This WACV 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.

DPBridge: Latent Diffusion Bridge for Dense Prediction

Haorui Ji Taojun Lin Hongdong Li
The Australian National University
{haorui.ji, taojun.lin, hongdong.li}@anu.edu.au

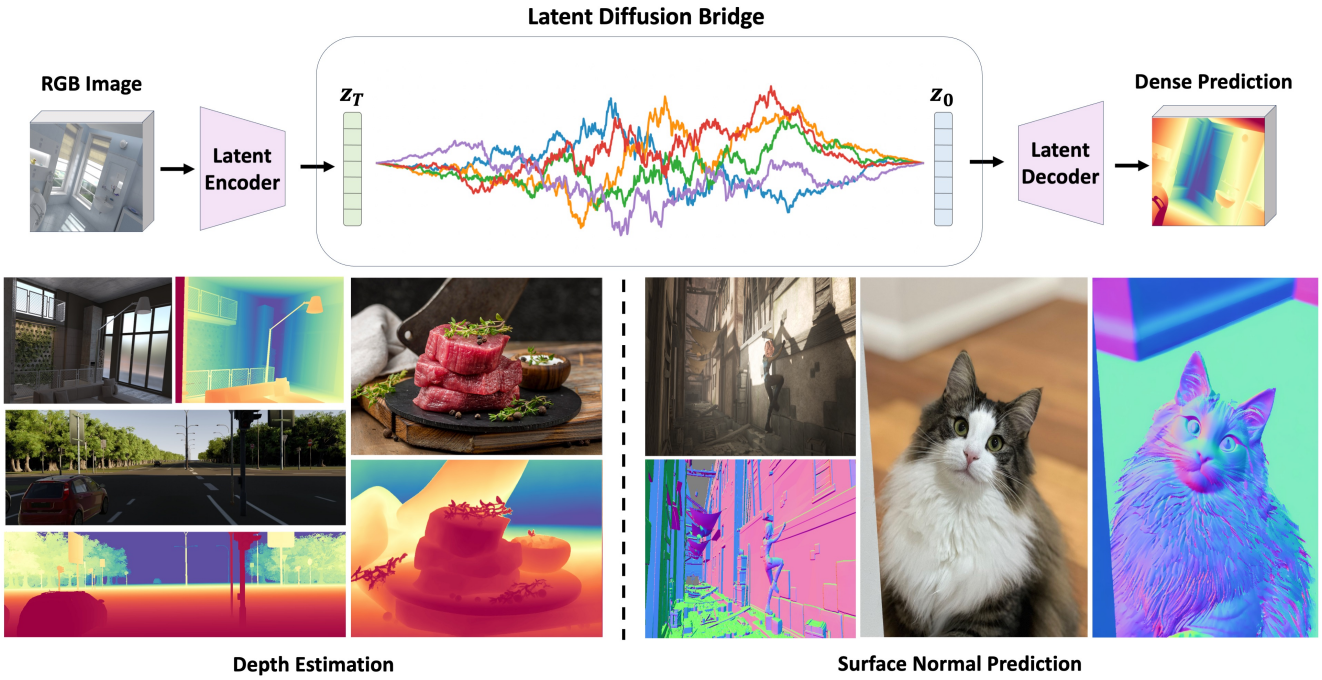


Figure 1. We present **DPBridge** to tackle general dense prediction tasks by establishing conditional probability transport between RGB images and corresponding dense maps in the latent space based on a tractable diffusion bridge process.

Abstract

Diffusion models demonstrate remarkable capabilities in capturing complex data distributions and have achieved compelling results in many generative tasks. While they have recently been extended to dense prediction tasks such as depth estimation and surface normal prediction, their full potential in this area remains underexplored. As target signal maps and input images are pixel-wise aligned, the conventional noise-to-data generation paradigm is inefficient, and input images can serve as a more informative prior compared to pure noise. Diffusion bridge models, which support data-to-data generation between two general data distributions, offer a promising alternative, but they typically fail to exploit the rich visual priors embedded in large pretrained foundation models. To address these limitations, we integrate diffusion bridge formulation with structured visual priors and introduce DPBridge, the first latent diffusion bridge framework for dense prediction tasks. To resolve the incompatibility between diffusion bridge models and pretrained diffusion backbones, we propose (1) a tractable reverse transition kernel for the diffusion bridge process, enabling maximum likelihood training scheme; (2) finetuning strategies including distribution-aligned normalization and image consistency loss. Experiments across extensive benchmarks validate that our method consistently achieves superior performance, demonstrating its effectiveness and generalization capability under different scenarios.

1. Introduction

Dense prediction refers to a class of problems in which the goal is to assign a prediction value for every pixel in the input image. It encompasses a variety of fundamental tasks, including depth estimation [14, 37, 52], surface normal pre-

diction [2, 3, 21], semantic segmentation [8, 47, 49], optical flow [22, 45, 50], etc.

Dense prediction has traditionally relied on discriminative models [2, 37, 60], which map RGB inputs to dense signal maps directly via feed-forward manners. While these methods offer high inference speed, they usually struggle in ambiguous scenarios-such as occluded regions or dynamic surface normals of glossy objects. This limitation is largely due to the lack of uncertainty modeling. Recently, generative models such as diffusions have been applied to mitigate these drawbacks by framing dense prediction as a conditional generation task. Leveraging the ability of modeling complex data distributions, diffusions serve as robust priors and deliver promising performance as different tasks [15, 16, 27]. Nevertheless, diffusion-based generative approaches suffer from inherent limitations rooted in their core mechanism: their denoising process initiates from Gaussian noise—an initial prior that contains minimal information about the target dense signal distribution. In such cases, the distributional gap between initial Gaussian prior and target distribution remains large, necessitating more sampling steps and often reliance on manually designed post-hoc guidance [9, 10, 43] or customized generation pipeline [33]. This not only increases computational overhead but can also degrade performance and prolong convergence.

