2. Scale and Shift Parameterization

Proper parameterization of scale and shift is crucial to achieve high-quality results. Notably, expressing metric depth through an affine-invariant prediction does not yield a unique decomposition, as multiple affine transformations can represent the same depth structure. However, since Marigold outputs are in the unit interval, given affine parameters $\hat{a}$ and $\hat{b}$, only metric values within the range $[\hat{b}, \hat{a} + \hat{b}]$ can be predicted. If this interval is not expressive enough, the loss will push points in the affine-invariant space toward its boundaries, leading to irrecoverable saturation.

We observe that a least squares fit to the condition $\mathbf{c}$ often produces a range notably smaller than the full set of available depth values, mainly because the initial geometry is not fully accurate, especially at the far plane. The opposite is also true: overshooting the range forces optimization towards a prediction with a distribution of values concentrated in a sub-range, even though Marigold has been trained to utilize the entire range of the decoder. To enable meaningful, non-saturating updates to all optimized components, we parameterize the scale and shift as follows (*min-max* initialization):

$$\hat{a} = \alpha^2 \cdot (\mathbf{c}_{\text{max}} - \mathbf{c}_{\text{min}}) \qquad \hat{b} = \beta^2 \cdot \mathbf{c}_{\text{min}} \tag{2}$$

where $\mathbf{c}_{\text{max}}$ and $\mathbf{c}_{\text{min}}$ are the maximum and minimum depth available as sparse conditioning, and α and β, initialized to one, are the parameters that receive the actual gradient updates. We ablate other initialization methods in Sec. 4.4.

3.3. Ensembling Procedure

Even with anchoring at guidance points, inherent variability in the final depth maps persists with a generative method, depending on the initial noise – similar to the unguided setting. This variability is particularly pronounced in challenging areas, such as reflective surfaces, edges, and distant planes, where depth points are often missing due to sensor limitations or range constraints. We leverage the variability via a simple ensemble method in metric space: after linearly scaling the predictions from multiple individual inferences, we compute the pixel-wise median to produce the final result. This gives us more robust estimates and, as an added benefit, generates an uncertainty map based on the median absolute deviation (MAD) between predictions. We use 10 separate predictions for evaluation, as suggested for Marigold.

4. Experiments

4.1. Evaluation Datasets

We evaluate Marigold-DC in a zero-shot setting on 6 real-world datasets unseen by the base model [33], which was trained exclusively on synthetic data from **Hypersim** [59] and **Virtual KITTI** [4]. The evaluation datasets span both indoor and outdoor scenes, covering various image resolutions, sparse depth densities, acquisition devices, and noise levels. **NYU-Depth V2** [50] consists of indoor scenes captured with an RGB-D Kinect sensor. We use the original test split of 654 samples. Images are downsampled to 320×240 and then center-cropped to 304×228, following established practice [6, 47, 54]. The sparse depth input is generated by sampling 500 random points from the ground truth depth map. **ScanNet** [13] contains room scans collected with a commodity RGB-D sensor. Following the filtering in [69], we select 745 samples from the official 100 scenes for testing. Images are resized to 640×480 to align with the depth resolution, and 500 random points are sampled as sparse guidance. The **VOID** [77] dataset includes synchronized RGB and depth streams of indoor and outdoor scenes at a resolution of 640×480, acquired via active stereo. We utilize all 800 frames from the 8 designated test sequences, and their provided sparse depth maps with three density levels of 150, 500, and 1500 points. **iBims-1** [35] is a high-quality indoor RGB-D dataset captured with a laser scanner, characterized by its low noise level, sharp depth transitions, precise details, and extended depth range up to 50 meters. We employ all 100 available images at 640×480 resolution and sample 1000 random depth points from the intersection of valid pixel masks (invalid, transparent, missing) as per the official evaluation protocol. The **KITTI DC** [75] dataset comprises driving scenes with paired RGB images and sparse LiDAR depth measurements captured at a resolution of 1216×352. The semi-dense ground truth is obtained by temporally accumulating multiple consecutive LiDAR frames with error filtering [21, 75]. We use the original validation split of 1000 samples and remove outliers [11] from the guidance points based on distance from the minimum depth within a local 7×7 patch, as in [2]. **DDAD** [25] is an autonomous driving dataset featuring a 360° multi-camera setup, capturing long-range LiDAR depth up to 250 meters. The official validation

