LumiDepth: Stable Monocular Depth in Multi-Illumination Scenes

Anqi Cheng[1], Zhiyuan Yang[1], Tianjiao Li[1], Haiyue Zhu[2], and Kezhi Mao[1†]

[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
anqi002,zhiyuan002@e.ntu.edu.sg, tianjiao.li,ekzmao@ntu.edu.sg
[2] SIMTech, Agency for Science, Technology and Research (A*STAR), Singapore
zhu_haiyue@a-star.edu.sg

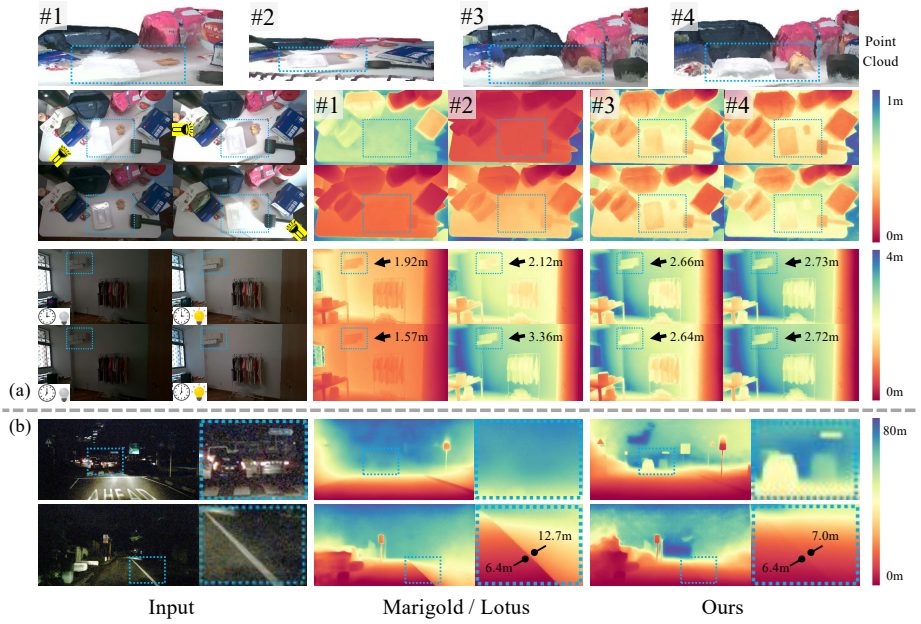

Fig. 1: Multi-illumination depth estimation. Compared with recent depth foundation models [20, 26] (middle), our method (right) preserves accurate, geometry-consistent depth. (a) ReMID-Tabletop/ReMID-Indoor: Marigold [26] is distracted by shadows, specular highlights, and exposure changes, yielding warped surfaces and missing fine structures; the point clouds (#1-2) are similarly distorted, while ours (#3-4) remain coherent across illumination variants. (b) NuScenes-Night [4]: under low light, Lotus [20] produces spurious geometry and misses critical objects or road markings, whereas our approach recovers robust, illumination-invariant depth.

† Corresponding author.

Abstract. Depth estimation in multi-illumination scenes with multiple, spatially varying light sources remains a crucial yet less-explored problem. Illumination changes introduce shadows, specular highlights, and exposure shifts that distort local appearance cues, causing severe depth inconsistency or even failure. Existing depth foundation models, trained predominantly on uniformly lit data, degrade sharply under such conditions. However, direct adaptation is challenging because ground truth depth is typically limited for multi-illumination datasets, while synthetic relighting often incurs geometric distortions. To address these challenges, we propose **LumiDepth**, a framework that learns from multi-illumination RGB images. First, a Disagreement-Calibrated Probabilistic Pseudo Supervision (DCPS) module constructs high-quality pseudo labels while preserving diversity. Second, a Frequency-aware Consistency and Distillation (FaCD) module improves cross-illumination stability without over-smoothing by enforcing low-frequency geometric consistency and distilling high-frequency structural details bi-directionally. To enable systematic evaluation, we introduce **ReMID**, a real-world multi-illumination RGB-D benchmark, together with stability metrics that quantify average and worst-case depth variation. Experiments across diverse datasets demonstrate that LumiDepth achieves state-of-the-art overall performance, markedly improving both consistency and accuracy by reducing depth variation by 30.2% and absolute relative error by 24.8%. We further show our target-domain label-free design remains effective for depth under other appearance shifts such as weather and sensor noise.*

Keywords: Monocular depth estimation · Multi-illumination scenes · Depth foundation model · Label-free training

1 Introduction

Monocular depth estimation (MDE) [10,11,14,18,31,74], the task of predicting a dense depth map from a single RGB image, has made significant strides with the advent of depth foundation models [12,20,26,43,64] trained on large-scale RGB-D datasets. While these models perform well in uniformly lit environments, we observe that they often struggle in complex real-world scenarios involving multiple light sources. In such multi-illumination scenes [4, 27, 39, 41], lighting variations caused by shadows, specular reflections, and exposure shifts alter object appearances in ways that are inconsistent with their true geometry [46,51,69]. As shown in Fig. 1, this limitation becomes particularly problematic in precision- and safety-critical scenarios, such as robotic navigation and manipulation at workspaces with mixed artificial lighting, or autonomous driving at nighttime under shifting headlights and street lamps.

* The ReMID dataset is available at `https://drive.google.com/drive/folders/1-3txE3KiFiscACxs1Xch8UO1Ax82VBNg?usp=sharing`. Accessed 26 June 2026.

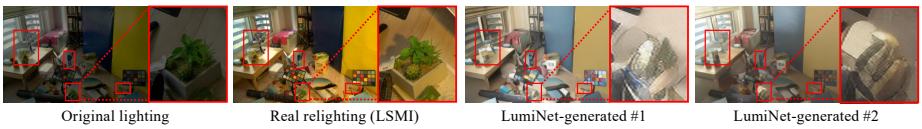

Fig. 2: Relighting results of [60] compared to real-world capture [27]. While producing plausible global illumination and overall appearance, it may introduce local texture artifacts that can alter depth-related fine geometric cues.

A natural solution might involve direct adaptation with large-scale datasets that capture scenes under diverse lighting with ground truth depth [35] or synthetically augmenting existing RGB-D datasets with relighting techniques. However, both approaches face limitations. On one hand, multi-illumination RGB-D datasets are scarce due to the difficulty of capturing accurate depth under varying lighting setups. On the other hand, although some recent relighting models [2, 21, 28, 34, 60, 70, 71] have achieved progress in synthesizing photorealistic images under diverse lighting conditions, they are primarily designed for visual appearance editing rather than geometry-aware applications, and thus lack the physical accuracy and illumination-geometry consistency required for reliable depth estimation (Fig. 2). Unlike synthetic perturbations such as rain or Gaussian noise applied in robust MDE methods [51, 56, 63, 67, 69], illumination changes affect not only global tone but also local structural cues, making data augmentation insufficient for bridging this gap.