Instead of starting from noise, a more effective alternative is to directly initiate the generative process from the input. Since RGB images are pixel-wise aligned with their corresponding dense signal maps, they shall act as highly informative priors. This motivates the exploration of diffusion bridge models [13, 31, 32, 35], another line of computational frameworks based on stochastic processes with fixed start and end states, designed to model probability flows between two given distributions. We argue that diffusion bridge models are more suitable for dense prediction tasks, as their data-to-data generation paradigm naturally leverages the structural correspondence between inputs and outputs, thereby mitigating inefficiencies inherent in the conventional noise-to-data scheme. Despite theoretical feasibility, diffusion bridge-based methods [29, 30, 61] have not yet demonstrated competitive performance compared with standard diffusion counterparts. This performance gap primarily arises because most bridge models are trained from scratch via score matching objectives, making them incompatible with the widely adopted pretrained diffusion backbones. Consequently, these methods fail to leverage the embedded semantic and spatial prior knowledge, constraining their expressive power and generalization capability in dense prediction scenarios.

In this paper, we propose **D**ense **P**rediction **Bridge** (**DPBridge**), the first latent diffusion bridge model tailored for general dense prediction tasks by integrating the strengths of diffusion bridge modeling with visual priors from pretrained backbones. Specifically, we derive a fully tractable bridge process to connect the latent representations of RGB images and their corresponding dense signal maps, and formulate the generation process as a sequence of transformations that progressively synthesize the target maps from input data. Furthermore, we introduce finetuning strategies to better adapt pretrained diffusion backbones for DPBridge construction. Leveraging rich visual priors not only enhances perceptual performance and generalization ability, but also significantly accelerates training convergence. Figure 1 provides a high-level overview of the DPBridge framework and illustrates its effectiveness across two representative tasks: depth estimation and surface normal prediction. These results demonstrate that our method is both theoretically feasible and practically adaptable to diverse scenarios.

Contributions of this paper can be summarized as follows:

- We propose DPBridge, a generative framework to address general dense prediction tasks by formulating them as instances of latent diffusion bridges.
- We derive a tractable reverse transition kernel for diffusion bridge process, enabling maximum likelihood training scheme. This is compatible with pretrained backbones, thus improving both training stability and the exploitation of visual prior knowledge.
- We introduce finetuning strategies to improve finetuning performance, including a distribution-aligned normalization technique that mitigates distributional discrepancies between the bridge and standard diffusion processes, and an auxiliary image consistency loss to better preserve fine-grained spatial details.
- Our method shows competitive performance across different tasks in terms of both quantitative metrics and visual inspection.

2. Related Works

Dense Prediction. Common solutions to dense predictions are categorized into discriminative and generative ones. Discriminative methods directly map input data to structured outputs in a feed-forward manner. [37] introduces transformer architecture and fusion mechanism to better capture both local and global context features. [2] incorporates per-pixel ray directions and models the relative rotations between neighboring pixels. Recently, diffusion-based generative methods have been applied in this field and have shown remarkable performance. [27] finetunes pretrained diffusion backbone, leveraging the rich visual knowledge within to enhance depth estimation. [15] takes advantage of diffusion priors to jointly estimate depth and surface normals from single image, enhancing detail preservation by decoupling complex data distributions into distinct sub-distributions. [18] exploits flow matching to facilitate the distribution shift between images and depth maps. Our work follows the conventional generative pipeline, but instead of starting from

pure noise, we establish direct connections between input image and target signal map distributions, and model conditional probability transport in an iterative denoising manner.

Diffusion and Bridge Models. Diffusion models [19, 44] are a family of generative models that excel at capturing complex data distributions through a sequence of forward and reverse processes. They have driven the advancement of modern generative systems, and have consistently achieved state-of-the-art performance across various domains, such as image and video synthesis [20, 39], controlled editing [4, 58], inverse problem solving [9, 43], 3D assets generation [51, 59], and human-centered applications [23–25].

In recent years, diffusion bridge models have gained growing interest in generative modeling [13, 31, 32, 35, 61]. They offer a novel perspective for translating between distributions and enable a more effective data-to-data generation paradigm. Recent works have explored different types of bridge processes and their applications. [29, 30, 56] develop Schrödinger bridge, Brownian bridge, and Ornstein-Uhlenbeck bridge frameworks, applying them to image restoration tasks such as inpainting, super-resolution. [11] proposes a modified inference procedure that enforces data consistency. [7] introduces bridge models to text-to-speech systems, enhancing synthesis quality.

To the best of our knowledge, no prior work has explored the use of bridge models for dense prediction tasks to facilitate distribution transport between images and dense signal maps, despite their structural closeness. Our work aims to fill this research gap. Conceptually, DPBridge is most related to DDBM [61], but the two differ in several key aspects: DDBM is trained from scratch using a score matching objective, whereas DPBridge is finetuned from large pretrained diffusion backbones using a maximum likelihood objective. In addition, we introduce strategies for efficient finetuning, enabling better use of visual priors and significantly improving performance on dense prediction settings.

3. Preliminary

3.1. Diffusion Process

A diffusion process $\mathbb{Q} = \{\mathbf{z}_t \in \mathbb{R}^d : t \in [0, T]\}$ on space $\mathbb{R}^d$ follows the stochastic differential equation of form

$$\mathrm{d}\mathbf{z}_t = f(\mathbf{z}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t \tag{1}$$

where $\mathbf{w}_t$ is a standard Wiener process, $f : \mathbb{R}^d \times [0, T] \to \mathbb{R}$ and $g : [0, T] \to \mathbb{R}$ are the drift and diffusion coefficients that describe the evolution of this stochastic process. The transition kernel is specified as $q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\alpha_t\mathbf{z}_0, \eta_t^2\mathbf{I})$ and the noise schedule comprising differentiable functions α_t, η_t depends on the design choice of drift and diffusion coefficients. It is also common to assess the diffusion process in terms of the signal-to-noise ratio (SNR), defined as $\mathrm{SNR}_t = \alpha_t^2/\eta_t^2$.

