Re-Depth Anything: Test-Time Depth Refinement via Self-Supervised Re-lighting

Ananta R. Bhattarai
Bielefeld University

Helge Rhodin
Bielefeld University

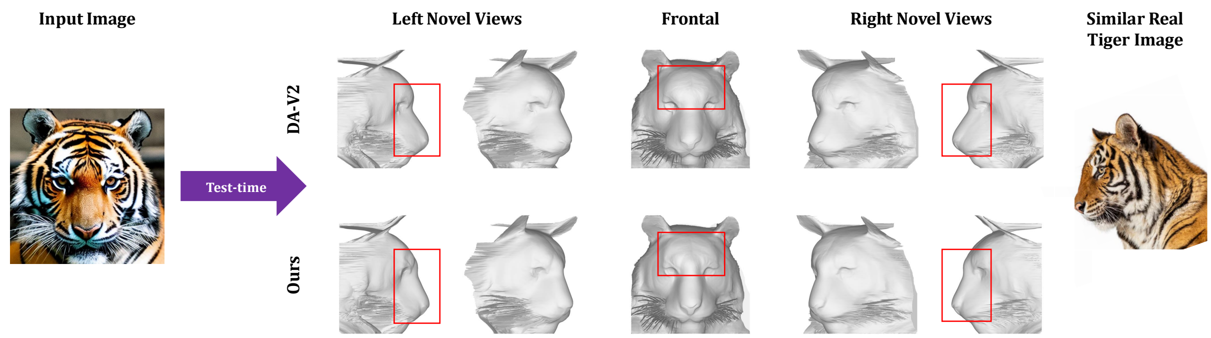

Figure 1. **Re-Depth Anything** refines the predictions of Depth Anything V2 [53] by re-lighting the reconstructed geometry and extracting knowledge from diffusion models in a self-supervised manner. In this example, test-time optimization enhances facial detail (see frontal view) and refines the nose shape to look more like a tiger (side view), correcting the dog-like initial resemblance likely originating from a biased training distribution. The key contribution is a re-synthesis method that replaces photometric reconstruction for self-supervision.

Abstract

Monocular depth estimation remains challenging, as foundation models such as Depth Anything V2 (DA-V2) struggle with real-world images that are far from the training distribution. We introduce Re-Depth Anything, a test-time self-supervision framework that bridges this domain gap by fusing foundation models with the powerful priors of large-scale 2D diffusion models. Our method performs label-free refinement directly on the input image by re-lighting the predicted depth map and augmenting the input. This re-synthesis method replaces classical photometric reconstruction by leveraging shape from shading (SfS) cues in a new, generative context with Score Distillation Sampling (SDS). To prevent optimization collapse, our framework updates only intermediate embeddings and the decoder's weights, rather than optimizing the depth tensor directly or fine-tuning the full model. Across diverse benchmarks, Re-Depth Anything yields substantial gains in depth accuracy and realism over DA-V2, and applied on top of Depth Anything 3 (DA3) achieves state-of-the-art results, showcasing new avenues for self-supervision by geometric reasoning.

Code: https://github.com/anantarb/Re-Depth-Anything

1. Introduction

Monocular Depth Estimation (MDE) aims to predict dense per-pixel depth from a single RGB image, enabling numerous applications, including 3D reconstruction [13, 26, 55], autonomous driving [46], robotic navigation [50], and virtual or augmented reality [34]. Lately, high-quality depth maps for diverse images have been enabled by foundation models using Vision Transformers (ViTs) [7] with dense prediction heads [33]. Crucially, MiDaS [32] pioneered training on a myriad of labeled datasets, and Depth Anything V2 (DA-V2) [53] showed that sparse and often noisy depth labels can be enhanced with a teacher model trained on synthetic data. While these foundational models are setting new state-of-the-art performance, inaccuracies in in-the-wild reconstruction remain (see Fig. 1). Monocular depth estimation remains one of the fundamental yet challenging problems in computer vision.

In this work, we introduce Re-Depth Anything, a test-time optimization framework designed to close the domain gap for feed forward models such as DA-V2 through self-supervision with 2D generative models. Given a single input image, our framework adapts the pre-trained model to the specific image content. The core idea is to re-light the feed-forward depth map under random illumination conditions and superimpose these renderings onto the input image. The depth map is then refined using a 2D diffusion model as a prior to score how realistic the augmented shading is. This plausibility estimate is backpropagated to the depth map through a differentiable renderer using the Blinn-Phong illumination model [3] and the SDS loss. Crucially, instead of directly optimizing the depth map or fine-tuning the entire network, we propose a targeted optimization strategy that jointly optimizes only the intermediate feature embeddings fed to the Dense Prediction Transformer (DPT) decoder and the decoder's weights. This approach preserves the strong geometric knowledge encoded in the embeddings while refining the final output.

3D knowledge distillation from 2D image diffusion models [30, 36, 37] using the SDS loss and shading cues was pioneered by DreamFusion [31] for text-to-3D generation. Their key ingredient is optimizing a 3D NeRF representation [26] such that its 2D renderings are perceived as realistic by the diffusion model. This and subsequent works [20, 23, 43, 47, 62] have enabled reconstruction of real images by pairing virtual views with photometric reconstruction of the real one. However, this line of purely self-supervised learning from geometric relations suffers from cue ambiguities and lags behind supervised models. Our key advance is to apply the benefits of self-supervised learning on top of supervised methods by re-lighting the predicted depth map. This re-synthesis and augmentation of the input image is fundamentally different from photometric reconstruction with a full-fledged NeRF [26] or Gaussian Splatting [13] renderer in prior self-supervised work, and alleviates the problem of reconstructing appearance pixel-perfect. Our main contributions are:

- Re-Depth Anything, a novel test-time optimization framework that adapts a pre-trained feed-forward model to real-world images using a 2D diffusion prior on re-synthesized depth predictions (w/o labeled data).
- We propose a single-image re-lighting model that differentiably links the predicted depth map to the input image, enabling the use of an SDS loss for self-supervised geometry refinement from a single view.
- We introduce a targeted optimization scheme that jointly optimizes the decoder's input embeddings and its weights, which we show is crucial for avoiding overfitting and preserving geometric structure.
- The method is developed on DA-V2, and we verify generality on Depth Anything 3 (DA3) [21].