In this paper, we propose **LumiDepth**, a framework for illumination-invariant monocular depth estimation that learns directly from multi-illumination RGB observations, without requiring target GT depth or synthetic relighting. Given a group of images depicting the same static scene under different lighting, we first build supervision from the model's own depth hypotheses. Specifically, our **Disagreement-Calibrated Probabilistic Pseudo Supervision (DCPS)** measures cross-illumination disagreement to identify unreliable predictions, filters them with a distribution-adaptive rule, and then samples pseudo labels from the remaining high-confidence set to preserve supervision diversity while avoiding brittle single-choice selection. On top of pseudo supervision, we introduce **Frequency-aware Consistency and Distillation (FaCD)** to improve cross-illumination stability without sacrificing geometric fidelity: we enforce low-frequency agreement to align global shape, while transferring high-frequency structures via a bidirectional stop-gradient distillation objective with reliability weighting, which mitigates error reinforcement when an illumination condition produces corrupted edges. Together, DCPS and FaCD enable stable training under illumination shifts and avoid the "consistent but blurry" failure mode of naive consistency regularization.

To bridge the gap between existing depth benchmarks and multi-illumination scenarios, we build **ReMID**, a real-world RGB-D dataset that captures scenes under diverse lighting conditions (*e.g.*, sunlight, room light, and torchlight). ReMID enables systematic evaluation of depth robustness and cross-illumination

consistency when appearance cues change but geometry remains fixed. Moreover, to complement conventional depth accuracy metrics and flow-based temporal consistency metrics targeting video settings, we introduce two complementary metrics that quantify both the average and worst-case geometric stability under varying illumination. Extensive experiments on both multi-illumination and standard RGB-D benchmarks validate the effectiveness of our approach. LumiDepth substantially improves depth accuracy and illumination consistency across diverse scenes, outperforming existing foundation models while maintaining strong generalization to uniformly lit datasets. Beyond lighting variation, our DCPS+FaCD training paradigm can extend to other settings that demand consistency under appearance shifts, *e.g.*, adverse weather, sensor noise, or other domain corruptions where ground truth depth is limited. Our contributions are:

- We propose LumiDepth for stable, accurate depth in multi-illumination scenes, where recent depth foundation models struggle. LumiDepth addresses the key bottleneck of limited ground truth depth, enabling adaptation from multi-illumination RGB information.
- We introduce DCPS module to build reliable, diversity-preserving pseudo labels from grouped, same-geometry, different-illumination data, and FaCD module to improve stability without over-smoothing via frequency-aware consistency and distillation.
- We build ReMID, a real-world multi-illumination RGB-D benchmark, and propose stability metrics for average and worst-case depth variation. Experiments show consistent gains in both stability and accuracy across multi-illumination and standard benchmarks.

2 Related Works

2.1 Monocular Depth Estimation

Monocular depth estimation (MDE) aims to infer a dense depth map from a single RGB image. Early supervised approaches [1, 10, 11, 31, 33, 66] relied on annotated datasets [16, 48] to train convolutional or transformer-based networks for dense depth prediction. To reduce the dependence on extensive labeled data, self-supervised methods [8, 14, 17, 18, 24, 72, 74] emerged that exploit photometric consistency across video frames for supervision.

Recent depth foundation models have greatly advanced zero-shot monocular depth estimation by scaling both data and model capacity. MiDaS [43] demonstrated strong cross-dataset generalization through relative depth supervision, establishing a foundation for transferable depth prediction. Depth Anything [36, 64, 65] further expanded this paradigm with large-scale data aggregation, achieving high accuracy across diverse domains. Beyond discriminative approaches, recent studies leverage generative priors for depth estimation. Marigold [26] first repurposed a diffusion backbone for conditional depth prediction, inspiring follow-up works that embed geometric reasoning within generative frameworks [12, 19, 20, 49, 61]. These models exploit diffusion formulations to preserve fine details and enable efficient, data-scalable depth prediction.

2.2 Consistency-Aware Depth Estimation

General MDE models often degrade under challenging visual conditions such as nighttime, adverse weather, or non-Lambertian surfaces [37, 50, 53, 57]. To address this, recent works have introduced various data synthesis and augmentation strategies to improve robustness. Image translation-based methods [15, 51, 52, 56, 67, 69] expand daytime datasets to nighttime or adverse conditions. For instance, md4all [15] employs translation networks [73] and distillation from clean images, while Tosi *et al*. [52] use diffusion models [40] with text-guided synthesis for self-distillation. D4RD [56] leverages diffusion backbones for synthetic-to-real adaptation in adverse weather. Other works propose targeted augmentations: random tone-mapping for non-Lambertian surfaces [69], realistic low-light simulation [67], and perturbation-based augmentation with spatially constrained distillation [51]. Despite these efforts, most approaches depend on a clean reference image and synthetic degradations, which limits their applicability to multi-illumination settings, where no canonical lighting condition exists and translation-based augmentations often distort true geometry.

Some works in video depth estimation [5, 23, 25, 30, 32, 38, 59, 68] encourage temporal consistency by enforcing smoothness or leveraging optical-flow-based alignment. However, these methods assume constant lighting conditions with moving cameras or dynamic objects, whereas our setting involves static scenes with varying illumination. Furthermore, video consistency metrics [32, 58, 68] depend on temporal correspondence through motion cues, which do not apply to static multi-illumination scenes. These differences separate our problem setting from video depth estimation and highlight the necessity for novel metrics and methods specifically designed for multi-illumination depth consistency.

A closely related study is [35], which investigates multi-illumination monocular depth using a synthetic RGB-D dataset rendered with fixed viewpoints under controlled lighting conditions, and improves robustness via an attention-augmented framework with a lightweight self-attention weighting block. Their direction is complementary to ours: it primarily enhances illumination robustness through supervised learning on rendered multi-illumination data, whereas our focus is adaptation in real multi-illumination settings where dense multi-illumination depth supervision is typically unavailable.