3.2. Diffusion Bridge Process

Diffusion bridge process is a special type of diffusion process that is conditioned to end at a specified terminal point at timestep T. Its dual fixed-endpoint characteristics enable a bidirectional distribution transport. Its construction can be derived from a reference diffusion process conditioned on terminal constraints. We can use Doob's h-transform technique [34, 40] to modify a reference process $\mathbb{Q}$ defined in Equation (1) to ensure that the event $\mathbf{z}_T = \mathbf{x}$ happens, where $\mathbf{x}$ is a predefined value. The conditioned process $\mathbb{Q}^{\mathbf{x}}(\cdot) := \mathbb{Q}(\cdot|\mathbf{z}_T = \mathbf{x})$ thus follows the law of

$$\mathrm{d}\mathbf{z}_t = \left[f(\mathbf{z}_t, t) + g^2(t)h(\mathbf{z}_t, t, \mathbf{x}, T)\right]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t \tag{2}$$

where $h(\mathbf{z}_t, t, \mathbf{x}, T) = \nabla_{\mathbf{z}_t} \log q(\mathbf{z}_T = \mathbf{x}|\mathbf{z}_t)$. Intuitively, the additional drift term $h(\mathbf{z}_t, t, \mathbf{x}, T)$ plays the role of steering $\mathbf{z}_t$ towards the target point $\mathbf{z}_T = \mathbf{x}$, and it is ensured that $\mathbb{Q}^{\mathbf{x}}$ invariably passes $\mathbf{x}$ at timestep T with probability 1 for any initial point $\mathbf{z}_0$.

The transition kernel of $\mathbb{Q}^{\mathbf{x}}$, given by $q(\mathbf{z}_t|\mathbf{z}_0, \mathbf{z}_T)$, can be subsequently derived through Kolmogorov equation:

$$\begin{aligned} q(\mathbf{z}_t|\mathbf{z}_0, \mathbf{z}_T) &= \mathcal{N}(m_t\mathbf{z}_0 + n_t\mathbf{z}_T, \sigma_t^2\mathbf{I}) \\ m_t = \alpha_t\left(1 - \frac{\mathrm{SNR}_T}{\mathrm{SNR}_t}\right), &\quad n_t = \frac{\mathrm{SNR}_T}{\mathrm{SNR}_t}\frac{\alpha_t}{\alpha_T}, \\ \sigma_t &= \eta_t\sqrt{1 - \frac{\mathrm{SNR}_T}{\mathrm{SNR}_t}} \end{aligned} \tag{3}$$

where $\mathrm{SNR}_t, \alpha_t, \sigma_t^2$ are all defined in the reference process $\mathbb{Q}$. In Equation (3), m_t and n_t are coefficients that enforce the bridge's dual endpoints: at $t = 0, m_0 = 1, n_0 = 0 \to z_t = z_0$; at $t = T, m_T = 0, n_T = 1 \to z_t = z_T$.

4. Method

4.1. Problem Formulation

We formulate different dense prediction tasks as instances of image-conditioned generation problems and propose DPBridge to model the conditional distributions $p(\mathbf{y}|\mathbf{x})$, where $\mathbf{y} \in \mathbb{R}^{H\times W\times C}$ denotes the target dense signal map (e.g. depth, surface normal, segmentation map), and $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ is the input RGB image. In the following sections, we'll elaborate the details of the DPBridge framework, which is constructed based on pretrained diffusion backbone through proper finetuning.

4.2. Dense Prediction Bridge

Given paired training data, a straightforward approach is to construct the diffusion bridge model from scratch, following the formulation in Sec. 3.2. However, this approach is extremely resource-demanding, making it hard to implement with limited computation power. What's more, it cannot to harness rich visual prior embedded in the foundation models

pretrained with large-scale data, which have been shown to significantly enhance both perceptual quality and generalization capability in dense prediction tasks. To address these limitations and fully exploit the visual prior, our DP-Bridge is built upon pretrained diffusion backbones through finetuning. We explore several strategies to handle the misalignment between bridge process and standard diffusion process, facilitating more efficient knowledge transfer. An overview of our training pipeline is provided in Figure 2, followed by a detailed explanation of each component.

4.3. Latent Space Implementation

We follow common practice to perform diffusion steps in a compressed yet perceptually aligned latent space, provided by an independently trained variational autoencoder (VAE). Specifically, both the dense signal map and RGB image are encoded using the VAE encoder, yielding latent representations $\mathbf{z}_0 = \mathcal{E}(\mathbf{y})$ and $\mathbf{z}_T = \mathcal{E}(\mathbf{x})$, which serve as the initial and terminal points of the diffusion bridge process. In this way the diffusion bridge process is operated entirely in the latent space. To recover the original signal map, we use the VAE decoder to decode the predicted latent $\hat{\mathbf{y}} = \mathcal{D}(\hat{\mathbf{z}}_0)$. For clarity, all subsequent derivations in this paper are conducted in latent space.

4.4. Training Objective

Maximum Likelihood Loss. A widely adopted approach for learning the reverse process associated with Equation (2) follows the theoretical framework of [1], where a neural network is trained to approximate the score function [61]. In contrast, we observe that in the diffusion bridge setting, the reverse transition kernel admits a tractable analytical form. This insight enables us to adopt a maximum likelihood training paradigm in place of score matching. Since widely used latent diffusion models like Stable Diffusion [39] are pretrained under a likelihood-based loss function, finetuning with a consistent objective is more compatible, which helps preserve prior knowledge and improves training stability.

Specifically, the training process is performed by optimizing the Evidence Lower Bound (ELBO). Given the encoded training pair $\left(\mathbf{z}_0, \mathbf{z}_T\right)$, the expectation of data log-likelihood possesses the following ELBO formulation:

$$\begin{aligned} \textit{ELBO} =& \mathbb{E}_{p(\mathbf{z}_0)}\Big[\mathbb{E}_{p(\mathbf{z}_1|\mathbf{z}_0)}\big[\log p_\theta(\mathbf{z}_0 \mid \mathbf{z}_1, \mathbf{z}_T)\big] - \\ & \sum_{t=2}^{T} \mathbb{E}_{p(\mathbf{z}_t|\mathbf{z}_0)}\big[\mathbb{D}_{KL}\big(q(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}_0, \mathbf{z}_T) \,||\, p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}_T)\big)\big]\Big] \end{aligned} \tag{4}$$

where $q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0, \mathbf{z}_T)$ and $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_T)$ are the transition kernel of true reverse process and our modeled generative process. $\mathbb{D}_{KL}$ is the KL divergence. Utilizing Bayes' theorem and Markov chain property, we can obtain:

$$\begin{aligned} q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0, \mathbf{z}_T) &= \frac{q(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{z}_T) q(\mathbf{z}_{t-1}|\mathbf{z}_0, \mathbf{z}_T)}{q(\mathbf{z}_t|\mathbf{z}_0, \mathbf{z}_T)} \\ &= \mathcal{N}(\tilde{\boldsymbol{\mu}}_t(\mathbf{z}_t, \mathbf{z}_0, \mathbf{z}_T), \tilde{\sigma}_t^2 \mathbf{I}) \end{aligned} \tag{5}$$