set includes 3950 samples for each camera at 1936×1216 resolution. Following [2, 92], we use only the front-facing view and sample approximately 20% of the available depth measurements as sparse input, applying the same filtering as done for KITTI DC raw LiDAR [11].

4.2. Evaluation Protocol

As mentioned, Marigold-DC is a zero-shot approach that requires no task-specific training, unlike most baselines, which rely on depth-completion checkpoints trained specifically for indoor (NYU-Depth V2) or outdoor (KITTI DC) environments. Our proposition for fair evaluation is to transfer the indoor checkpoints to the indoor benchmarks (ScanNet, iBims-1, and VOID) and the outdoor checkpoints to the outdoor benchmark (DDAD). For evaluation on NYU-Depth V2 and KITTI DC, we avoid expensive retraining of baselines on other datasets and instead reuse the available zero-shot results from the VPP4DC [2] paper, reporting the best of all examined training configurations and leaving the rest blank.

We run inference on each dataset at its original resolution, except for NYU and DDAD, where we resize the images to a 768-pixel longer side while preserving the aspect ratio. This is done for our method and Marigold, due to the image sizes being too small and too large, respectively. In these cases, we resize the output to the original resolution for guidance, enabling processing at the finest level.

Following recent work [2, 69, 92], we report Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as performance metrics.

4.3. Comparison with Other Methods

We compare Marigold-DC to 8 baselines that have achieved strong results in standard evaluation settings on NYU-Depth V2 and KITTI DC, with some also claiming good generalization. NLSPN [54], SpAgNet [12], CompletionFormer [88] and BP-Net [74] are variants of the multi-modal fusion with depth refinement strategy. VPP4DC [2] reformulates the problem as fictitious stereo matching and OGNI-DC [92] optimizes a depth gradient field. DepthLab [46] uses latent diffusion for densification and Prompt Depth Anything [39] adapts the depth foundation model from [82] for completion tasks. To show the effectiveness of our guidance framework, we also compare to vanilla Marigold [33] using only the RGB input with ensemble size 10, followed by (i) a least-squares estimate and (ii) a L1 + L2 optimization for scale and shift based on the sparse depth. These methods are referred to as "Marigold + LS" and "Marigold + optim" below.

As shown in Tab. 1, Marigold-DC outperforms the baselines in most cases and secures the highest overall ranking by a significant margin. Notably, this is achieved despite relying on a frozen base model trained on synthetic datasets for a different task, offering a true plug-and-play solution largely independent of guidance patterns. We argue that the crucial factor is to exploit the rich monocular depth estimation prior provided by the base model, while simultaneously supplying it with sparse evidence. This is in line with our hypothesis that depth completion benefits more from a strong visual prior than from task-specific training.