3 Method

3.1 Problem Formulation

We build on latent diffusion-based MDE methods [13, 20, 26]. Given an RGB image $\mathbf{x}$, we encode it once into a conditioning latent $\mathbf{c}$. For a depth map $\mathbf{d}$, the VAE encoder produces a depth latent $\mathbf{z}^{(0)} = \mathcal{E}(\mathbf{d})$. The forward diffusion perturbs $\mathbf{z}^{(0)}$ into $\mathbf{z}^{(t)}$ at timestep t. Following Lotus [20], we adopt an $\mathbf{x}_0$-prediction parameterization. U-Net f_θ is trained with:

$$\mathcal{L}_{\mathrm{GT}} = \mathbb{E}_{\mathbf{z}^{(0)},t,\boldsymbol{\epsilon}} \left\| \mathbf{z}^{(0)} - f_\theta(\mathbf{z}^{(t)}, \mathbf{c}, t) \right\|_2^2 . \tag{1}$$

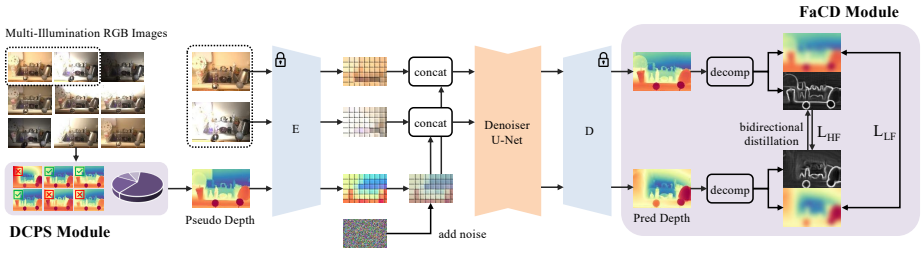

Fig. 3: Overview of LumiDepth training. Given an RGB-only multi-illumination group, we first obtain multiple depth hypotheses and construct pseudo supervision with **DCPS**. We then train the diffusion-based depth predictor with **FaCD** that improves cross-lighting stability while preserving geometry by enforcing low-frequency agreement and performing bidirectional distillation of high-frequency structures.

The Supplementary Material details the forward process. At inference, starting from $\mathbf{z}^{(T)} \sim \mathcal{N}(0, \mathbf{I})$, a sampler produces $\hat{\mathbf{z}}^{(0)}$ and the final depth $\hat{\mathbf{d}} = \mathcal{D}(\hat{\mathbf{z}}^{(0)})$.

Motivated by the limited availability of multi-illumination depth supervision, we consider two training sources: (i) conventional RGB-D pairs for $\mathcal{L}_{\text{GT}}$, and (ii) RGB-only multi-illumination groups $\mathcal{G} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of the same static scene. Our target is to train a predictor $f(\cdot)$ that estimates both illumination-invariant and geometrically correct depth:

$$f(\mathbf{x}_i) \;\approx\; f(\mathbf{x}_j) \;\approx\; \mathbf{d}^{\text{gt}}, \quad \forall i, j, \tag{2}$$

where $\mathbf{d}^{\text{gt}}$ is available only during evaluation.

In the following sections, we instantiate this objective by constructing target-domain label-free supervision from cross-illumination relations within each group and enforcing geometry-preserving paired consistency during training.

3.2 Disagreement-Calibrated Probabilistic Pseudo Supervision (DCPS)

Fig. 3 overviews our training pipeline. Given $\mathcal{G}$, we first obtain N depth predictions $\{\mathbf{d}_1, \ldots, \mathbf{d}_N\}$ [3] using a pretrained depth model (i.e., the same model being adapted for fair and controlled comparison). For each hypothesis $\mathbf{d}_i$, we quantify its reliability by measuring how much it deviates from its peers: $v_i = \frac{1}{N-1} \sum_{j \neq i} \|\mathbf{d}_i - \mathbf{d}_j\|_2^2$, where v_i serves as a disagreement-based uncertainty score. Intuitively, predictions that are geometrically stable across illuminations yield lower disagreement.

A naive pseudo label choice such as mean/median aggregation mixes in unreliable hypotheses and often blurs structures, while selecting a single "best" candidate can be brittle and may overfit to illumination-specific artifacts (Tab. 6, Fig. 4). To remove clear outliers without introducing fragile thresholds, we adopt

[3] Unless stated otherwise, $\mathbf{d}$ denotes normalized depth.

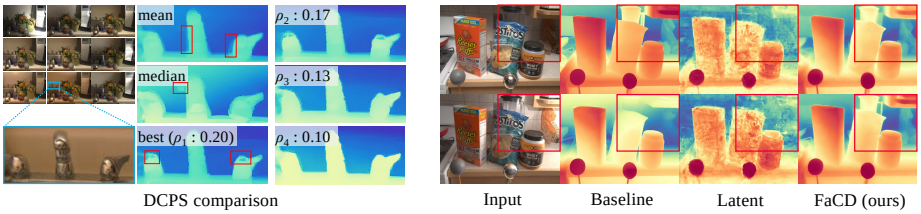

Fig. 4: Left: pseudo label construction from a multi-illumination group. Mean/median aggregation mixes unreliable hypotheses and blurs structures, while selecting a single best candidate is brittle and can inherit illumination-specific artifacts (see Sec. 3.2). **Right**: direct latent-space matching ("Latent") tends to over-regularize and suppress fine geometry, whereas FaCD aligns low-frequency shape while distilling high-frequency edges to preserve sharp structures (see Sec. 3.3 and Tab. 7).

a distribution-adaptive filtering rule based on the per-sample variance distribution: $\tau = \text{median}(\{v_i\}_{i=1}^{N})$. We keep candidates with $v_i \leq \tau$. This guarantees a non-empty candidate set (at least $N/2$ samples) and is stable to scene-dependent disagreement scales.

After median thresholding, we retain K candidates and re-index them as $\{\mathbf{d}_1, \ldots, \mathbf{d}_K\}$, with corresponding uncertainties $\{u_1, \ldots, u_K\}$. We convert uncertainties into normalized confidence weights via a temperature-scaled softmax:

$$\rho_k = \frac{\exp(-u_k/T)}{\sum_{j=1}^{K} \exp(-u_j/T)}, \qquad k = 1, \ldots, K, \tag{3}$$

where smaller u_k yields larger ρ_k. We then sample one candidate to construct the pseudo label:

$$\mathbf{d}^{\text{pseudo}} = \mathbf{d}_{k^*}, \quad k^* \sim \text{Categorical}(\rho_1, \ldots, \rho_K). \tag{4}$$

This stochastic, confidence-weighted supervision acts as an implicit ensemble over reliable illumination-conditional hypotheses across training iterations, while avoiding the bias accumulation and edge blurring typical of deterministic averaging. The reliability signal is most informative when each group contains sufficiently diverse, spatially varying illumination; we provide qualitative examples and further discussion in the Supplementary Material.

For each multi-illumination training sample, we encode $\mathbf{d}^{\text{pseudo}}$ into latent space $\mathbf{z}^{(0),\text{pseudo}}$ and supervise with the reconstruction objective:

$$\mathcal{L}_{\text{pseudo}} = \mathbb{E}_{\mathbf{z},t,\boldsymbol{\epsilon}} \left\| \mathbf{z}^{(0),\text{pseudo}} - f_\theta(\mathbf{z}^{(t),\text{pseudo}}, \mathcal{E}(\mathbf{x}), t) \right\|_2^2, \quad \mathbf{x} \in \mathcal{G}. \tag{5}$$

3.3 Frequency-aware Consistency and Distillation

Beyond pseudo supervision, we apply cross-illumination constraints to explicitly suppress illumination-induced depth drift while preserving sharp structures.