For simplicity, we define the following notations: $a_t = m_t / m_{t-1}, b_t = n_t - a_t \cdot n_{t-1}, \delta_t^2 = \sigma_t^2 - a_t^2 \cdot \sigma_{t-1}^2$. Based on Equation (3), we further derive the mean value term:

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_t(\mathbf{z}_t, \mathbf{z}_0, \mathbf{z}_T) &= k_1 \mathbf{z}_t + k_2 \mathbf{z}_0 + k_3 \mathbf{z}_T \\ k_1 = \frac{\sigma_{t-1}^2}{\sigma_t^2} \cdot a_t, \quad & k_2 = \frac{\delta_t^2}{\sigma_t^2} \cdot m_{t-1}, \\ k_3 = \frac{\delta_t^2}{\sigma_t^2} \cdot n_{t-1} &- \frac{\sigma_{t-1}^2}{\sigma_t^2} \cdot a_t \cdot b_t, \end{aligned} \tag{6}$$

and the variance term:

$$\tilde{\sigma}_t^2 = \frac{\sigma_{t-1}^2}{\sigma_t^2} \cdot \delta_t^2 \tag{7}$$

Assuming $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_T)$ also follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{z}_t, \mathbf{z}_T, t), \boldsymbol{\Sigma}_\theta(\mathbf{z}_t, \mathbf{z}_T, t))$ with fixed variance $\boldsymbol{\Sigma}_\theta(\mathbf{z}_t, \mathbf{z}_T, t) = \tilde{\sigma}_t^2$ for simplicity of derivation. Following the strategy of DDPM [19], we reparameterize the model $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{z}_T, t)$ to predict noise instead of posterior mean. Under this formulation, the ELBO objective is equivalent to:

$$\begin{aligned} \mathcal{L}_{elbo} &= \mathbb{E}_{t, \mathbf{z}_0, \mathbf{z}_t, \mathbf{z}_T}\left[\frac{1}{2\tilde{\sigma}_t^2} \left\|\tilde{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_\theta(\mathbf{z}_t, \mathbf{z}_T, t)\right\|^2\right] \\ &\cong \mathbb{E}_{t, \mathbf{z}_0, \mathbf{z}_t, \mathbf{z}_T}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{z}_T, t)\right\|^2\right] \end{aligned} \tag{8}$$

While Equation (8)'s MSE-like form resembles DDPM's noise loss, its derivation relies on the analytical reverse kernel of the bridge process. The inclusion of z_T is not "standard conditional generation" but a requirement for enforcing the bridge's dual-endpoint constraint. For more derivation details, please refer to the supplementary materials.

Image Consistency Loss. In addition to the latent-space likelihood loss, we also incorporate an image-level consistency loss, where we first decode the latent prediction back to image space, and then evaluate its consistency with the ground truth dense map using a simple MSE.

$$\mathcal{L}_{ic} = \left\|\mathcal{D}(\hat{\mathbf{z}}_0) - \mathbf{y})\right\|^2 \tag{9}$$

This additional supervision addresses the loss of high-frequency structures caused by latent-space compression, which can impair the model's ability to preserve fine-grained details. By introducing this auxiliary image-space loss, we encourage the model to retain spatial information critical for accurate per-pixel prediction. This dual-loss strategy combines the efficiency of latent-space modeling with the fidelity of image-space supervision, leading to improved performance. The final training objective is given by:

$$\mathcal{L} = \omega_1 \mathcal{L}_{elbo} + \omega_2 \mathcal{L}_{ic} \tag{10}$$

where ω_1, ω_2 are hyperparameters to balance the loss.

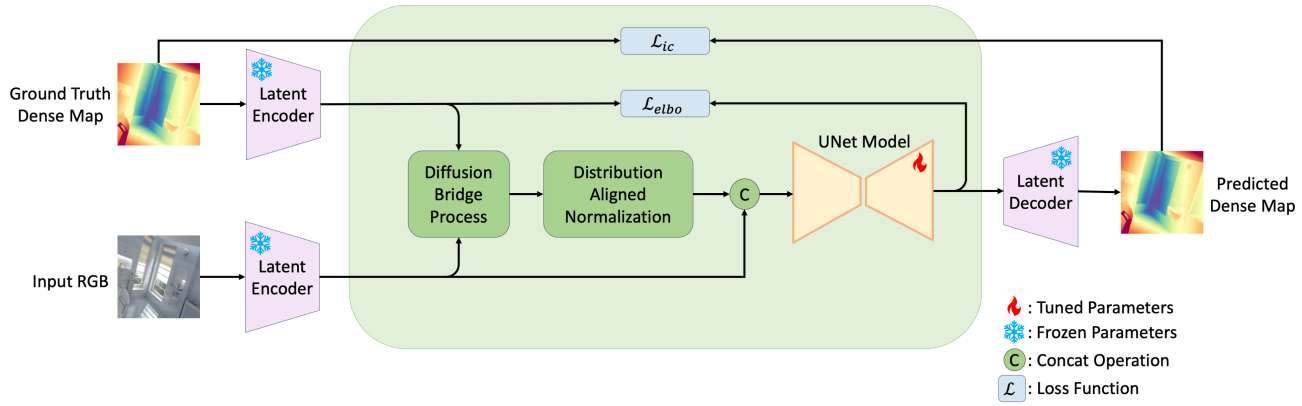


Figure 2. Overview of our training pipeline. We derive the maximum likelihood training objective of diffusion bridge process, propose a distribution-aligned normalization strategy, and assist the training with an auxiliary image consistency loss.