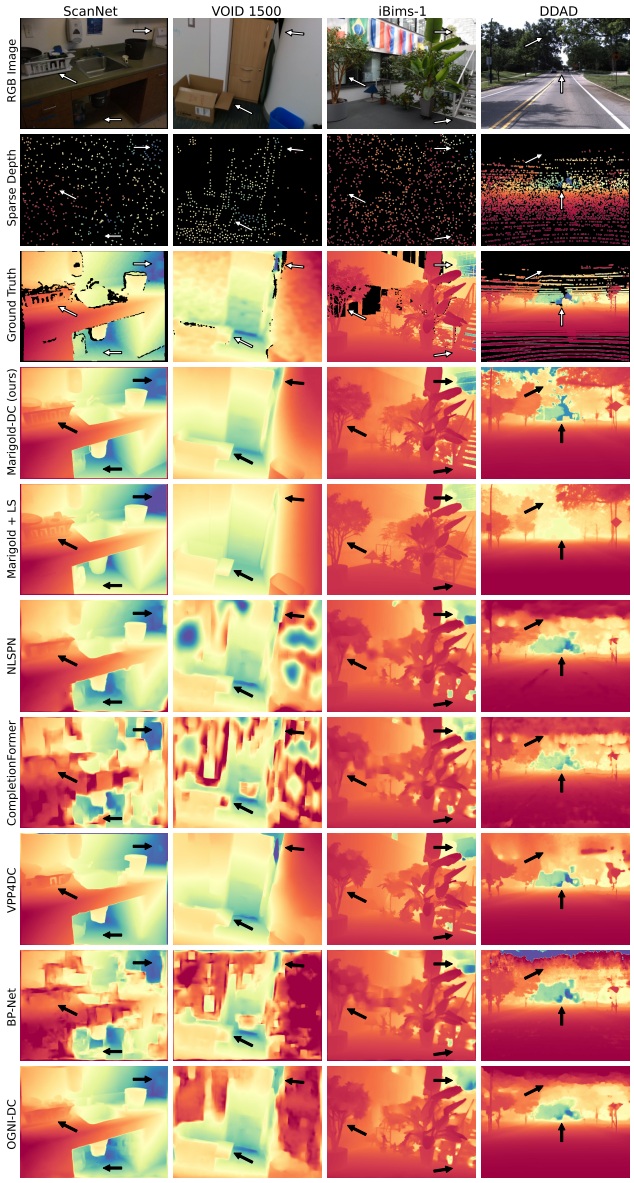

Figure 3. **Qualitative comparison** of benchmarked methods on samples from four datasets. Non-generative methods struggle with dataset-specific biases, such as input resolution lock or variations in guidance sparsity. Marigold-DC demonstrates high-quality metric depth densification with strong generalization. Sampling patterns and noise characteristics vary across datasets. Black regions indicate missing depth values. Arrows suggest key areas of interest.

In Fig. 3, we present a qualitative comparison of samples from several evaluation datasets. In these examples, although the depths predicted with "Marigold + LS" are rich in de-

Table 1. **Quantitative comparison** of Marigold-DC with state-of-the-art depth completion methods on several zero-shot benchmarks. All metrics† are presented in absolute terms; **bold** numbers are best, underscored second best. In most cases, our method outperforms other approaches in both indoor and outdoor scenes, despite not having seen a real depth sample nor being trained for the depth completion task.