A straightforward consistency objective, e.g. $\|\hat{\mathbf{z}}_1 - \hat{\mathbf{z}}_2\|_2^2$ [4] (or its decoded-depth counterpart), often over-regularizes the predictions: illumination-specific appearance changes can be mistakenly treated as geometric disagreement, and the easiest way to satisfy the constraint is to attenuate high-frequency content (edges, thin structures), producing blurred depth (Fig. 4). We address this by (i) enforcing agreement only on low-frequency geometry, and (ii) transferring high-frequency details via a stop-gradient distillation that avoids coupled collapse.

We first decode the predicted latents into depth space using the VAE decoder:

$$\hat{\mathbf{d}}_1 = \mathcal{D}(\hat{\mathbf{z}}_1), \qquad \hat{\mathbf{d}}_2 = \mathcal{D}(\hat{\mathbf{z}}_2). \tag{6}$$

Let $\mathcal{LPF}(\cdot)$ be a low-pass operator. We define low-frequency consistency as:

$$\mathcal{L}_{\text{LF}} = \left\| \mathcal{LPF}(\hat{\mathbf{d}}_1) - \mathcal{LPF}(\hat{\mathbf{d}}_2) \right\|_1, \tag{7}$$

which aligns global shape while leaving local discontinuities unconstrained. We enforce FaCD in the decoded depth domain instead of the latent domain for stability and interpretability.

We then distill high-frequency structure using depth gradients (instead of image gradients to reduce illumination sensitivity; see the Supplementary Material for details). We define edge magnitude maps $\mathbf{g}_i = \|\nabla \hat{\mathbf{d}}_i\|$ and edge masks $\mathbf{M}_i = \mathbf{1}(\mathbf{g}_i > \tau_d)$. The high-frequency objective is:

$$\mathcal{L}_{\text{HF}} = \rho_2 \left\| \nabla \hat{\mathbf{d}}_1 - sg[\nabla \hat{\mathbf{d}}_2] \right\|_{1,\mathbf{M}_2} + \rho_1 \left\| \nabla \hat{\mathbf{d}}_2 - sg[\nabla \hat{\mathbf{d}}_1] \right\|_{1,\mathbf{M}_1}, \tag{8}$$

where $sg[\cdot]$ stops gradient flow [7]. More reliable predictions provide stronger teacher signals, which mitigates error reinforcement when one lighting condition yields corrupted edges. This bi-directional stop-gradient distillation distinguishes our method from prior approaches [6,43] that enforce strict agreement.

Furthermore, to prevent bias accumulation from pseudo labels, we include the standard diffusion reconstruction loss on RGB-D pairs with ground truth depth captured under uniform lighting (Eq. (1)). The overall objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GT}} + \lambda_1 \mathcal{L}_{\text{pseudo}} + \lambda_2 \mathcal{L}_{\text{LF}} + \lambda_3 \mathcal{L}_{\text{HF}}, \tag{9}$$

where we set λ_1=0.5, λ_2=0.5, and λ_3=0.1 in all experiments.

4 Experiments

4.1 Implementation Details

LumiDepth is built on the Stable Diffusion v2 [45] architecture. We initialize the multi-step and single-step variants from Marigold [26] and Lotus [20] checkpoints, respectively. During training, we freeze the latent encoder, decoder, and the first two downsampling blocks of the denoiser U-Net. We use a batch size of

[4] We denote $\hat{\mathbf{z}}_i$ as the predicted latent under pseudo supervision at timestep t.

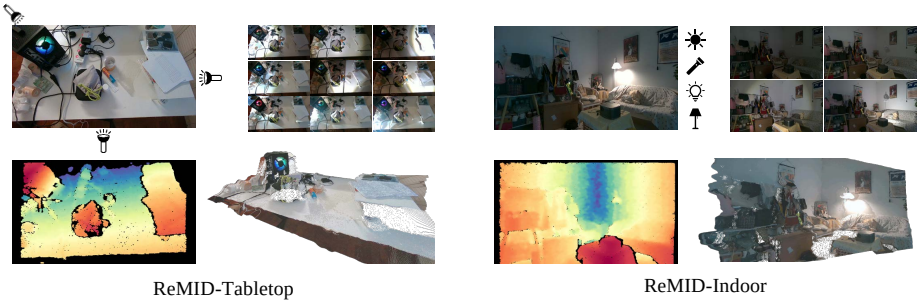

Fig. 5: Samples from the ReMID dataset, containing real-world multi-illumination RGB-D data. See the Supplementary Material for more details.

16 with gradient accumulation to enable training on a single NVIDIA RTX 4090 GPU. We use a Gaussian filter (kernel size 11, $\sigma = 10$) as the low-pass filter, and generate edge masks by suppressing gradients below a threshold τ_d of 0.05. Additional details are provided in the Supplementary Material.

4.2 Datasets

Dataset statistics are shown in Tab. 2. Detailed statistics and preprocessing steps are provided in the Supplementary Material.

In training, for RGB-D datasets, we follow the same setup as baselines and do not include any additional ground-truth depth information beyond Hypersim [44] and Virtual KITTI [3]. For multi-illumination scenes without depth annotations, we use MIIW [41], LSMI [27] and KITTI-C [29].

In evaluation, to bridge the gap between standard depth estimation and multi-illumination scenarios, we introduce *ReMID*, a new evaluation set containing 1000 samples from 100 scenes under diverse lighting conditions with dense depth annotations. As illustrated in Fig. 5, each scene is captured under several lighting variants (*e.g.*, sunlight, room light, torchlight) using a RealSense camera for dense RGB-D information. We evaluate zero-shot performance under nighttime conditions using NuScenes-Night [4] and RobotCar-Night [39]. In-domain results on MIIW and LSMI are provided in the Supplementary Material. For completeness, we also report zero-shot performance on standard depth estimation benchmarks.

We also considered the synthetic multi-illumination dataset released with [35]. However, the public version available at the time of our experiments contains only two synthetic scenes and does not provide an official training/evaluation protocol or public checkpoints, which limits diversity and makes benchmarking unreliable. We therefore do not include it in this paper.

Table 1: Quantitative comparison of affine-invariant depth estimation on **zero-shot** multi-illumination and nighttime datasets. We compare against both general and robust MDE methods. § denotes methods leverage DA series prior. Best and second best results are indicated by **bold** and underline. *st* denotes direct self-training on RGB-only data; see Sec. 4.4 for details. Our LumiDepth achieves state-of-the-art overall performance without relying on any additional ground truth depth supervision.