4.5. Distribution Alignment Normalization

A key challenge during finetuning is the misalignment between the marginal distribution $q(\mathbf{z}_t)$ of intermediate samples in DPBridge and that of the diffusion backbone, due to the difference in the forward transition kernels. This misalignment weakens the model's denoising ability and hinders effective knowledge transfer. To address this issue, we propose a distribution alignment normalization technique that transforms intermediate latents computed from our bridge process Equation (3) to match the distributional form of a standard diffusion process.

$$\mathbf{z}'_t = \frac{\mathbf{z}_t - m_t \mathbf{z}_T}{\sqrt{n_t^2 + \sigma_t^2}} = \frac{n_t}{\sqrt{n_t^2 + \sigma_t^2}} \mathbf{z}_0 + \frac{\sigma_t}{\sqrt{n_t^2 + \sigma_t^2}} \boldsymbol{\epsilon}_t \quad (11)$$

Since the square sum of coefficients is equal to 1, the marginal distribution of the normalized bridge process $\{\mathbf{z}'_t : t \in [0, T]\}$ can thus align with that of the standard diffusion process, while preserving bridge-specific dynamics. Therefore, by feeding the normalized sample $\mathbf{z}'_t$ in the network $\boldsymbol{\epsilon}_\theta$, we can fully exploit the denoising capability from the pretrained diffusion backbone. We note that a similar strategy was independently proposed in FrameBridge [48], where it is applied in the context of image-to-video generation. Our method, developed concurrently, is specifically designed to facilitate adaptation in dense prediction tasks under the DPBridge framework. An overview of the full training and inference procedures is summarized in the supplementary materials.

5. Experiments

We conduct evaluation primarily on two representative dense prediction tasks: depth estimation and surface normal prediction. We begin with a brief overview of experimental settings, followed by both quantitative and qualitative analyses. Next, we present ablation studies to investigate the contribution of each design component. Implementation details are presented in the supplementary materials for reference.

5.1. Experiments Setup

Datasets. The basic setting for depth estimation and surface normal prediction experiments are similar. We use two synthetic datasets, HyperSim [38] and Virtual KITTI [6], for training, which cover both indoor and outdoor scenes with high-quality ground-truth annotations. To comprehensively evaluate the model performance and demonstrate its generalization capability, we conduct zero-shot evaluations on real-world datasets. For depth estimation, this includes NYUv2 [42], KITTI [17], ScanNet [12], ETH3D [41] and DIODE [46]. For surface normal prediction, we additionally incorporate iBims [28] and Sintel [5] datasets. For fair comparison, we use the same validation split as in previous methods [27].

Evaluation Metrics. For depth estimation, we follow the affine-invariant evaluation protocol, where we first align the prediction results to the ground truth with least squares fitting, and then use Absolute Mean Relative Error (AbsRel) and δ_1 accuracy to assess the performance. Denote the ground truth depth $\mathbf{d}$, predicted depth $\hat{\mathbf{d}}$, and the number of valid pixels N, the AbsRel is calculated as $\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{d}_i - \hat{\mathbf{d}}_i\|/\mathbf{d}_i$, and δ_1 accuracy is measured using the ratio of pixels satisfying $\max(\mathbf{d}_i/\hat{\mathbf{d}}_i, \hat{\mathbf{d}}_i/\mathbf{d}_i) < 1.25$. For the metrics of surface normal task, we report the commonly used mean angular error between the ground-truth normal vectors and the prediction results, as well as the percentage of pixels with an angular error below 11.25 degrees.

Method	Model Type			NYUv2		KITTI		ETH3D		ScanNet		DIODE		Avg.
	FF	DM	BM	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	Rank
DiverseDepth [54]	✓			11.7	87.5	19.0	70.4	22.8	69.4	10.9	88.2	37.6	63.1	12.2
MiDaS [36]	✓			11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5	11.8
LeReS [55]	✓			9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.1	27.1	76.6	9.5
Omnidata [14]	✓			7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.2	33.9	74.2	9.2
HDN [57]	✓			6.9	94.8	11.5	86.7	12.1	83.3	8.0	93.6	24.6	78.0	7.2
DPT [37]	✓			9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.8	18.2	75.8	7.0
Depth Anything V2 [52]	✓			**4.4**	**97.9**	**7.5**	**94.8**	13.2	84.2	**4.4**	**98.0**	**6.5**	**95.4**	**2.3**
Marigold [27]		✓		6.0	95.9	10.5	90.4	7.1	95.1	6.9	94.5	31.0	71.2	6.0
DepthFM [18]		✓		6.5	95.6	8.3	93.4	-	-	-	-	22.5	80.0	3.8
Geowizard [15]		✓		5.7	96.2	14.4	82.0	7.5	94.3	6.1	95.8	31.4	77.1	6.2
E2E-FT [16]		✓		5.4	96.4	9.6	92.1	**6.4**	**95.9**	5.8	96.5	30.3	77.6	3.6
I2SB [30]			✓	23.5	63.6	35.1	46.4	53.4	40.5	21.1	67.4	55.5	58.2	14.0
BBDM [29]			✓	33.2	49.6	58.6	25.6	52.0	40.6	25.2	61.2	70.1	46.9	15.2
DDBM [61]			✓	26.4	62.2	55.1	27.6	69.5	33.8	23.2	65.6	52.3	56.9	14.5
DPBridge (SD1.5)			✓	6.9	95.7	11.2	87.3	8.1	93.2	7.5	94.7	29.7	78.6	5.4
DPBridge (SD2.1)			✓	**4.4**	97.6	9.8	91.6	6.9	94.9	4.5	97.4	23.8	84.0	2.7

Table 1. Quantitative comparison on affine-invariant depth estimation benchmarks. All metrics are presented in percentage terms. FF, DM, and BM correspond to feed-forward, diffusion models, and bridge models. The best and second-best results are highlighted in **bold** and underline formats.