Method	ScanNet		IBims-1		VOID 150		VOID 500		VOID 1500		NYU-Depth V2		KITTI DC		DDAD	
	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
NLSPN [54] (ECCV '20)	0.036	0.127	0.049	0.191	0.492	0.963	0.301	0.783	0.210	0.668	0.440	0.716	1.335	2.076	2.498	9.231
SpAgNet [12] (WACV '23)	∅	∅	∅	∅	0.408	0.866	0.326	0.752	0.244	0.706	0.158	0.292	0.518	1.788	4.578	13.236
CompletionFormer [88] (CVPR '23)	0.120	0.232	0.058	0.206	0.487	0.956	0.385	0.821	0.261	0.726	0.186	0.374	0.952	1.935	2.518	9.471
VPP4DC [2] (3DV '24)	0.023	0.076	0.062	0.228	0.245	0.690	0.187	0.582	**0.148**	**0.543**	0.077	0.247	**0.413**	1.609	**1.344**	6.781
BP-Net [74] (CVPR '24)	0.122	0.212	0.078	0.289	0.471	0.936	0.370	0.793	0.270	0.742	⟳	⟳	⟳	⟳	2.270	8.344
OGNI-DC [92] (ECCV '24)	0.029	0.094	0.059	0.186	0.261	0.693	0.198	0.589	0.175	0.593	⟳	⟳	⟳	⟳	1.867	6.876
DepthLab [46] (arXiv preprint '24)	0.051	0.081	0.098	0.198	0.268	0.689	0.223	0.590	0.214	0.602	0.184	0.276	0.921	2.171	4.498	8.379
Prompt Depth Anything [39] (CVPR '25)	0.042	0.079	0.088	0.196	0.248	0.681	0.202	0.589	0.191	0.605	0.110	0.233	0.934	2.803	2.107	7.494
Marigold + optim [33] (CVPR '24)	0.091	0.141	0.167	0.300	0.279	0.687	0.261	0.625	0.261	0.652	0.194	0.309	1.765	3.361	22.872	32.661
Marigold + LS [33] (CVPR '24)	0.083	0.129	0.154	0.286	0.266	0.670	0.243	0.606	0.238	0.628	0.190	0.294	1.709	3.305	8.217	14.728
Ours (w/o ensemble)	0.020	0.063	0.062	0.205	0.201	0.629	0.167	0.546	0.157	0.557	0.057	0.142	0.558	1.676	2.985	7.905
Ours (w/ ensemble)	**0.017**	**0.057**	**0.045**	**0.166**	**0.194**	**0.622**	**0.158**	**0.535**	0.152	0.551	**0.048**	**0.124**	0.434	**1.465**	2.364	**6.449**

† Metrics highlighted in gray are sourced from OGNI-DC [92] and VPP4DC [2], reported with the best training setup. For the others, we evaluated all baselines in zero-shot settings (*cf*. Sec. 4.2). ⟳ indicates that generating zero-shot results would require retraining on an unidentified set of datasets, as the available checkpoints were trained on these benchmarks. ∅ denotes cases where neither training code nor checkpoints are available, making evaluation impossible.

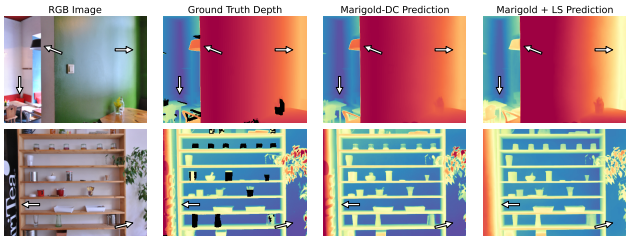

Figure 4. **Comparison between vanilla Marigold and guided predictions** on iBims-1 [35] samples. With guidance from sparse points, the scene geometry is correctly adjusted (first row) and the depth of challenging protruding text can be recovered (second row).

tail, they frequently exhibit layout distortions (Fig. 4) and struggle to position the far plane accurately. In such cases, merely achieving correct depth ordering is insufficient, and the overall scene layout becomes critical.

Performance degradation occurs rapidly in methods that rely on spatial propagation, leading to visible artifacts in the predictions. We hypothesize that this is primarily due to their inability to handle inputs that are sparser than those encountered during training.

Our depth completion approach produces realistic results in regions lacking depth measurements and preserves fine details where other methods resort to coarse interpolation. This makes our method particularly effective in sparse settings with only a few hundred or even a few dozen points.

4.4. Ablation Studies

We analyze the impact of some key design choices on overall performance. Aspects not targeted in each respective experiment are fixed to our reference settings: a learning rate of 0.05 for the depth latent and 0.005 for affine parameters and min-max initialization for scale and shift. We re-use some of the default settings of Marigold [33], namely 50 DDIM denoising steps and an ensemble size of 10. All studies are conducted on a randomly selected subset of 100 samples from the training split of NYU-Depth V2.