2. Related Work

Our work, Re-Depth Anything, builds upon progress in three primary research areas: monocular depth estimation, test-time adaptation for monocular depth estimation, and the use of 2D diffusion models as priors for 3D reconstruction. Below, we review the most related methods from each of these domains.

Monocular Depth Estimation. Monocular depth estimation has been a long-standing challenge in computer vision. Early approaches [1, 8, 15, 19, 40, 51, 54, 56, 58] relied heavily on supervised learning, training on datasets with ground-truth depth, such as KITTI [9] for outdoor driving and NYU Depth V2 [41] for indoor scenes. More recently, the field has shifted towards building general-purpose foundation models for depth. MiDaS [32] enabled joint training on multiple datasets by predicting disparity instead of depth and by normalizing predictions to a unit range, enabling zero-shot generalization to new domains. Even better performance is possible with DPT heads [33] and large diffusion models [12, 61]. DA [52] and its successor, DA-V2 [53], use the same relative prediction and achieved further improvements by training on massive-scale image datasets, aligning the predictions of a teacher model to sparse ground-truth measurements. This approach mitigates, but does not resolve, noise in LiDAR-based ground truth on in-the-wild images.

Another line of research predicts absolute depth without disparity normalization, from single images [4, 29, 44, 57] or multiple images [5, 14, 45]. We focus on relative depth prediction, as these models typically yield higher surface detail, which we aim to improve. Moreover, the absolute depth scale is invariant to shading cues. Hence, we use the recent and popular DA-V2 model as our foundation, aiming to correct its errors rather than retraining it, so that it applies to underrepresented and out-of-distribution inputs.

Test-Time Adaptation for MDE. Test-Time Adaptation (TTA) or Test-Time Optimization (TTO) aims to adapt a pre-trained model to a specific test input at inference time.

In the context of MDE, TTA often relies on self-supervision signals available from the input itself. For video inputs, temporal and photometric consistency between frames is a powerful signal used to fine-tune a depth network [16, 18, 28].

TTA for a single image is challenging because self-supervision cues are scarce. While existing methods rely on specific external priors—such as 3D meshes or sparse points—our method leverages a general-purpose 2D diffusion model to refine relative depth for any arbitrary scene. Crucially, instead of relying solely on the input image's features, we introduce a powerful prior that provides a dense, geometry-aware supervisory signal for adaptation.

2D Diffusion Models as Priors for 3D. 2D image diffusion models [30, 36, 37], trained on internet-scale image and text data, have learned incredibly rich priors about the visual world. A recent line of work has focused on leveraging these 2D priors for 3D tasks. The most influential is DreamFusion [31], which introduced the SDS loss, enabling the use of a pre-trained text-to-image diffusion model as a loss function to optimize a 3D NeRF representation from scratch using only a text prompt. This concept has been extended and improved by numerous follow-ups [23, 25, 43, 62], including using mesh representations [20, 47] and Gaussian Splatting [6, 42] as differentiable renderers. Reconstructing a real input image brings additional challenges. RealFusion [24] proposes a method that combines real and synthetic views through diffusion-model fine-tuning and carefully selects compatible virtual views, while others directly fine-tune a 2D diffusion model for multi-view reconstruction [22]. Our re-lighting principle is inspired by the recent virtual shape-from-texture approach [2], which utilizes shape-from-texture cues through virtual augmentation. However, these fully unsupervised methods have not demonstrated advantages over the latest supervised depth estimation techniques.

Our work builds directly on this idea but applies it in a novel context. Instead of generating a 3D shape or optimizing a NeRF from multiple views, we use the SDS loss on a single image to refine the parameters of a pre-trained, feed-forward depth estimation model (e.g., DA-V2) using re-lighting rather than photometric reconstruction. This test-time optimization adapts the model's prediction to the specific image content, guided by the diffusion model's knowledge about the shading of natural objects.

Single-View Geometry and Shading. Shape-from-Shading (SfS) [10, 60] is a classical attempt to recover 3D shape from shading variations in a single 2D image. The idea is to decompose the image into (piecewise-constant) albedo and shading components, e.g. using the diffuse and specular components of a Blinn-Phong [3] model. However, classical SfS [10, 60] and other shape-from-X methods, such as Shape-from-Texture (SfT) [48, 49], are highly ill-posed, relying on strong assumptions about lighting, texture regularity, and material properties that are rarely met in practice.

DreamFusion and RealFusion were the first to revisit the SfS principle in a modern context using generative models. Although they could lift some assumptions, RealFusion still follows the classical reconstruction principle of rendering a synthetic object optimized to reconstruct the input image. However, real illumination and material properties are complex, leading to artifacts around specular highlights when attempting to match them with simplistic shading models in deployed differentiable renderers.

By contrast, we do not attempt to solve the full, ill-posed inverse graphics problem. Instead, one of our key contributions is to augment the input with additional shading cues, which is achieved by modifying a simple, lightweight Blinn-Phong [3] renderer. This allows the diffusion prior to critique the plausibility of the 3D shape as expressed through its shading, linking the underlying geometry and image content without photometric reconstruction.