Method	Model Type			NYUv2		ScanNet		iBims-1		Sintel		Avg.
	FF	DM	BM	Mean↓	11.25° ↑	Mean↓	11.25° ↑	Mean↓	11.25° ↑	Mean↓	11.25° ↑	Rank
EESNU [3]	✓			16.2	58.6	-	-	20.0	58.5	42.1	11.5	6.8
Omnidata v1 [14]	✓			23.1	45.8	22.9	47.4	19.0	62.1	41.5	11.4	8.8
Omnidata v2 [26]	✓			17.2	55.5	16.2	60.2	18.2	63.9	40.5	14.7	6.2
DSINE [2]	✓			16.4	59.6	16.2	61.0	17.1	67.4	34.9	21.5	3.5
Marigold [27]		✓		18.8	55.9	17.7	58.8	18.4	64.3	39.1	14.9	6.4
Geowizard [15]		✓		17.0	56.5	15.4	61.6	**13.0**	65.3	40.4	13.2	4.2
E2E-FT [16]		✓		16.5	60.4	14.7	66.1	16.1	**69.7**	33.5	22.3	2.1
StableNormal [53]		✓		17.7	55.9	15.8	61.2	17.2	66.6	-	-	5.0
DPBridge (SD1.5)			✓	18.1	54.2	18.0	54.5	24.3	51.2	41.2	19.4	8.0
DPBridge (SD2.1)			✓	**14.8**	**65.3**	**13.4**	**68.5**	17.1	68.0	**25.0**	**46.3**	**1.4**

Table 2. Quantitative comparison on surface normal prediction benchmarks. All metrics are presented in percentage terms. FF, DM, and BM correspond to feed-forward, diffusion models, and bridge models. The best and second-best results are highlighted in **bold** and underline formats.

5.2. Quantitative and Qualitative Results

Tab. 1 and Tab. 2 summarize the quantitative results of DPBridge on zero-shot benchmarks across depth estimation and surface normal prediction tasks. We evaluate two variants of our model—DPBridge (SD1.5) and DPBridge (SD2.1)—built on top of Stable Diffusion v1.5 and v2.1-base [39], respectively. For depth estimation, DPBridge (SD2.1) achieves an AbsRel of 4.4 and δ_1 accuracy of 97.6% on NYUv2, and 4.5 AbsRel / 97.4% δ_1 on ScanNet, both among the best in class. Across all five datasets, it achieves an average ranking of 2.7—second only to Depth Anything V2 (rank 2.3), which is trained on over 1,000 times more data. For surface normal prediction, DPBridge also delivers state-of-the-art performance with a mean angular error of 14.8° and 65.3% accuracy under 11.25° on NYUv2, outperforming both feed-forward and diffusion-based methods. It also maintains strong results on iBims-1 (17.1, 68.0%) and Sintel (25.0, 46.3%). These results highlight DPBridge's versatility to generalize well across tasks and benchmarks. Notably, compared to prior bridge models such as I2SB, BBDM, and DDBM, DPBridge reduces average error by a large margin (e.g., 4.4 vs. 23.5 AbsRel in NYUv2 depth), underscoring the benefits and importance of incorporating pretrained visual priors to bridge construction. Furthermore, among diffusion-based approaches, DPBridge consistently achieves the best or second-best scores, validating our core hypothesis that bridge formulation provides clear advantages over standard diffusion frameworks in dense prediction tasks,

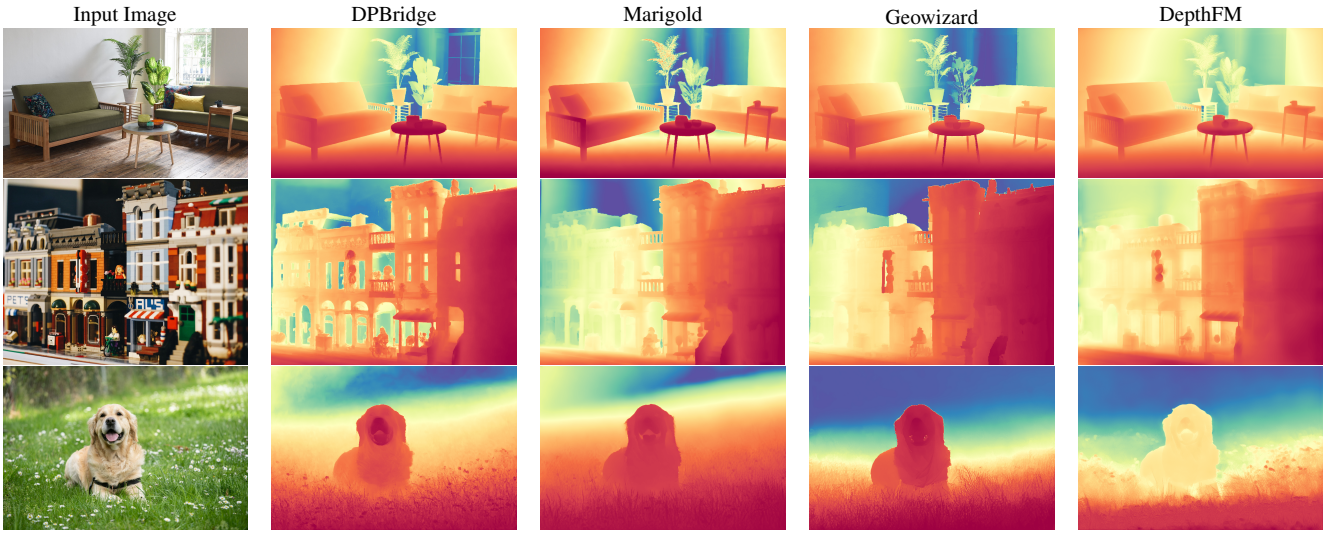


Figure 3. Qualitative comparison with diffusion-based depth estimation methods on in-the-wild images. DPBridge can generate depth maps with sharper boundaries and more coherent scene geometry.