Learning rates. Since our method involves test-time optimization, we evaluate the impact of different learning rates used to update the depth latent and affine parameters. For the study, we vary the value of λ_{base} on a log-scale from 0.005 to 0.5, setting the learning rate for the depth latent to this value and the learning rate for scale and shift to $\lambda_{\text{base}}/10$. The results are shown in Fig. 5: we observe a sweet spot around $\lambda_{\text{base}} = 0.05$, with performance degrading in both directions. Lower values result in weak guidance, whereas higher values lead to instability in the optimization process.

Number of denoising steps. We vary the number of denoising steps of the DDIM scheduler [70] and show results in Fig. 5. Unsurprisingly, increasing the number of denoising steps improves the results, though the relative gain diminishes between 25 and 50 steps, with performance saturating beyond that. Inference with less than 25 steps comes at the cost of reduced accuracy and geometric distortions. We anticipate greater speed improvements by having the base model trained on more complex, deeper scenes with enhanced far-plane supervision, since the discrepancy between the initial prediction and the true linear scaling heavily influences convergence speed. Additional speedup by an order of magnitude can be achieved with a TinyVAE [49], using mixed precision and enabling model compilation.

Test-time ensembling. We assess the effectiveness of the proposed test-time ensembling scheme by comparing different ensemble sizes. As shown in Tab. 1 and Fig. 5, a single prediction already yields competitive, often state-of-the-art results. Consistent with standard Marigold, we observe a performance boost through ensembling, with errors generally decreasing as ensemble size increases, albeit with a linear increase in runtime. The relative improvement diminishes for ensemble sizes > 10. This is a hyper-parameter that can be easily adjusted to balance between runtime and performance.

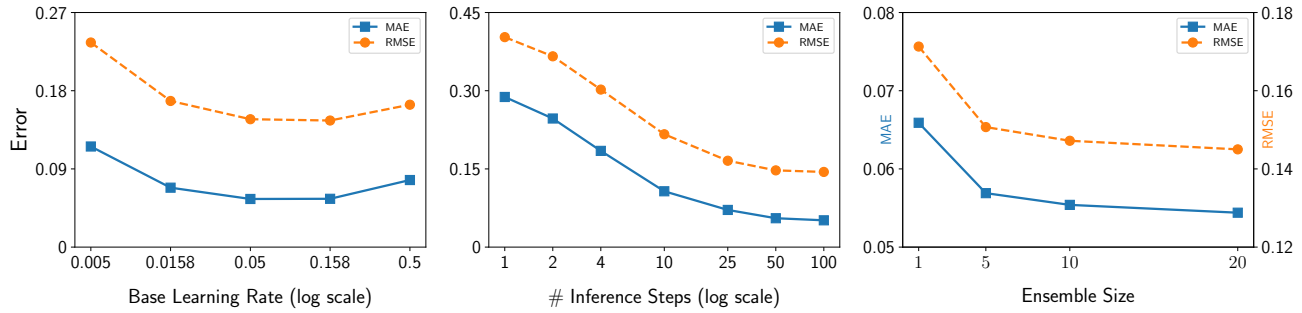

Figure 5. **Ablation of learning rate, number of inference steps, and ensemble size.** Left: Empirically, we identify an optimal learning rate that balances the trade-off between guidance strength and optimization stability. Middle: Performance consistently improves with more denoising iterations, showing saturation beyond 50 steps. Right: A monotonic improvement is seen with increasing ensemble size, diminishing after 10 predictions per sample.

Initialization	NYU-Depth V2 MAE ↓	NYU-Depth V2 RMSE ↓
Least Squares	0.065	0.165
Oracle	0.058	0.153
Extended Min-Max	0.057	0.148
Min-Max	**0.055**	**0.147**

Table 2. **Ablation of scale and shift initialization.** Min-max methods perform similarly to the oracle initialization, whereas least squares performs significantly worse.