3. Preliminaries and Notation

Score Distillation Sampling. Given a 2D image $\mathbf{X}$, rendered from a differentiable representation with parameters θ, SDS [31] utilizes a pre-trained diffusion model ϕ to optimize θ via gradient descent. Specifically, a gradient toward a more likely image is obtained from the noise ϵ_ϕ predicted by ϕ given a noisy image $\mathbf{X}_t$, text embedding c, and noise level t,

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\mathbf{X}, c) = \mathop{\mathbb{E}}_{\epsilon,t}\left[w(t)\left(\epsilon_\phi(\mathbf{X}_t; c, t) - \epsilon\right)\frac{\partial \mathbf{X}}{\partial \theta}\right], \quad (1)$$

where $w(t)$ is a weighting function and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In DreamFusion [31], the SDS loss matches 2D renderings from random angles to the text prompt. In our work, we apply the SDS loss to the re-lit image and use it to optimize both the intermediate feature embeddings and decoder weights in DA-V2 to enhance depth prediction.

Depth Anything V2 Architecture. DA-V2 [53] follows the recent trend of using ViT encoders, as in DINOv2 [27] and DPT [33], with a DPT-based disparity decoder. Specifically, the encoder transforms input tokens into new feature representations using L transformer layers. Features from four selected layers are extracted and passed to the DPT head for disparity prediction, with layer selection depending on the ViT variant. For example, layers $l = 3, 6, 9, 12$ are used in the ViT-Small configuration. Given an input image $\mathbf{I}$, we denote the extracted feature representations as embeddings $\mathbf{W}$. The final disparity $\hat{\mathbf{D}}_{\text{disp}} = f(\mathbf{W}; \theta)$ is predicted by the DPT head $f(\cdot)$, parameterized by pre-trained weights θ and taking embeddings $\mathbf{W}$ as input.

4. Method

Given an input image $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$, our goal is to refine an initial disparity map estimate $\hat{\mathbf{D}}_{\text{init}} \in \mathbb{R}^{H\times W}$ predicted by the pre-trained DA-V2 [53] model. As illustrated in Fig. 2, Re-Depth Anything is a self-supervised test-time optimization framework that adapts the DA-V2 model to the specific input image.

Key to our method is how we use the SDS [31] loss as a 2D diffusion prior for 3D refinement. To this end, we first introduce a differentiable rendering function that links the predicted disparity map $\hat{\mathbf{D}}_{\text{disp}}$ to a re-illuminated image $\hat{\mathbf{I}}$ through augmentation. We then use the SDS loss on $\hat{\mathbf{I}}$ to jointly optimize the decoder's input embeddings $\mathbf{W}$ and the weights θ to better align the re-illuminated and original images. We describe each component in detail below.

4.1. Shaded Depth Rendering

To leverage a 2D diffusion prior, we must establish a differentiable link between the depth map $\hat{\mathbf{D}}$ and a 2D image. A full inverse-rendering approach, which would decompose

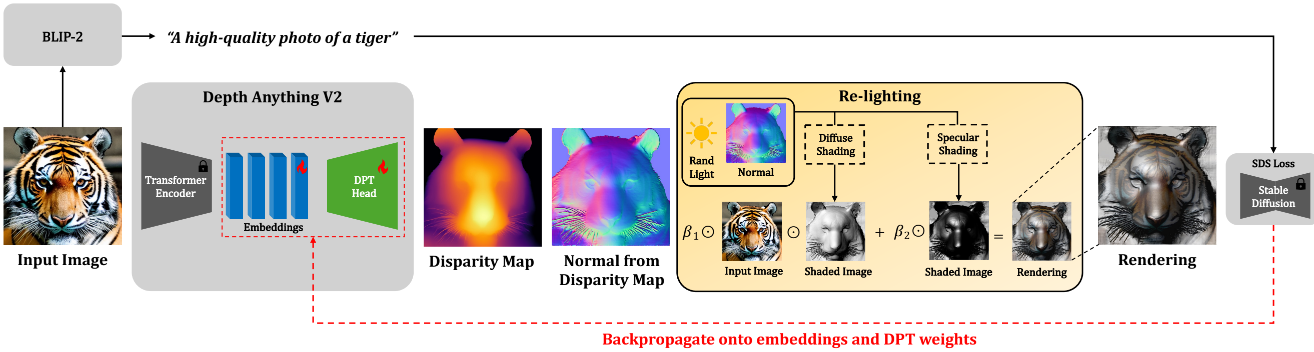

Figure 2. **Re-Depth Anything overview.** Our main contribution is the re-lighting module, which randomizes light conditions and shades the estimated geometry on the input. Notably, the re-lighting does not need to look physically accurate, as we are only augmenting the image, not photometrically reconstructing it. Key is also the optimization of the embeddings and decoder, while leaving the encoder frozen.

the scene into albedo, lighting, materials, and 3D geometry, is highly ill-posed from a single image.

Instead, we propose augmenting the input image with additional shading effects through re-lighting. We synthesize diffuse and specular reflectance maps with the classical Blinn-Phong shading model [3] as a function of the normals $\mathbf{N}$ of the depth map. Computing the normals at pixel coordinates u, v requires a camera model. We test scaled orthographic and perspective projections with intrinsic matrices $\mathbf{K}_{\text{orth}}$ and $\mathbf{K}_{\text{persp}}$, respectively. We then unproject the depth map element-wise into a 3D mesh with vertices

$$\mathbf{X} = \mathbf{K}_{\text{persp}}^{-1} \begin{pmatrix} \mathbf{U}\hat{\mathbf{D}} \\ \mathbf{V}\hat{\mathbf{D}} \\ \hat{\mathbf{D}} \end{pmatrix} \text{ or } \mathbf{X} = \mathbf{K}_{\text{orth}}^{-1} \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \\ \hat{\mathbf{D}} \end{pmatrix}, \quad (2)$$

where $\mathbf{U}$ and $\mathbf{V}$ are the tensors of horizontal and vertical pixel coordinates ranging from -1 to 1. The normal $\mathbf{N}$ is orthogonal to the spatial gradients $(\nabla\mathbf{X}_u, \nabla\mathbf{X}_v)$, which are computed across the entire image using the element-wise cross product.