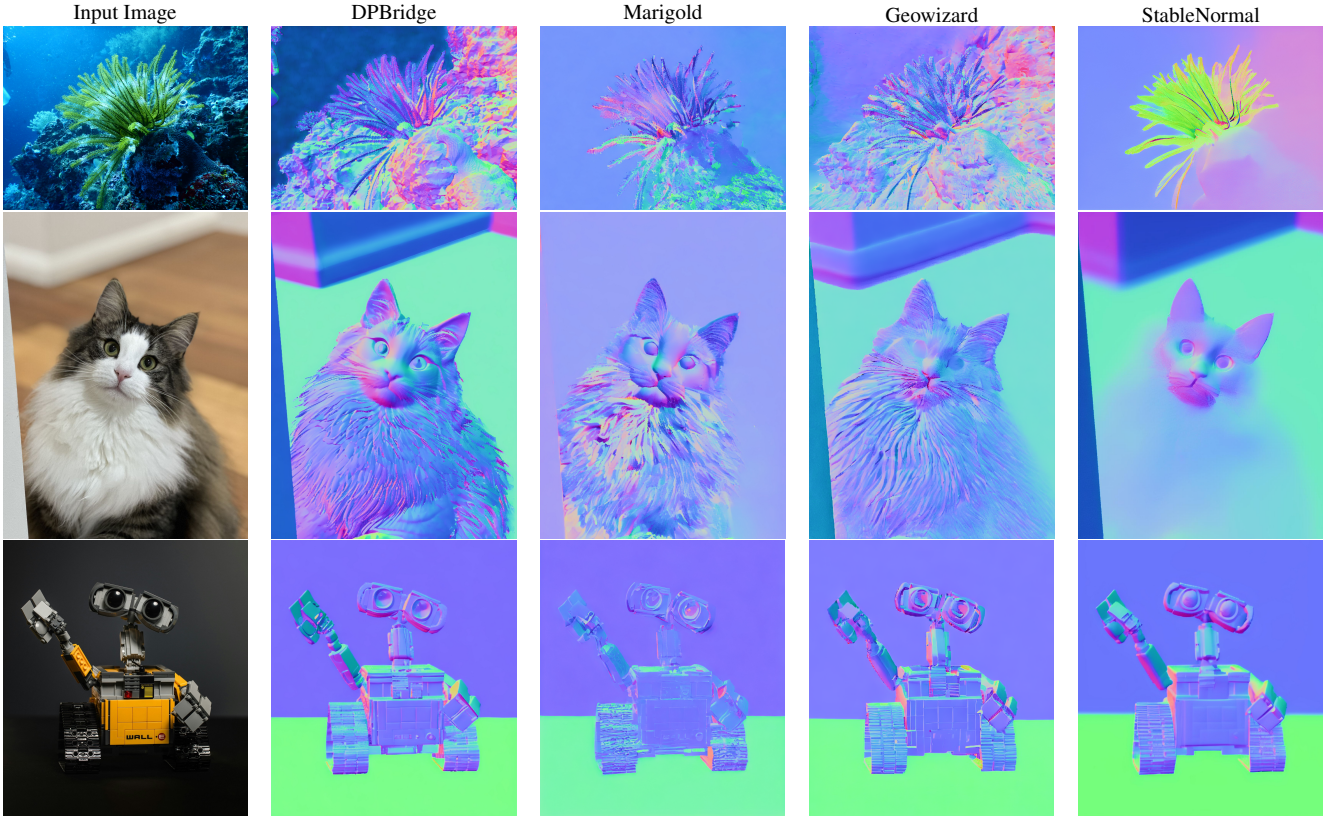


Figure 4. Qualitative comparison with diffusion-based surface normal prediction methods on in-the-wild images. DPBridge exhibits more fine-grained geometric details, such as the coral and reef's contour, the kitten's fur details, and the robot's body composition curvature.

and combining bridge modeling with pretrained visual prior offers a more efficient and generalizable framework.

Figure 3 and Figure 4 present qualitative comparisons of DPBridge with SOTA diffusion-based methods on in-the-wild images for depth estimation and surface normal prediction, respectively. As shown in Figure 3, DPBridge generates depth maps with sharper object boundaries and more coherent scene geometry, particularly in challenging

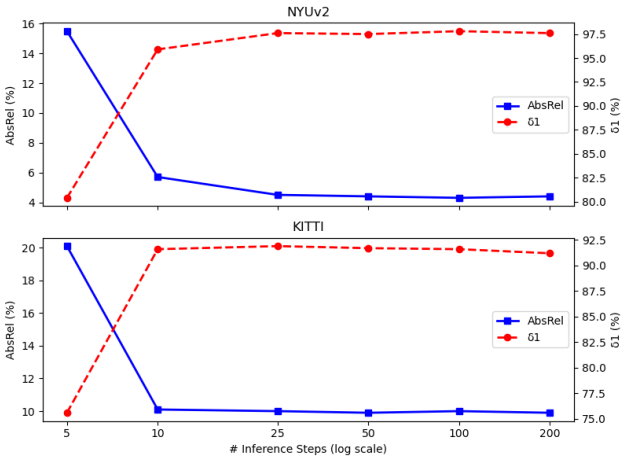


Figure 5. Ablation of sampling steps. The performance improves as the number of steps increases, while observing saturation after 10 steps.

Design Components			NYUv2		KITTI	
FT	*DAN*	*IC*	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
	✓	✓	21.8	83.7	30.6	62.7
✓			6.4	94.0	11.2	87.0
✓		✓	6.2	95.8	11.0	87.5
✓	✓		4.7	97.1	10.0	91.0
✓	✓	✓	4.4	97.6	9.8	91.6

Table 3. Ablation studies regarding different design components. *FT* stands for finetuning pretrained backbone. *DAN* denotes distribution aligned normalization technique. *IC* indicates image consistency loss.

cases such as urban scenes or complex foreground objects like animals, grass, and flowers. In contrast, competing methods like Geowizard and DepthFM often produce oversmoothed outputs or fail to preserve depth discontinuities. In Figure 4, DPBridge demonstrates stronger fidelity in fine-grained geometric details, such as the contour of coral and reef, the fur details of the cat, and the body composition curvature of the robot. Competing method like Marigold tends to focus on the center object in the image and ignore the environment, while StableNormal produces unreliable outputs. These visual results reinforce the quantitative findings and highlight the effectiveness of DPBridge in modeling structured output distributions with both global coherence and fine details.