Depth Latent Score Function	Depth Latent Direct Optim.	Scale & Shift Direct Optim.	NYU-Depth V2 MAE ↓	NYU-Depth V2 RMSE ↓
✓	✗	✓	0.058	0.150
✗	✓	✓	**0.055**	**0.147**

Table 3. **Ablation of guidance method.** Direct optimization of the depth latent proves more effective than modifying the score function. Affine parameters are optimized separately in both cases.

Scale and shift initialization. Proper initialization of scale and shift is crucial to converge to a solution that aligns well with the sparse depth conditioning, especially when significant adjustments to the initial prediction are necessary. We compare four initialization methods, each derived from the affine-invariant depth map estimate obtained after the first denoising iteration: (1) least-squares fit to align with sparse depth; (2) min-max scaling to match the minimum and maximum values of the sparse input; (3) extended min-max scaling, adjusting the range by 5% beyond the far plane and 5% closer to the near plane, assuming the sparse points provide only a lower bound on scene depth; (4) an oracle using the union of sparse conditioning and ground truth, unavailable in practice but serving as a near-perfect baseline.

As shown in Tab. 2, (extended) min-max and oracle initialization achieve similar performance. Least-squares, however, performs noticeably worse. We acknowledge that the min-max method may struggle when the point distribution is not fully representative of the actual range (*e.g.* the degenerate case of guidance only in a narrow range). However, we are not aware of any depth completion method that does not suffer from this issue. Within our framework, the most robust solution would be an improved model-based initialization or a switch between min-max and least squares whenever the point distribution is considered unrepresentative.

Guidance method. We compare our fully end-to-end optimization framework to an alternative guidance approach that adjusts the score function with a conditional term [1, 15, 84] (*i.e.*, adding the gradient of the likelihood to the predicted noise) while optimizing scale and shift separately. Both approaches are evaluated at their optimal settings, with results reported in Tab. 3. Our optimization-based method outperforms the mixed variant, demonstrating both higher performance and greater stability.

5. Conclusion

We have introduced Marigold-DC, which effectively combines monocular depth estimators' generalization and robustness with the anchoring needed to solve depth completion tasks. By leveraging a pretrained, affine-invariant diffusion-based model and dynamically incorporating sparse depth measurements during inference, Marigold-DC merges rich monocular depth priors with reliable sensor data. Without task-specific training, the method achieves state-of-the-art performance across six zero-shot benchmarks spanning both indoor and outdoor environments. With our work, we aim to inspire further research on methods that prioritize generalization and robustness, to extend depth completion beyond specific training domains and make it more applicable in real-world settings. Our diffusion-based backbone introduces computational overhead due to the iterative nature of denoising diffusion models and its ensembling process. Future research directions include reducing inference time and supporting alternative guidance cues.

References

[1] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 843–852, 2023. 2, 3, 5, 8

[2] Luca Bartolomei, Matteo Poggi, Andrea Conti, Fabio Tosi, and Stefano Mattoccia. Revisiting depth completion from a stereo matching perspective for cross-domain generalization. In *2024 International Conference on 3D Vision (3DV)*, pages 1360–1370. IEEE, 2024. 3, 5, 6, 7

[3] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 2

[4] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*, 2020. 5

[5] Yun Chen, Bin Yang, Ming Liang, and Raquel Urtasun. Learning joint 2d-3d representations for depth completion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10022–10031, 2019. 2

[6] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *European Conference on Computer Vision (ECCV)*, pages 103–119, 2018. 2, 5

[7] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. CSPN++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *AAAI Conference on Artificial Intelligence*, 2019. 2

[8] Nathaniel Chodosh, Chaoyang Wang, and Simon Lucey. Deep convolutional compressed sensing for lidar depth completion. In *Asian Conference on Computer Vision (ACCV)*, pages 499–513, 2019. 2