Method	ReMID				NuScenes-N		RobotCar-N		**Avg.**
	MeanDV↓	MaxPIV↓	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	**Rank↓**
general MDE methods									
Pixel-Perfect-Depth [61]§	0.068	0.120	0.150	0.751	0.275	0.552	0.238	0.675	11.0
DepthMaster [49]§	0.059	0.092	0.142	0.803	0.252	0.646	0.233	0.582	8.5
DAv2 [65]§	0.051	0.076	0.119	0.857	0.262	0.707	0.242	0.514	7.5
DAv3 [36]§	0.049	0.075	0.118	0.862	0.270	0.695	0.235	0.520	6.9
GenPercept [62]	0.052	0.088	0.111	0.880	0.256	0.584	0.228	0.663	5.8
DistillAnyDepth [22]§	0.047	0.074	0.147	0.835	**0.220**	0.718	0.233	0.527	5.8
GeoWizard [12]	0.045	0.072	0.110	0.879	0.385	0.381	0.219	**0.683**	5.6
robust MDE methods									
Robust-Depth [47]	0.081	0.130	0.195	0.710	0.301	0.583	0.508	0.510	15.0
WeatherDepth [55]	0.063	0.100	0.178	0.733	0.292	0.602	0.469	0.512	13.2
D4RD [56]	0.058	0.088	0.149	0.764	0.276	0.638	0.434	0.525	11.0
DA-AC [51]§	0.047	0.080	0.117	0.876	0.241	**0.719**	0.227	0.557	4.8
Marigold [26] (*st*)	0.109	0.198	0.197	0.702	0.315	0.498	0.282	0.588	15.5
Marigold [26]	0.101	0.188	0.181	0.726	0.308	0.506	0.267	0.605	14.0
LumiDepth (n-step)	0.057	0.097	0.138	0.825	0.274	0.585	0.254	0.632	9.3
Improvement (%)	43.6	48.4	23.8	13.6	11.0	15.6	4.9	4.5	33.6
Lotus [20] (*st*)	0.048	0.085	0.120	0.835	0.306	0.554	0.289	0.632	9.6
Lotus [20]	0.043	0.074	0.113	0.850	0.280	0.580	0.245	0.656	6.7
LumiDepth (1-step)	**0.030**	**0.055**	**0.085**	**0.914**	0.223	0.636	**0.217**	0.676	**2.0**
Improvement (%)	30.2	25.7	24.8	7.5	20.3	9.7	11.4	3.0	70.1

4.3 Evaluation Metrics

We report AbsRel↓ and δ_1 ↑ on datasets with ground truth. To measure stability across N illumination conditions, we define a per-illumination deviation

$$v_i = \text{mean}_p \frac{\left|\tilde{\mathbf{d}}_i(p) - \bar{\mathbf{d}}(p)\right|}{\bar{\mathbf{d}}(p) + \epsilon}, \quad \bar{\mathbf{d}}(p) = \tfrac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{d}}_i(p). \tag{10}$$

where p is the pixel index and each $\tilde{\mathbf{d}}_i$ is aligned to ground truth. We then use mean depth variation **MeanDV**$= \text{mean}_i\, v_i$ to summarize the *average* stability, and max per-image variation **MaxPIV**$= \max_i v_i$ to capture the *extreme-case* (worst-illumination) stability (lower is better). See the Supplementary Material for details and discussion.

4.4 Results

Evaluation on Multi-Illumination Datasets Tab. 1 reports our *zero-shot* performance on the multi-illumination ReMID benchmark and two nighttime

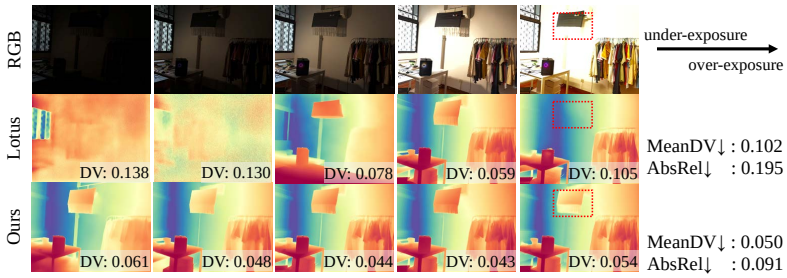

Fig. 6: Zero-shot results on ReMID dataset, covering extreme illumination conditions (leftmost and rightmost). Our method not only improves overall accuracy but also substantially reduces cross-illumination depth variation, yielding more stable predictions.

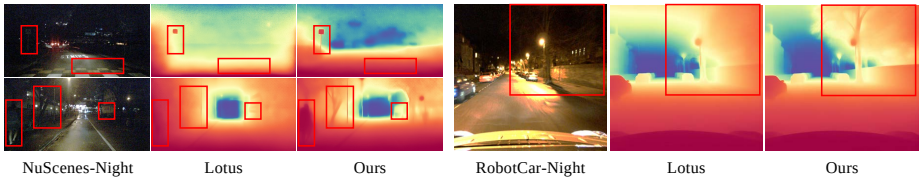

Fig. 7: Zero-shot results on NuScenes-Night [4] and RobotCar-Night [39].

driving datasets (NuScenes-Night and RobotCar-Night). We achieve state-of-the-art overall results, consistently improving both depth accuracy (AbsRel/δ_1) and illumination stability (MeanDV/MaxPIV), *without* using any additional ground-truth depth supervision during training. In particular, compared with Lotus [20], our single-step model reduces MeanDV from 0.043 to 0.030 (-30.2%) and AbsRel from 0.113 to 0.085 (-24.8%) on ReMID, while improving δ_1 from 0.850 to 0.914 ($+7.5\%$). As an additional baseline, we include a direct self-training (*st*) variant that uses 14K RGB-only multi-illumination data, but removes our DCPS and FaCD modules. This comparison isolates the effect of simply scaling RGB-only adaptation from our consistency design. Notably, LumiDepth is also competitive with methods explicitly tailored for nighttime robustness, despite those approaches typically relying on large labeled datasets, whereas our framework remains label-free for the target domains. Overall, results demonstrate strong generalization to complex and unseen lighting conditions.

Fig. 1, Fig. 6, and Fig. 7 provide qualitative comparisons. Across diverse lighting conditions, LumiDepth consistently yields illumination-invariant and structurally coherent depth maps. In particular, Fig. 6 stresses *extreme* illumination changes in ReMID, where strong specularities, cast shadows, and low-light regions often induce appearance-driven failures, while Fig. 7 highlights robustness in real nighttime driving scenes. In contrast, recent depth foundation models [20,26] tend to misinterpret illumination artifacts as geometric cues, leading to distorted geometry and missing structures. Additional qualitative results are provided in the Supplementary Material.

Table 2: Main datasets used in our study. Our real-world ReMID dataset bridges the depth (D) and multi-illumination (MI) gap to enable comprehensive evaluation. *No additional depth supervision* beyond the baseline setup is used.