$$\mathbf{N} = \frac{\nabla\mathbf{X}_v \times \nabla\mathbf{X}_u}{||\nabla\mathbf{X}_v \times \nabla\mathbf{X}_u||_2}. \quad (3)$$

Crucial for our re-lighting of the input image is using the inverse-tonemapped input image, $\tau^{-1}(\mathbf{I})$, as a proxy for the scene's diffuse albedo, and we assume that specular highlights are colorless and not affected by albedo. While not physically accurate, this approach exploits the fact that illumination effects are linear, and there remains ambiguity between light and surface color, providing a sufficient, differentiable connection between geometry and re-lit appearance. Specifically, we synthesize a re-illuminated image $\hat{\mathbf{I}} \in \mathbb{R}^{C\times H\times W}$ using Blinn-Phong shading [3],

$$\hat{\mathbf{I}} = \tau\big(\beta_1 \max(\mathbf{N}\cdot\mathbf{l}, 0)\odot\tau^{-1}(\mathbf{I}) + \beta_2 \max(\mathbf{N}\cdot\mathbf{h}, 0)^{\alpha}\big), \quad (4)$$

where $\mathbf{N} \in \mathbb{R}^{3\times H\times W}$ is the per-pixel normal map derived from the depth gradients of $\hat{\mathbf{D}}$, $\mathbf{l} \in \mathbb{R}^3$ is the light direction, $\mathbf{h} \in \mathbb{R}^3$ is the halfway vector between $\mathbf{l}$ and the view direction $\mathbf{v} = [0, 0, 1]^T$, and α, β_1, and β_2 are material and light-intensity parameters that are determined in the next section. The tone-mapping function $\tau(\mathbf{I}) = \mathbf{I}^{1/\gamma}$, with $\gamma = 2.2$, ensures that shading is performed in the linear RGB color space.

Note that using $\mathbf{I}$ as the albedo can double the shading effects (e.g. existing shadows get darker), and the addition of specular maps may saturate the image; however, this rarely occurs in our experiments and is similar to specular highlights appearing white in photographs. We keep values within bounds by clamping the rendered output $\hat{\mathbf{I}}$ to the range $[10^{-3}, 1]$.

Handling normalized relative depth. DA-V2 outputs normalized relative depths in the form of the normalized disparity map $\hat{\mathbf{D}}_{\text{disp}} = (1/\hat{\mathbf{D}} - m)s$, where m is the minimum disparity and s is the inverse of the maximum–minimum disparity range. Converting to absolute depth requires

$$\hat{\mathbf{D}} = \frac{1}{\hat{\mathbf{D}}_{\text{disp}}/s + m} = \frac{s}{\hat{\mathbf{D}}_{\text{disp}} + ms}, \quad (5)$$

where neither the scaling s nor the offset m is known at test time. However, the normal is invariant to global scale, and hence we only have to optimize the unknown scalar $b = ms$ alongside depth refinement. Notably, we found the optimization to be insensitive to these parameters, likely because shading depends on the relative angle between the light and the normal, not absolute orientation. Specifically, it is sufficient to fix $b = 0.1$ with scaled orthographic projection.

4.2. Augmentation Objective

Our goal is to refine the depth map to produce shading effects that yield plausible re-lightings of the input image. This augmentation principle lets us choose random light and material properties instead of estimating parameters for potentially absent or complex shading effects in the input image, as required for photometric reconstruction. At each optimization step, we randomly sample the light direction $\mathbf{l}$, diffuse and specular intensities (β_1, β_2), and exponent α to ensure the refined geometry is consistent with the image across diverse shading conditions.

Loss Function. Plausibility of the augmentation is measured by the total loss, which combines the generative prior with a smoothness regularizer,

$$\mathcal{L}(\hat{\mathbf{I}}, c, \hat{\mathbf{D}}_{\text{disp}}) = \mathcal{L}_{\text{SDS}}(\hat{\mathbf{I}}, c) + \frac{\lambda_1}{hw} \sum_{i,j} \|\Delta \hat{\mathbf{D}}^{i,j}_{\text{disp}}\|^2, \quad (6)$$

where $\mathcal{L}_{\text{SDS}}$ is the SDS loss defined in Eq. (1), and $\hat{\mathbf{I}}$ is the image rendered from $\hat{\mathbf{D}}$ using Eq. (4). The second term is an ℓ_1 regularizer on the disparity gradients, which encourages smoother surfaces and prevents noisy artifacts.

To obtain the conditioning text prompt c, we employ BLIP-2 [17], a state-of-the-art image-to-text model, to generate a descriptive caption for the input image $\mathbf{I}$.

4.3. Optimization Scheme

Our goal is to refine the output depth map of feed-forward depth estimators, specifically the DA-V2 model. However, directly optimizing the depth map with the proposed re-lighting loss remains an ambiguous problem with many plausible solutions. Instead, we propose optimizing the latent feature space and weights of DA-V2, thereby leveraging the prior on 3D shapes learned at training time.

Candidates for optimization are the entire DA-V2 weights or its components. We found that fine-tuning the entire DA-V2 tends to fall into poor local minima or causes the geometry to overfit to image textures. To address this, we jointly optimize only the intermediate embeddings $\mathbf{W}$ (the intermediate feature embeddings of the frozen ViT encoder) and the DPT head's weights θ.

$$\mathbf{W}^*, \theta^* = \arg\min_{\mathbf{W},\theta} \mathcal{L}(\hat{\mathbf{I}}, c, \hat{\mathbf{D}}_{\text{disp}}) \quad (7)$$

where $\hat{\mathbf{D}}_{\text{disp}} = f(\mathbf{W}; \theta)$ and $\hat{\mathbf{I}}$ is a function of $\hat{\mathbf{D}}_{\text{disp}}$.