[9] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023. 3, 4, 5

[10] Hyungjin Chung, Jong Chul Ye, Peyman Milanfar, and Mauricio Delbracio. Prompt-tuning latent diffusion models for inverse problems. In *Proceedings of the 41st International Conference on Machine Learning*, pages 8941–8967. PMLR, 2024. 3

[11] Andrea Conti, Matteo Poggi, Filippo Aleotti, and Stefano Mattoccia. Unsupervised confidence for LiDAR depth maps and applications. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022. IROS. 5, 6

[12] Andrea Conti, Matteo Poggi, and Stefano Mattoccia. Sparsity agnostic depth completion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5871–5880, 2023. 3, 6, 7

[13] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[14] Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G. Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems, 2024. 3, 5

[15] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021. 2, 3, 5, 8

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021. 2

[17] Yiqun Duan, Xianda Guo, and Zheng Zhu. DiffusionDepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023. 3

[18] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 4

[19] Hugging Face. The hugging face diffusion models course. https://huggingface.co/learn, 2022. Online, accessed on 2024-09-01. 4

[20] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. *arXiv preprint arXiv:2409.11355*, 2024. 4

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5

[22] Asya Grechka, Guillaume Couairon, and Matthieu Cord. Gradpaint: Gradient-guided inpainting with diffusion models. *arXiv preprint arXiv:2309.09614*, 2023. 4

[23] Jakub Gregorek and Lazaros Nalpantidis. SteeredMarigold: Steering diffusion towards depth completion of largely incomplete depth maps. *arXiv preprint arXiv:2409.10202*, 2024. 3

[24] Ming Gui, Johannes S. Fischer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*, 2024. 3

[25] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 2, 3

[29] Lee Hyoseok, Kyeong Seon Kim, Kwon Byung-Ki, and Tae-Hyun Oh. Zero-shot depth completion via test-time alignment with affine-invariant depth prior. In *AAAI Conference on Artificial Intelligence*, 2025. 3

[30] Saif Imran, Yunfei Long, Xiaoming Liu, and Daniel Morris. Depth coefficients for depth completion. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[31] Saif Imran, Xiaoming Liu, and Daniel Morris. Depth completion with twin-surface extrapolation at occlusion boundaries. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[32] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. DDP: Diffusion model for dense visual prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21741–21752, 2023. 3

[33] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9492–9502, 2024. 2, 3, 5, 6, 7

[34] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 4

[35] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 331–348, 2019. 5, 7

[36] Jason Ku, Ali Harakeh, and Steven L Waslander. In defense of classical image processing: Fast depth completion on the cpu. In *Conference on Computer and Robot Vision (CRV)*, pages 16–22, 2018. 2

[37] Ang Li, Zejian Yuan, Yonggen Ling, Wanchao Chi, Chong Zhang, et al. A multi-scale guided cascade hourglass network for depth completion. In *IEEE Conference on Applications of Computer Vision (WACV)*, pages 32–40, 2020. 2

[38] Haotian Lin, Yixiao Wang, Mingxiao Huo, Chensheng Peng, Zhiyuan Liu, and Masayoshi Tomizuka. Joint pedestrian trajectory prediction through posterior sampling. *arXiv preprint arXiv:2404.00237*, 2024. 4

[39] Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation, 2025. 3, 6, 7

[40] Yuankai Lin, Hua Yang, Tao Cheng, Wending Zhou, and Zhouping Yin. DySPN: Learning dynamic affinity for image-guided depth completion. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. 2

[41] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[42] Lina Liu, Xibin Song, Xiaoyang Lyu, Junwei Diao, Mengmeng Wang, Yong Liu, and Liangjun Zhang. FCFR-Net: Feature fusion based coarse-to-fine residual learning for depth completion. In *AAAI Conference on Artificial Intelligence*, 2020. 2

[43] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. In *Advances in Neural Information Processing Systems*, 2017. 2