Dataset	D	MI	Domain	Train	Eval
Hypersim [44]	✓	✗	Syn.	54K	0
VKITTI [3]	✓	✗	Syn.	20K	0
NYUv2 [48]	✓	✗	Real	0	654
KITTI [16]	✓	✗	Real	0	697
NuScenes-N [4]	✓	✗	Real	0	500
RobotCar-N [39]	✓	✗	Real	0	186
KITTI-C [29]	✗	✓	Syn.	3K	0
MIIW [41]	✗	✓	Real	10K	300
LSMI [27]	✗	✓	Real	1K	148
ReMID (ours)	✓	✓	Real	0	1000

Table 3: Quantitative comparison on zero-shot general depth benchmarks (AbsRel↓). † denotes RGB-only multi-illumination data.

Method	Data	NYUv2	KITTI	ETH3D	DIODE
discriminative					
MiDaS [43]	2M	0.111	0.236	0.184	0.332
Omnidata [9]	12.2M	0.074	0.149	0.166	0.339
DPT [42]	1.4M	0.098	0.100	**0.078**	**0.182**
DA [64]	62.6M	**0.043**	0.076	0.127	0.260
DAv2 [65]	62.6M	0.045	**0.074**	0.131	0.265
generative					
GeoWizard [12]	280K	**0.052**	0.097	0.064	0.297
DepthFM [19]	74K	0.060	0.091	0.065	**0.224**
GenPercept [62]	90K	0.054	0.102	0.071	0.312
E2E FT [13]	74K	**0.052**	0.096	0.062	0.302
Marigold [26]	74K	0.055	0.099	0.065	0.308
Lotus [20]	59K	0.054	0.085	**0.059**	0.229
Lotus [20] (*st*)	59K+14K†	0.057	0.097	0.063	0.278
LumiDepth	59K+14K†	**0.052**	**0.079**	**0.059**	0.230

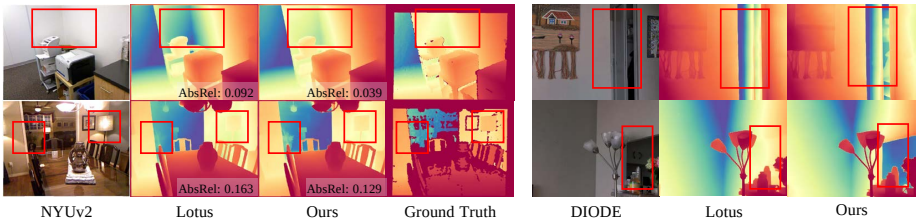

Fig. 8: Zero-shot results on general depth benchmarks [48, 54]. Our method produces sharper and more accurate depth maps compared to existing approaches.

Evaluation on General RGB-D Benchmarks We further evaluate on general RGB-D benchmarks (Tab. 3) to verify that LumiDepth generalizes beyond multi-illumination scenes to standard uniform-lighting settings. Following [13,20,26,62], we compare against both generative and discriminative models. Generative approaches (diffusion-based) synthesize depth via conditional denoising and often transfer well without large-scale retraining, whereas discriminative models directly regress depth and typically rely on extensive labeled supervision for competitive performance. Overall, results show that LumiDepth maintains strong performance on general benchmarks while consistently outperforming the direct *st* baseline, confirming that the gains are not attributable to naive self-training alone. Beyond the quantitative results, Fig. 8 shows that LumiDepth produces sharper depth boundaries and cleaner structures, while avoiding artifacts around saturated highlights and preserving accurate depth in underexposed regions.

Table 4: Backbone-agnostic plug-in on discriminative method.

Method	MeanDV↓	AbsRel↓
DAv2 [65]	0.051	0.119
DAv2+FaCD	0.048	0.112
DAv2+DCPS	0.040	0.105
DAv2+DCPS+FaCD	0.036	0.102
Ours	**0.030**	**0.085**

Table 5: Quantitative results on RoboDepth [29]. We report the average over all 18 corruption types.

Method	MeanDV↓	MaxPIV↓	AbsRel↓	δ_1 ↑
DAv2 [65]	0.056	0.094	0.138	0.837
DA-AC [51]	0.048	0.089	0.130	0.845
Lotus [20]	0.051	0.103	0.135	0.842
Ours	**0.038**	**0.072**	**0.115**	**0.853**

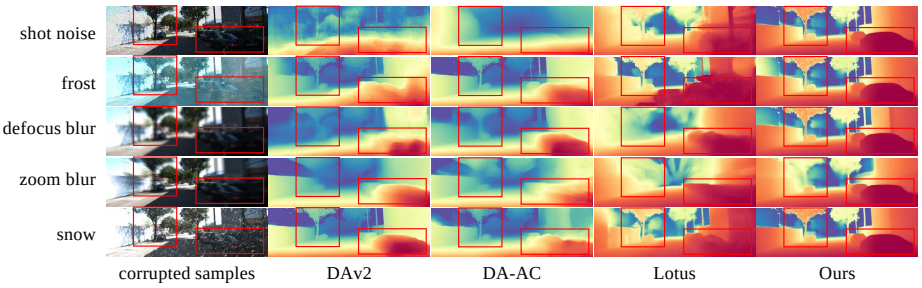

Fig. 9: Qualitative results on non-illumination corruptions from RoboDepth [29].

Backbone-agnostic plug-in. Our modules are model-agnostic and can be integrated into both generative and discriminative predictors. To verify this, we apply them to DAv2 and observe consistent improvements (Tab. 4). In this work, we primarily instantiate LumiDepth on a diffusion pipeline, as our experiments indicate that the iterative denoising process tends to better mitigate appearance-driven artifacts under illumination changes than direct image-to-depth regression in a discriminative backbone under comparable adaptation settings.

Beyond multi-illumination: other consistency-critical scenarios. Although multi-illumination serves as our primary setting, the same consistency challenge appears whenever the scene geometry remains fixed while the visual evidence changes across conditions (e.g., weather effects, motion-induced artifacts, or processing distortions). Importantly, our modules are not illumination-specific. They only assume access to paired or grouped observations of the same scene with different appearances, and can therefore transfer to other consistency-critical depth scenarios with the same supervision-free setup, as shown quantitatively in Tab. 5 and qualitatively in Fig. 9. Additional implementation details and extended discussion are provided in the Supplementary Material.

4.5 Ablation Studies

Table 6 compares DCPS with three alternative pseudo-label aggregation strategies. *Mean* and *Median* average hypotheses from the pretrained model and therefore inevitably absorb unreliable outliers, yielding supervision that is less accurate across illuminations. *Best* selects the single hypothesis with the lowest

Table 6: Ablation on DCPS selection. *Mean* and *Median* replace DCPS by using the mean or median of pretrained-model predictions, respectively, while *Best* selects the prediction with the minimum cross-illumination variance (Fig. 4).