Depth Map Ensembling. The stochastic nature of the SDS loss, primarily due to random sampling of noise ϵ and timestep t, can lead to high variance in the optimization results. Consequently, disparity predictions can vary noticeably across different runs. To stabilize the final prediction, inspired by the ensembling in Marigold [12], we perform the optimization N times with different random seeds. We then aggregate the resulting disparity maps using a simple mean operation. The final disparity map $\mathbf{D}_{\text{disp}}$ is obtained as

$$\mathbf{D}_{\text{disp}} = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{W}^*_i; \theta^*_i), \quad (8)$$

where $(\mathbf{W}^*_i, \theta^*_i)$ are the optimized embeddings and decoder weights from the i-th run.

5. Experiments

We study the effectiveness and accuracy of our self-supervised re-lighting approach on three benchmarks, demonstrating consistent improvements over the DA-V2 and DA3 baselines across established metrics, including a relative error improvement of up to 11.4% and a normal map improvement of 14.7%. Fig. 3 shows exemplary cases that are improved by removing noise from flat areas and adding missing details. Additional results are shown in the supplemental document.

Baselines. We utilize the small variant of the DA-V2 architecture as our base model, which is one of the most popular monocular relative depth estimation methods. Besides various baseline variants that use only parts of our contributions, we also compare against a simple Shape-from-Shading implementation to demonstrate the advantage of re-lighting over photometric reconstruction.

Implementation Details. For all experiments, input images are first resized, maintaining their original aspect ratio, such that at least one side measures 518 pixels to match the training resolution of the DA-V2 model. Before applying Stable Diffusion, the images are further zero-padded if they have a non-square aspect ratio and resized to 512×512.

Our entire pipeline is implemented in PyTorch. For SDS guidance, we employ v1.5 of Stable Diffusion [36]. We optimize the encoder's embeddings and DPT weights for 1000 iterations using the AdamW optimizer. We set a learning rate of 1×10^{-3} for the embeddings and 2×10^{-6} for the DPT weights. We set the regularization weight λ_1 to 1.0. At each optimization step, we uniformly sample two coefficients $(\beta_1, \beta_2) \sim \mathcal{U}[0, 1]$ and an exponent α, where $\alpha = 2^k$ and $k \sim \mathcal{U}[2, 8]$. β_1 and β_2 are subsequently normalized to ensure their sum is 1.0 (i.e. $\beta_1 + \beta_2 = 1$). Similarly, the light direction vector $\mathbf{l} = (L_x, L_y, L_z)$ is sampled by drawing the X and Y coordinates from a uniform distribution, $L_x, L_y \sim \mathcal{U}[-1, 1]$, while fixing the Z coordinate to $L_z = 1$. The resulting vector $\mathbf{l}$ is then ℓ_2-normalized. In the absence of known camera parameters for in-the-wild images, our default is a scaled orthographic camera, and we optimize the scaling starting from 7.0. For a perspective camera, the focal length is initialized to 2.0 and refined alongside disparity optimization. The final prediction is generated by aggregating the results from 10 optimization runs. Each of these 10 runs is initialized from the original pre-trained weights of the DA-V2 model. We conduct our evaluation at the resolution of the initial 518-pixel-side resized input image. One run takes approximately 80 seconds on a single NVIDIA RTX 5000.

Datasets. We evaluate Re-Depth Anything on three standard benchmarking datasets. CO3Dv2 [35] contains multi-view images of several objects across 50 categories, with camera poses, 3D point clouds, foreground masks, and depth maps that we utilize as sparse ground truth depth.

Table 1. Comparison with DA-V2 across datasets. Relative error reduction of **Ours** over DA-V2 is shown in the last row of each dataset.

Dataset	Method	*Higher is better* ↑			*Lower is better* ↓					
		δ_1	δ_2	δ_3	AbsRel	RMSE	log10	RMSE log	SI log	SqRel
CO3D	DA-V2	**1.0**	**1.0**	**1.0**	0.00227	0.0602	0.000985	0.00321	0.321	0.000244
	Ours + DA-V2	**1.0**	**1.0**	**1.0**	**0.00223**	**0.0588**	**0.000968**	**0.00314**	**0.314**	**0.000235**
	Rel. Δ *(%)*	-	-	-	**1.75**	**2.26**	**1.74**	**2.24**	**2.24**	**3.66**
KITTI	DA-V2	0.568	0.796	0.902	0.305	7.01	0.118	0.348	33.6	2.49
	Ours + DA-V2	**0.593**	**0.818**	**0.917**	**0.283**	**6.71**	**0.110**	**0.319**	**30.7**	**2.20**
	Rel. Δ *(%)*	**5.73**	**10.9**	**15.3**	**7.10**	**4.29**	**6.55**	**8.51**	**8.51**	**11.4**
ETH3D	DA-V2	0.884	0.956	0.978	0.113	0.955	0.0448	0.153	15.1	0.391
	Ours + DA-V2	**0.898**	**0.965**	**0.982**	**0.104**	**0.875**	**0.0413**	**0.143**	**14.1**	**0.347**
	Rel. Δ *(%)*	**12.2**	**21.1**	**19.5**	**8.30**	**8.39**	**7.72**	**6.44**	**6.22**	**11.1**

From each sequence with a valid depth map, we randomly selected two images. This preprocessing step yielded a total of 80 images from 20 object categories. KITTI [9] is a large-scale autonomous driving dataset featuring sparse metric depth captured by a LiDAR sensor. We randomly sampled 10 images from each sequence in the official validation set, resulting in a total of 130 images. ETH3D [39] is a high-resolution benchmark with both indoor and outdoor scenes. We randomly sampled ten raw images and their corresponding depth maps from 13 scenes. This process resulted in a total of 130 images for evaluation.