[44] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[46] Zhiheng Liu, Ka Leong Cheng, Qiuyu Wang, Shuzhe Wang, Hao Ouyang, Bin Tan, Kai Zhu, Yujun Shen, Qifeng Chen, and Ping Luo. DepthLab: From partial to complete. *arXiv preprint arXiv:2412.18153*, 2024. 3, 6, 7

[47] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4796–4803, 2018. 2, 5

[48] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from LiDAR and monocular camera. *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3288–3295, 2018. 2

[49] Madebyollin. TAESD. https://github.com/madebyollin/taesd, 2025. Accessed: 2025-03-07. 7

[50] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision (ECCV)*, 2012. 5

[51] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021. 3

[52] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[53] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[54] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 5, 6, 7

[55] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. DeepLiDAR: Deep surface normal guided depth prediction for outdoor scene from sparse LiDAR data and single color image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen

Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 3

[57] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2020. 2

[58] Kyeongha Rho, Jinsung Ha, and Youngjung Kim. GuideFormer: Transformers for image guided depth completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6250–6259, 2022. 2, 3

[59] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV)*, 2021. 5

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3

[61] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 3

[62] Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 5

[63] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 3

[64] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH*, pages 1–10, 2022. 3

[65] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. 2, 3

[66] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J. Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *arXiv preprint arXiv:2306.01923*, 2023. 3

[67] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 3

[68] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294, 2022. 3

[69] Yunxiao Shi, Manish Kumar Singh, Hong Cai, and Fatih Porikli. DeCoTR: Enhancing depth completion with 2d and 3d attentions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10736–10746, 2024. 2, 5, 6

[70] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4, 7

[71] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019. 3

[72] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3

[73] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30:1116–1129, 2020. 2

[74] Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9763–9772, 2024. 2, 6, 7

[75] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant CNNs. In *International Conference on 3D Vision (3DV)*, 2017. 2, 5

[76] Yufei Wang, Bo Li, Ge Zhang, Qi Liu, Gao Tao, and Yuchao Dai. LRRU: Long-short range recurrent updating networks for depth completion. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[77] Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters*, 5(2):1899–1906, 2020. 5

[78] Xin Xiong, Haipeng Xiong, Ke Xian, Chen Zhao, Zhiguo Cao, and Xin Li. Sparse-to-dense depth completion revisited: Sampling strategy and graph construction. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[79] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. Depth completion from sparse LiDAR data with depth-normal constraints. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2811–2820, 2019. 2

[80] Zhiqiang Yan, Yuankai Lin, Kun Wang, Yupeng Zheng, Yufei Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Tri-perspective view decomposition for geometry-aware depth completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4874–4884, 2024. 2

[81] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the

power of large-scale unlabeled data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10371–10381, 2024. 2

[82] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 3, 6

[83] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020. 2

[84] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. *arXiv preprint arXiv:2406.09394*, 2024. 4, 5, 8

[85] Zhu Yu, Zehua Sheng, Zili Zhou, Lun Luo, Si-Yuan Cao, Hong Gu, Huaqi Zhang, and Hui-Liang Shen. Aggregating feature point cloud for depth completion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8698–8709, 2023. 2, 3

[86] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 2, 3

[87] Yinda Zhang and Thomas A. Funkhouser. Deep depth completion of a single rgb-d image. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–185, 2018. 2

[88] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. CompletionFormer: Depth completion with convolutions and vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18527–18536, 2023. 2, 3, 6, 7

[89] Shanshan Zhao, Mingming Gong, Huan Fu, and Dacheng Tao. Adaptive context-aware multi-modal network for depth completion. *IEEE Transactions on Image Processing (TIP)*, 30:5264–5276, 2021. 2

[90] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv:2303.02153*, 2023. 3

[91] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (NTIRE)*, 2023. 3

[92] Yiming Zuo and Jia Deng. Ogni-dc: Robust depth completion with optimization-guided neural iterations. In *European Conference on Computer Vision (ECCV)*, pages 78–95, 2024. 3, 6, 7