5.3. Ablation Studies

To evaluate the effectiveness of each design component and training setting, we conduct ablation studies using depth estimation tasks on the NYUv2 and KITTI datasets. The results are summarized in Tab. 3 and Figure 5, with detailed discussions provided below.

Visual Prior Exploitation. To evaluate the usage of visual priors, we compare finetuning with training from scratch. As shown in the first row of Tab. 3, finetuning yields significantly better results under identical settings. This is attributed to stronger visual priors from foundation models and faster convergence. Additionally, the last rows of Tab. 1 and Tab. 2 show that SD2.1 outperforms SD1.5, highlighting that higher-quality priors further boost performance. These results confirm the advantage of building DPBridge on strong pretrained backbones.

Distribution Alignment Normalization. We next assess the benefit of aligning marginal distributions between the diffusion bridge and standard diffusion processes. As shown in the second and fourth rows of Tab. 3, adding a distribution alignment normalization step before the noise prediction network consistently improves performance.

Image Consistency Loss. As shown in the third and fourth rows of Tab. 3, the auxiliary image consistency loss improves performance but has a smaller impact than distribution alignment. This is likely because it mainly affects the boundaries or overlapping regions between masked and valid areas, which are limited in scope. Nonetheless, when combined with other components, it contributes to a notable overall performance gain.

Number of Sampling Steps. We evaluate performance under varying sampling steps during inference. As shown in Figure 5, perception quality improves rapidly with more steps and saturates after 10. Notably, compared to Tab. 1, where other diffusion-based methods use 50 steps, DPBridge achieves comparable results with significantly fewer steps, highlighting its efficiency.

6. Conclusion

In this paper, we propose DPBridge, a latent diffusion bridge framework for dense prediction tasks through integrating the strengths of diffusion bridge modeling with visual priors from pretrained foundation models. We derive a fully tractable bridge process to establish direct probability transport in the latent space, and propose finetuning strategies to better adapt pretrained diffusion backbones in bridge construction. While DPBridge achieves competitive performance across different tasks and benchmarks, several limitations remain. First, our method still requires iterative sampling, which makes it slower than feed-forward approaches. Exploring acceleration strategies such as distillation or single-step bridge solvers could further improve practical deployment, especially in real-time settings. Second, our formulation is restricted to certain bridge parameterization, and exploring alternatives to adapt to more advanced pretrained backbone is a promising future direction.

Acknowledgement

This research is funded in part by an ARC (Australian Research Council) Discovery Grant of DP220100800. Prof Hongdong Li holds concurrent appointments as a Full Professor with the ANU and as an Amazon Scholar with Amazon (parttime). This paper describes work performed at ANU and is not associated with Amazon.

References

[1] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. 4

[2] Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9535–9545, 2024. 2, 6

[3] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13137–13146, 2021. 2, 6

[4] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 3

[5] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012. 5

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

[7] Zehua Chen, Guande He, Kaiwen Zheng, Xu Tan, and Jun Zhu. Schrodinger bridges beat diffusion models on text-to-speech synthesis. *arXiv preprint arXiv:2312.03491*, 2023. 3

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2

[9] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 2, 3

[10] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022. 2

[11] Hyungjin Chung, Jeongsol Kim, and Jong Chul Ye. Direct diffusion bridge using data consistency for inverse problems. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 5

[13] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021. 2, 3

[14] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 1, 6

[15] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2025. 2, 6

[16] 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. 2, 6

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 5

[18] 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. 2, 6

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 3

[21] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. 2

[22] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *European conference on computer vision*, pages 668–685. Springer, 2022. 2

[23] Haorui Ji and Hongdong Li. 3d human pose analysis via diffusion synthesis. *arXiv preprint arXiv:2401.08930*, 2024. 3

[24] Haorui Ji, Hui Deng, Yuchao Dai, and Hongdong Li. Unsupervised 3d pose estimation with non-rigid structure-from-motion modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3314–3323, 2024.

[25] Haorui Ji, Rong Wang, Taojun Lin, and Hongdong Li. Jade: Joint-aware latent diffusion for 3d human generative modeling. *arXiv preprint arXiv:2412.20470*, 2024. 3

[26] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18974, 2022. 6

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

[28] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 5

[29] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 1952–1961, 2023. 2, 3, 6

[30] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I^2sb: Image-to-image schrödinger bridge. *arXiv preprint arXiv:2302.05872*, 2023. 2, 3, 6

[31] Xingchao Liu and Lemeng Wu. Learning diffusion bridges on constrained domains. In *international conference on learning representations (ICLR)*, 2023. 2, 3

[32] Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022. 2, 3

[33] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[34] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013. 3

[35] Michele Pavon, Giulio Trigila, and Esteban G Tabak. The data-driven schrödinger bridge. *Communications on Pure and Applied Mathematics*, 74(7):1545–1573, 2021. 2, 3

[36] 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*, 44(3):1623–1637, 2020. 6

[37] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 1, 2, 6

[38] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 5

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 4, 6

[40] Simo Särkkä and Arno Solin. *Applied stochastic differential equations*. Cambridge University Press, 2019. 3

[41] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 5

[42] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 5

[43] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. 2, 3

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[45] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[46] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 5

[47] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 2

[48] Yuji Wang, Zehua Chen, Xiaoyu Chen, Jun Zhu, and Jianfei Chen. Framebridge: Improving image-to-video generation with bridge models. *arXiv preprint arXiv:2410.15371*, 2024. 5

[49] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, et al. Step: Segmenting and tracking every pixel. *arXiv preprint arXiv:2102.11859*, 2021. 2

[50] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. 2

[51] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong

Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 3

[52] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 1, 6

[53] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 43 (6):1–18, 2024. 6

[54] 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. 6

[55] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021. 6

[56] Conghan Yue, Zhengwei Peng, Junlong Ma, Shiyan Du, Pengxu Wei, and Dongyu Zhang. Image restoration through generalized ornstein-uhlenbeck bridge. *arXiv preprint arXiv:2312.10299*, 2023. 3

[57] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. *Advances in Neural Information Processing Systems*, 35:14128–14139, 2022. 6

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[59] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43 (4):1–20, 2024. 3

[60] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 2

[61] Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. *arXiv preprint arXiv:2309.16948*, 2023. 2, 3, 4, 6