Evaluation protocol. We apply the widely adopted metrics for assessing the quality of monocular depth estimation: δ_1, δ_2, δ_3, AbsRel, RMSE, log10, RMSE log, SI log, and SqRel. We compute these metrics on absolute depth maps, obtained by first finding the least-squares affine fit to the labels in disparity, then converting to depth, and subsequently finding the affine fit in depth, as is typical for relative depth prediction on these benchmarks [53].

5.1. Quantitative Evaluation

Table 1 shows that our test-time refinement improves upon DA-V2 across CO3D, KITTI, and ETH3D on all nine evaluation metrics. Notably, we achieve significant relative reductions in error, including 8.5% in SI log and RMSE log on KITTI, alongside an 8.4% reduction in AbsRel on ETH3D.

On CO3D, errors are smaller overall, leading to smaller yet still consistent improvements. This robustness across nine different metrics validates the efficacy of our self-supervised approach, demonstrating considerable potential for fine-tuning foundational models.

Notably, the improvements are consistent across CO3D, which covers single objects in a close-up setting, street scenes in the KITTI dataset, and indoor scenes in the ETH3D dataset. This highlights the strong generalization capability of our test-time re-lighting approach, inherited from the robustness of generative image models. The nature of the improvement is best explained with examples.

Table 2. **Quantitative comparison against DA3.**

Dataset	Method	AbsRel	SqRel	Normal MSE
CO3D	DA3	0.00251	0.000317	0.000479
	Ours + DA3	**0.00238**	**0.000294**	**0.000409**
	Rel. Δ *(%)*	**4.83**	**7.39**	**14.65**
ETH3D	DA3	0.0444	0.0853	0.00057
	Ours + DA3	**0.0441**	**0.0847**	**0.00054**
	Rel. Δ *(%)*	**0.54**	**0.71**	**3.74**

Re-Depth generalization to DA3. In Table 2, we show the effectiveness of our self-supervised refinement strategy on a different backbone by applying it to the recent DA3 [21] model (DA3MONO-LARGE). Our method consistently improves DA3 on CO3D and ETH3D, achieving state-of-the-art results. In particular, normal errors are improved by up to 15%, verifying strong detail improvement. Further results and details are given in the supplemental document.

5.2. Qualitative Evaluation

For visual assessment, we present a qualitative comparison against DA-V2 in Fig. 3. Re-Depth Anything produces visibly superior results by enhancing fine-grained details such as the threads on a ball (first image), balcony railings, and electricity wires (second-to-last image), while also removing noise from flat surfaces, as indicated by an arrow in the fourth example. These qualitative improvements are consistent with the quantitative gains reported in Table 1.

We also compare to classical shape-from-shading, which fails drastically when its strict assumptions are violated. For instance, even in the relatively simple ball example in Fig. 4 (first row), discoloration of the leather leads to spurious and noisy normals. This highlights the importance of our re-lighting augmentation strategy, which does not assume albedo constancy and does not suffer from seam artifacts typically associated with shape-from-shading.