Selection	ReMID		NuScenes-N	NYUv2
	AbsRel↓	δ_1 ↑	AbsRel↓	AbsRel↓
Mean	0.092	0.909	0.265	0.058
Median	0.091	0.911	0.262	0.057
Best	0.088	0.912	0.245	0.056
DCPS	**0.085**	**0.914**	**0.223**	**0.052**

Table 7: Ablation on FaCD. *Latent* and *Depth* indicate enforcing cross-illumination consistency in latent or depth space (Eq. (6)). This table uses n-step training for complementary analysis.

Variant	ReMID		NuScenes-N	NYUv2
	AbsRel↓	δ_1 ↑	AbsRel↓	AbsRel↓
Latent	0.365	0.629	0.738	0.179
Depth	0.274	0.692	0.520	0.127
$\mathcal{L}_{\text{LF}}$	0.145	0.818	0.282	0.057
$\mathcal{L}_{\text{HF}}$	0.170	0.766	0.289	0.056
FaCD	**0.138**	**0.825**	**0.274**	**0.054**

Table 8: Component-removal ablation on the final training objective.

Selection	ReMID				NuScenes-N		NYUv2	
	MeanDV↓	MaxPIV↓	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
FULL	0.030	0.055	**0.085**	**0.914**	**0.223**	**0.636**	**0.052**	**0.965**
– $\mathcal{L}_{\text{GT}}$	**0.029**	**0.053**	0.092	0.904	0.264	0.615	0.057	0.954
– $\mathcal{L}_{\text{pseudo}}$	0.035	0.062	0.090	0.909	0.262	0.618	0.056	0.962
– ($\mathcal{L}_{\text{LF}}$+$\mathcal{L}_{\text{HF}}$)	0.042	0.076	0.096	0.899	0.258	0.621	0.053	0.962

cross-illumination disagreement but is brittle: it may lock onto an overconfident candidate that matches illumination-specific appearances rather than true geometry (Fig. 4). In contrast, DCPS performs confidence-guided *probabilistic* selection, assigning higher sampling probability to low-uncertainty candidates while suppressing highly uncertain ones. This preserves multiple plausible hypotheses during training, improving robustness and reducing the risk of drifting toward illumination-biased pseudo supervision.

Table 7 presents results when consistency is enforced without full FaCD. Applying a naive consistency loss directly in the latent or depth domain indeed enhances stability but severely impairs geometric accuracy (Fig. 4). This confirms that frequency-agnostic constraints encourage over-smoothing and suppress important edge details. By decoupling low-frequency geometric alignment and high-frequency edge distillation, FaCD preserves fine structural cues while maintaining cross-illumination consistency.

Table 8 analyzes the contribution of each loss component. Removing $\mathcal{L}_{\text{GT}}$ increases bias toward pseudo labels, showing that this term mainly serves to calibrate and stabilize training. Excluding $\mathcal{L}_{\text{pseudo}}$ leads to simultaneous drops in accuracy and stability, highlighting that DCPS pseudo labels offer illumination-adaptive supervision complementary to $\mathcal{L}_{\text{GT}}$. Eliminating frequency-based consistency terms notably degrades ReMID robustness, confirming FaCD's unique role in maintaining geometric and edge fidelity under diverse lighting conditions.

Additional experiments on other components, including hyperparameter selection, edge mask, frequency operators, gradient stoppage, dataset configurations, and freezing network layers, are provided in the Supplementary Material.

5 Conclusion

We presented **LumiDepth** that improves both *accuracy* and *cross-illumination consistency* for monocular depth under severe appearance shifts. Given only RGB observations of the same static scene captured under different lighting, **DCPS** builds reliable pseudo supervision by leveraging cross-illumination disagreement to select high-confidence hypotheses, while **FaCD** enforces stability by aligning low-frequency geometry and distilling reliable high-frequency details without over-smoothing. We also introduce **ReMID**, a real-world multi-illumination benchmark, together with stability metrics that quantify average and worst-case depth variation across illuminations. Extensive experiments show consistent gains across diverse illumination and corruption settings, and the same formulation can be applied to other consistency-critical scenarios with paired or grouped observations but limited depth supervision.

References

1. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4009–4018 (2021)
2. Bhattad, A., Soole, J., Forsyth, D.A.: Stylitgan: Image-based relighting via latent control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4231–4240 (2024)
3. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
5. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22831–22840 (2025)
6. Chen, X., Li, T.H., Zhang, R., Li, G.: Frequency-aware self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5808–5817 (2023)
7. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
8. Cheng, A., Yang, Z., Zhu, H., Mao, K.: Gam-depth: self-supervised indoor depth estimation leveraging a gradient-aware mask and semantic constraints. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5367–5374. IEEE (2024)
9. Eftekhar, A., Sax, A., Malik, J., Zamir, A.: 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. pp. 10786–10796 (2021)
10. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)

11. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2002–2011 (2018)
12. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In: European Conference on Computer Vision. pp. 241–258. Springer (2024)
13. Garcia, G.M., Abou Zeid, K., Schmidt, C., De Geus, D., Hermans, A., Leibe, B.: Fine-tuning image-conditional diffusion models is easier than you think. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 753–762. IEEE (2025)
14. Garg, R., Bg, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European conference on computer vision. pp. 740–756. Springer (2016)
15. Gasperini, S., Morbitzer, N., Jung, H., Navab, N., Tombari, F.: Robust monocular depth estimation under challenging conditions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8177–8186 (2023)
16. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)
17. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
18. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019)
19. Gui, M., Schusterbauer, J., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B.: Depthfm: Fast generative monocular depth estimation with flow matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3203–3211 (2025)
20. He, J., Li, H., Yin, W., Liang, Y., Li, L., Zhou, K., Zhang, H., Liu, B., Chen, Y.: Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In: International Conference on Learning Representations. vol. 2025, pp. 89454–89467 (2025)
21. He, K., Liang, R., Munkberg, J., Hasselgren, J., Vijaykumar, N., Keller, A., Fidler, S., Gilitschenski, I., Gojcic, Z., Wang, Z.: Unirelight: Learning joint decomposition and synthesis for video relighting. arXiv preprint arXiv:2506.15673 (2025)
22. He, X., Guo, D., Li, H., Cui, Y., Weng, L., Li, R., Zhang, C.: Distill any depth: Distillation creates a stronger monocular depth estimator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 591–601 (2026)
23. Hu, W., Gao, X., Li, X., Zhao, S., Cun, X., Zhang, Y., Quan, L., Shan, Y.: Depthcrafter: Generating consistent long depth sequences for open-world videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2005–2015 (2025)
24. Ji, P., Li, R., Bhanu, B., Xu, Y.: Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12787–12796 (2021)
25. Ke, B., Narnhofer, D., Huang, S., Ke, L., Peters, T., Fragkiadaki, K., Obukhov, A., Schindler, K.: Video depth without video models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7233–7243 (2025)

26. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9492–9502 (2024)
27. Kim, D., Kim, J., Nam, S., Lee, D., Lee, Y., Kang, N., Lee, H.E., Yoo, B., Han, J.J., Kim, S.J.: Large scale multi-illuminant (lsmi) dataset for developing white balance algorithm under mixed illumination. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2410–2419 (2021)
28. Kocsis, P., Philip, J., Sunkavalli, K., Nießner, M., Hold-Geoffroy, Y.: Lightit: Illumination modeling and control for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9359–9369 (2024)
29. Kong, L., Xie, S., Hu, H., Ng, L.X., Cottereau, B., Ooi, W.T.: Robodepth: Robust out-of-distribution depth estimation under corruptions. Advances in Neural Information Processing Systems **36**, 21298–21342 (2023)
30. Kopf, J., Rong, X., Huang, J.B.: Robust consistent video depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1611–1621 (2021)
31. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). pp. 239–248. IEEE (2016)
32. Li, S., Luo, Y., Zhu, Y., Zhao, X., Li, Y., Shan, Y.: Enforcing temporal consistency in video depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1145–1154 (2021)
33. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2041–2050 (2018)
34. Liang, R., Müller, N., Weber, E., Zauss, D., Vijaykumar, N., Kontschieder, P., Richardt, C.: Luxremix: Lighting decomposition and remixing for indoor scenes. arXiv preprint arXiv:2601.15283 (2026)
35. Liang, Y., Zhang, Z., Xian, C., He, S.: Delving into multi-illumination monocular depth estimation: A new dataset and method. IEEE Transactions on Multimedia **27**, 1018–1032 (2024)
36. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
37. Liu, L., Song, X., Wang, M., Liu, Y., Zhang, L.: Self-supervised monocular depth estimation for all day images using domain separation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12737–12746 (2021)
38. Luo, X., Huang, J.B., Szeliski, R., Matzen, K., Kopf, J.: Consistent video depth estimation. ACM Transactions on Graphics (ToG) **39**(4), 71–1 (2020)
39. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research **36**(1), 3–15 (2017)
40. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
41. Murmann, L., Gharbi, M., Aittala, M., Durand, F.: A dataset of multi-illumination images in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4080–4089 (2019)

42. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
43. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: 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)
44. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10912–10922 (2021)
45. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
46. Sarkar, A., Mai, H., Mahapatra, A., Lazebnik, S., Forsyth, D.A., Bhattad, A.: Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 28140–28149 (2024)
47. Saunders, K., Vogiatzis, G., Manso, L.J.: Self-supervised monocular depth estimation: Let's talk about the weather. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8907–8917 (2023)
48. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)
49. Song, Z., Wang, Z., Li, B., Zhang, H., Zhu, R., Liu, L., Jiang, P.T., Zhang, T.: Depthmaster: Taming diffusion models for monocular depth estimation. IEEE Transactions on Circuits and Systems for Video Technology (2026)
50. Spencer, J., Bowden, R., Hadfield, S.: Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14402–14413 (2020)
51. Sun, B., Jin, M., Yin, B., Hou, Q.: Depth anything at any condition. arXiv preprint arXiv:2507.01634 (2025)
52. Tosi, F., Ramirez, P.Z., Poggi, M.: Diffusion models for monocular depth estimation: Overcoming challenging conditions. In: European Conference on Computer Vision. pp. 236–257. Springer (2024)
53. Vankadari, M., Garg, S., Majumder, A., Kumar, S., Behera, A.: Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In: European Conference on Computer Vision. pp. 443–459. Springer (2020)
54. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463 (2019)
55. Wang, J., Lin, C., Nie, L., Huang, S., Zhao, Y., Pan, X., Ai, R.: Weatherdepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 4976–4982. IEEE (2024)
56. Wang, J., Lin, C., Nie, L., Liao, K., Shao, S., Zhao, Y.: Digging into contrastive learning for robust depth estimation with diffusion models. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 4129–4137 (2024)

57. Wang, K., Zhang, Z., Yan, Z., Li, X., Xu, B., Li, J., Yang, J.: Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16055–16064 (2021)
58. Wang, Y., Pan, Z., Li, X., Cao, Z., Xian, K., Zhang, J.: Less is more: Consistent video depth estimation with masked frames modeling. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 6347–6358 (2022)
59. Wang, Y., Shi, M., Li, J., Huang, Z., Cao, Z., Zhang, J., Xian, K., Lin, G.: Neural video depth stabilizer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9466–9476 (2023)
60. Xing, X., Groh, K., Karaoglu, S., Gevers, T., Bhattad, A.: Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 442–452 (2025)
61. Xu, G., Lin, H., Luo, H., Wang, X., Yao, J., Zhu, L., Pu, Y., Chi_, C., Sun, H., Wang, B., et al.: Pixel-perfect depth with semantics-prompted diffusion transformers. Advances in Neural Information Processing Systems **38**, 174731–174755 (2026)
62. Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., Shen, C.: What matters when repurposing diffusion models for general dense perception tasks? arXiv preprint arXiv:2403.06090 (2024)
63. Yan, W., Li, M., Li, H., Shao, S., Tan, R.T.: Synthetic-to-real self-supervised robust depth estimation via learning with motion and structure priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21880–21890 (2025)
64. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
65. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
66. Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: New crfs: Neural window fully-connected crfs for monocular depth estimation. arXiv preprint arXiv:2203.01502 (2022)
67. Zeng, L., Zhu, Z., Lu, R., Lu, M., Zheng, B., Yan, C., Xue, A.: Depthdark: Robust monocular depth estimation for low-light environments. arXiv preprint arXiv:2507.18243 (2025)
68. Zhang, H., Shen, C., Li, Y., Cao, Y., Liu, Y., Yan, Y.: Exploiting temporal consistency for real-time video depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1725–1734 (2019)
69. Zhang, J., Li, J., Huang, Y., Wang, Y., Zheng, J., Shen, L., Cao, Z.: Towards robust monocular depth estimation in non-lambertian surfaces. In: European Conference on Computer Vision. pp. 175–189. Springer (2024)
70. Zhang, L., Rao, A., Agrawala, M.: Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In: The Thirteenth International Conference on Learning Representations (2025)
71. Zhang, X., Gao, W., Jain, S., Maire, M., Forsyth, D., Bhattad, A.: Latent intrinsics emerge from training to relight. Advances in Neural Information Processing Systems **37**, 96775–96796 (2024)
72. Zhao, C., Zhang, Y., Poggi, M., Tosi, F., Guo, X., Zhu, Z., Huang, G., Tang, Y., Mattoccia, S.: Monovit: Self-supervised monocular depth estimation with a vision transformer. In: 2022 international conference on 3D vision (3DV). pp. 668–678. IEEE (2022)

73. Zheng, Z., Wu, Y., Han, X., Shi, J.: Forkgan: Seeing into the rainy night. In: European conference on computer vision. pp. 155–170. Springer (2020)
74. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1851–1858 (2017)