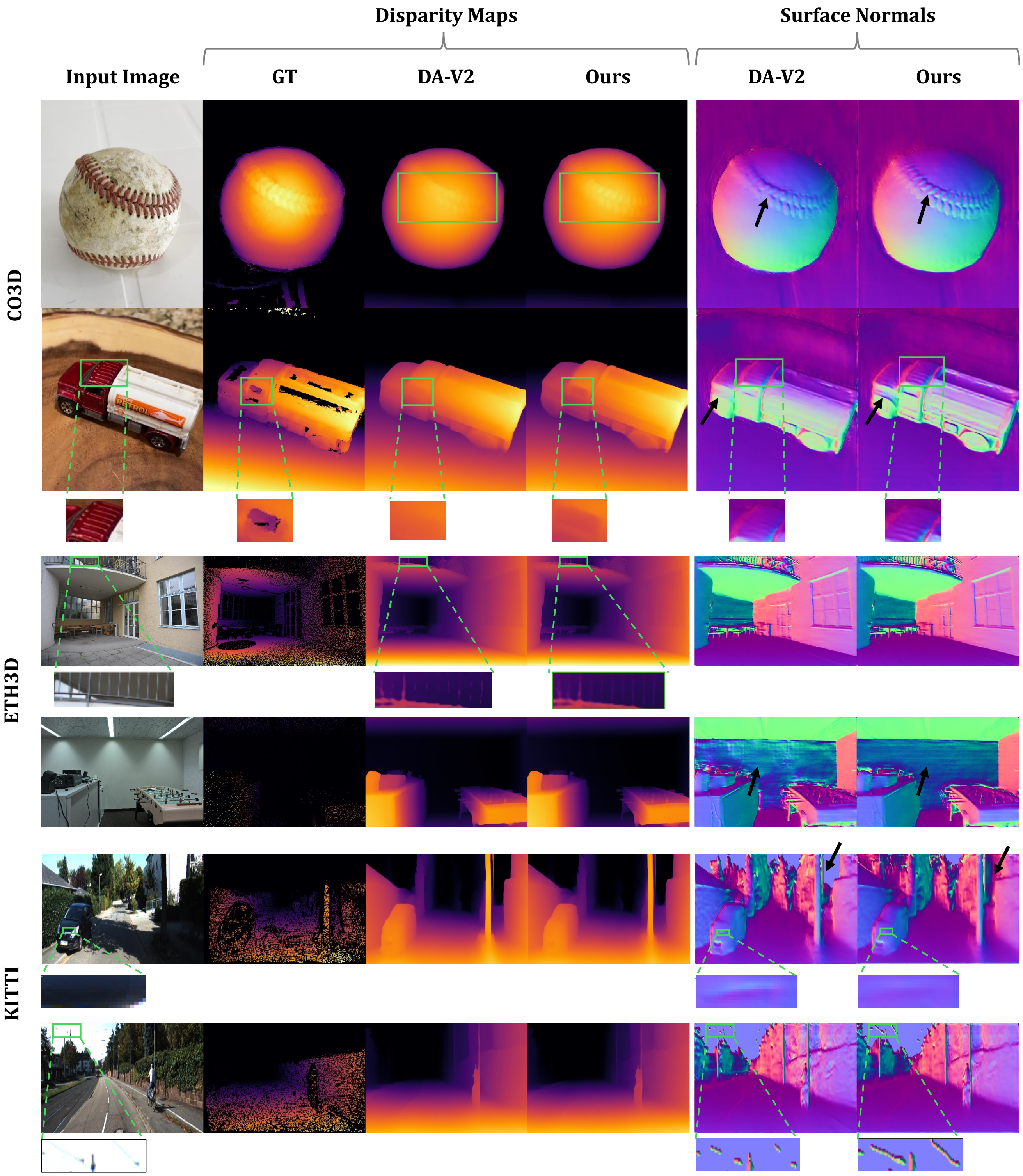

Figure 3. **Qualitative Comparison,** highlighting the added detail (rows 1,2,3,6) and noise-removal effects (rows 4,5).

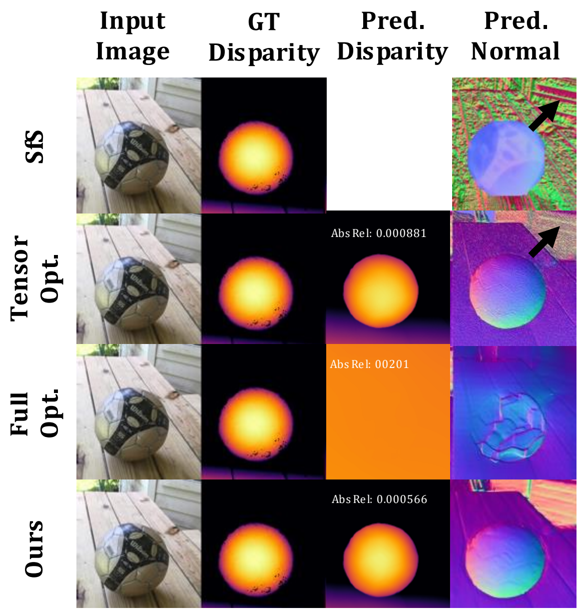

Figure 4. **Qualitative ablation** showing that directly optimizing depth or fine-tuning the whole network at once is detrimental. The listed error values relate to visual and quantitative improvements.

Table 3. **Ablation on the CO3D dataset** showing the significance of the major design choices.

Method	AbsRel ↓	RMSE ↓	log10 ↓	RMSE log ↓	SI log ↓	SqRel ↓
w/o $\mathcal{L}_{\text{SDS}}$	0.00427	0.0993	0.00185	0.00532	0.532	0.000661
Tensor Model	0.00226	0.0601	0.000985	0.00321	0.321	0.000244
Full Model	0.00331	0.0779	0.00143	0.00418	0.418	0.000412
Two Stage	0.00225	0.0597	0.000979	0.00319	0.319	0.000241
Ours	**0.00223**	**0.0588**	**0.000968**	**0.00314**	**0.314**	**0.000235**

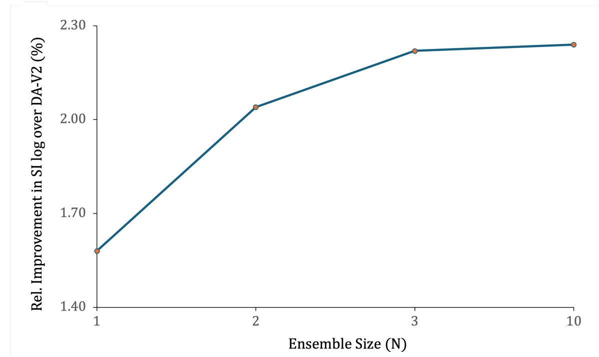

Figure 6. **Ensemble size** vs. improvement in SI log on CO3D.

Limitations. We rarely observed small hallucinated edges, such as the sticker on the truck in Fig. 3, which are plausible but incorrect. Sometimes, the method extends geometry into the sky and over-smooths fine details, such as trees in dark regions, as seen in the KITTI example (see inset Fig. 5), which could potentially be handled by thresholding. In general, we found that single objects in the CO3D dataset show stronger detail improvements, while in room and street scenes the largest gains come from removing suspicious details in the initial DA-V2 predictions that lead to unrealistic relighting highlights and are thus effectively removed by our method while preserving actual detail.

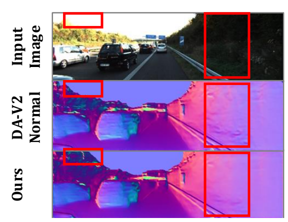

Figure 5. **Limitations.**

5.3. Ablation study

Optimization. Table 3 presents an ablation study on the CO3D evaluation set, comparing our full pipeline to a baseline lacking the SDS loss and four optimization variants: (1) direct depth pixel optimization, (2) full DA-V2 fine-tuning, (3) fine-tuning only embeddings and DPT weights, and (4) a two-stage version of (3) that first optimizes the embeddings and subsequently fine-tunes the decoder.

The qualitative results in Fig. 4 show that direct optimization (1) creates noise artifacts (first row), while full fine-tuning (2) causes the geometry to collapse (second row), even though we reduced the learning rate to a very small value of 2×10^{-6} for optimizing both the ViT encoder and the DPT decoder. Our chosen approach (3) strikes a balance, enhancing detail without degradation. These visual observations are corroborated by the quantitative metrics in Table 3. Although variants (3) and (4) are visually comparable, (3) achieves lower errors.

Ensembling predictions. We investigate the impact of ensembling predictions via mean aggregation. The results, shown in Fig. 6, demonstrate clear benefits but with rapidly diminishing returns. While a prediction from a single run achieves a 1.58% improvement in SI log over DA-V2, ensembling predictions from three runs significantly increases this to 2.22%. This benefit quickly saturates, with 10 predictions offering only a negligible further gain (2.24%).

6. Conclusion

Re-Depth Anything presents a new method for test-time optimization through re-lighting. The key contribution is using generative models to score shading–image alignment instead of requiring photometric reconstruction. This alleviates the need to construct a photorealistic renderer for inverse graphics and shows consistent improvements across all tested datasets and metrics. Notably, it sets a new state-of-the-art for detail reconstruction on the CO3D dataset using DA3 as the backbone.

Acknowledgment

We thank all members of the Visual AI for Extended Reality lab at Bielefeld University for providing crucial feedback and participating in helpful discussions.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[2] Ananta R Bhattarai, Xingzhe He, Alla Sheffer, and Helge Rhodin. Dreamtexture: Shape from virtual texture with analysis by augmentation. *ArXiv*, 2025. 3

[3] James F. Blinn. Models of light reflection for computer synthesized pictures. In *Computer Graphics and Interactive Techniques*, 1977. 2, 3, 4

[4] 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. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[5] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2025. 2

[6] Zilong Chen, Feng Wang, and Huaping Liu. Text-to-3d using gaussian splatting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[7] 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 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Neural Information Processing Systems (NeurIPS)*, 2014. 2

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2, 6

[10] Berthold K. P. Horn and Michael J. Brooks. *Shape from Shading*. MIT Press, Cambridge, Massachusetts, 1989. 3

[11] Berthold K. P. Horn and Michael J. Brooks. *The variational approach to shape from shading*, page 173–214. MIT Press, Cambridge, MA, USA, 1989. 4

[12] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2024. 2, 5, 3

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 2023. 1, 2

[14] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[15] Bo Li, Chunhua Shen, Yuchao Dai, Anton van den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[16] Huan Li, Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. On-site adaptation for monocular depth estimation with a static camera. In *British Machine Vision Conference (BMVC)*, 2023. 2

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023. 5

[18] Zhi Li, Shaoshuai Shi, Bernt Schiele, and Dengxin Dai. Test-time domain adaptation for monocular depth estimation. In *International Conference on Robotics and Automation (ICRA)*, 2023. 2

[19] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *IEEE Transactions on Image Processing*, 33: 3964–3976, 2024. 2

[20] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[21] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026. 2, 6, 3

[22] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[23] Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. Att3d: Amortized text-to-3d object synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[24] Luke Melas-Kyriazi, Iro Laina, C. Rupprecht, and Andrea Vedaldi. Realfusion 360° reconstruction of any object from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[25] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view

synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, 2024. 3

[28] Hyoungseob Park, Anjali W. Gupta, and Alex Wong. Test-time adaptation for depth completion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[29] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler. *ArXiv*, 2025. 2

[30] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[31] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3

[32] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *Transactions on Pattern Analysis and Machine Intelligence*, 44:1623–1637, 2019. 1, 2

[33] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3

[34] Alex Rasla and Michael Beyeler. The relative importance of depth cues and semantic edges for indoor mobility using simulated prosthetic vision in immersive virtual reality. In *ACM Symposium on Virtual Reality Software and Technology*, 2022. 1

[35] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021. 5

[36] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 5

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Neural Information Processing Systems (NeurIPS)*, 2022. 2

[38] Hanno Scharr. *Optimal operators in digital image processing*. PhD thesis, University of Heidelberg, Germany, 2000. 4

[39] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[40] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Nddepth: Normal-distance assisted monocular depth estimation. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[41] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, 2012. 2

[42] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[43] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Gregory Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[44] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Computer Vision and Pattern Recognition Conference (CVPR)*, 2025. 2

[45] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[46] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark E. Campbell, and Kilian Q. Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[47] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3

[48] Ryan White and David A Forsyth. Combining cues: Shape from shading and texture. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. 3

[49] Andrew P. Witkin. Recovering surface shape and orientation from texture. *Artificial Intelligence*, 17(1-3):17–45, 1981. 3

[50] Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *International Conference on Robotics and Automation (ICRA)*, 2019. 1

[51] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

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

the power of large-scale unlabeled data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[53] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2, 3, 6

[54] Xiaodong Yang, Zhuang Ma, Zhiyu Ji, and Zhe Ren. Gedepth: Ground embedding for monocular depth estimation. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[55] Chongjie Ye, Yinyu Nie, Jiahao Chang, Yuantao Chen, Yihao Zhi, and Xiaoguang Han. Gaustudio: A modular framework for 3d gaussian splatting and beyond. *ArXiv*, 2024. 1

[56] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[57] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[58] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[59] Ruo Zhang and Mubarak Shah. Shape from intensity gradient. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 29(3):318–325, 1999. 4

[60] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8): 690–706, 1999. 3

[61] Xiang Zhang, Bingxin Ke, Hayko Riemenschneider, Nando Metzger, Anton Obukhov, Markus Gross, Konrad Schindler, and Christopher Schroers. Betterdepth: Plug-and-play diffusion refiner for zero-shot monocular depth estimation. In *Neural Information Processing Systems (NeurIPS)*, 2024. 2, 4

[62] Junzhe Zhu, Peiye Zhuang, and Oluwasanmi Koyejo. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 3