Diffusion Models for Monocular Depth Estimation: Overcoming Challenging Conditions

Fabio Tosi, Pierluigi Zama Ramirez, and Matteo Poggi

University of Bologna, Bologna, Italy
{fabio.tosi5, pierluigi.zama, m.poggi}@unibo.it
https://diffusion4robustdepth.github.io/

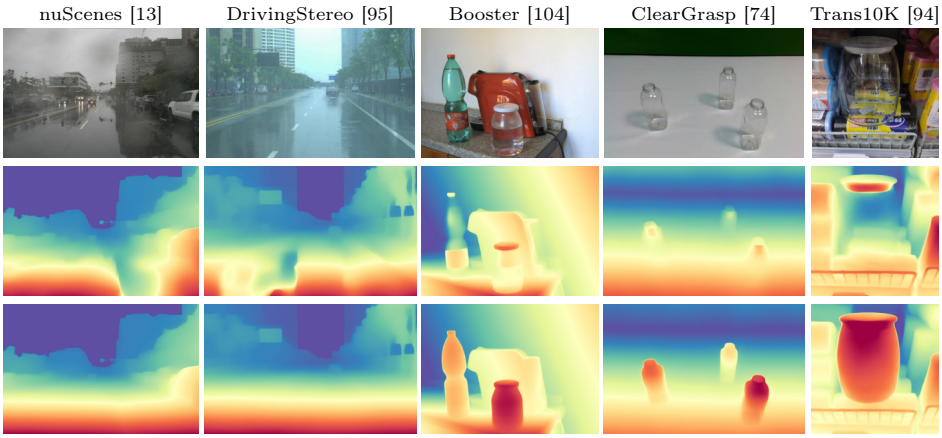


Fig. 1: Framework Results. From top to bottom: source image, depth predictions from the original Depth Anything [96], and results from our fine-tuned version.

Abstract. We present a novel approach designed to address the complexities posed by challenging, out-of-distribution data in the single-image depth estimation task. Starting with images that facilitate depth prediction due to the absence of unfavorable factors, we systematically generate new, user-defined scenes with a comprehensive set of challenges and associated depth information. This is achieved by leveraging cutting-edge text-to-image diffusion models with depth-aware control, known for synthesizing high-quality image content from textual prompts while preserving the coherence of 3D structure between generated and source imagery. Subsequent fine-tuning of any monocular depth network is carried out through a self-distillation protocol that takes into account images generated using our strategy and its own depth predictions on simple, unchallenging scenes. Experiments on benchmarks tailored for our purposes demonstrate the effectiveness and versatility of our proposal.

1 Introduction

Monocular depth estimation, a key computer vision task, has significantly advanced due to recent breakthroughs in deep learning techniques. This has wide-ranging applications, from enhancing robotics and augmented reality to improving autonomous driving safety and precision, where relying on multiple images for depth estimation may not be feasible due to resource or spatial constraints. However, while being practical, it contends with the challenge of inferring depth from a single image, a problem acknowledged for its ill-posed and severely under-constrained nature. Typically, addressing this challenge often involves training monocular depth networks through supervised methods [9, 15, 46, 61, 93, 98, 101] using annotations from active sensors or self-supervised techniques that exploit stereo image pairs [29] or monocular video sequences [112] at training time.

State-of-the-art models, such as DPT [69] and the newer Depth Anything [96], instead, combine insights from a large number of datasets, each with depth annotations extracted using different techniques. This extensive training protocol equips these models to excel in a wide range of real-world scenarios. Nevertheless, it is crucial to stress that even these models, while excelling in numerous settings, face significant challenges when dealing with data falling far from the distribution observed during training – such as, for instance, adverse conditions (*e.g.*, rain and nighttime), or objects featuring non-Lambertian surfaces. These challenges arise mainly from insufficient high-quality annotated data for robust model training, compounded by the limitations of existing vision-based depth extraction techniques as well as active sensors (*e.g.*, LiDAR, ToF, Kinect, etc.), which struggle in complex environments such as rain, snow, or materials with specific reflectivity properties. As a result, depth estimates in such settings tend to be unreliable, yielding severe implications for subsequent applications reliant on accurate 3D information. Tipically, current approaches tend to address these challenges independently. Some focus solely on resolving the issue of poor illumination and adverse weather [26, 27, 91], while others tackle the problem of non-Lambertian surfaces [18]. These disjointed approaches underscore the need for a unified methodology - a single framework capable of addressing all adverse scenarios simultaneously, providing a more robust and general solution.

In this work, we introduce diffusion models [20, 44], originally designed for image synthesis, as a pioneering strategy to address the demanding challenges posed by those images that fall in the long tail of the data distribution usually considered to train depth estimation models.

Building upon principles of text-to-image diffusion models with multi-modal controls [56, 106], we aim to create a diverse collection of highly realistic scenes that accurately replicate the 3D structure of a specific reference setting, but are intentionally enriched with various adverse factors. Importantly, these conditions are purely arbitrary and can be customized with user-defined text prompts based on the specific application of interest.

More specifically, our approach begins by selecting images that initially depict scenes devoid of the complexities associated with adverse conditions. These samples can be obtained either from an existing real-world dataset [17, 28, 57],

through custom collections, or even generated using generative models [2, 59]. With the preselected images, we employ any readily available monocular depth estimation network to provide an initial 3D representation of the scenes. Importantly, such a model can be pre-trained on different large-scale datasets or tailored to a specific domain based on the application requirements.

Subsequently, we apply text-to-image diffusion models to transform the initial unchallenging images into more complex ones while preserving the same underlying 3D scene structure (*i.e.* depth). After combining complex and simple imagery, the pre-trained depth network used for the 3D data generation enters the fine-tuning phase. In this stage, we expose the model to the composed dataset, providing it with challenging training images and their corresponding depth maps obtained in the initial step. This fine-tuning process refines the ability of the monocular network to infer depth, enabling it to better handle adverse settings, as clearly shown in Fig. 1.

We summarize our main contributions as follows:

- We pioneer the use of diffusion models as a novel solution to address the challenges of single-image depth estimation, particularly in scenarios involving adverse weather conditions and non-Lambertian surfaces.
- By distilling the knowledge of diffusion models, our approach improves the robustness of existing monocular depth estimation models, especially in challenging out-of-distribution settings.
- Our approach tackles adverse weather and non-Lambertian challenges at once, demonstrating the potential to address multiple challenging scenarios simultaneously while achieving competitive results compared to specialized solutions [18, 27] that rely on additional training information.

2 Related work

2.1 Monocular Depth Estimation

Monocular depth estimation has undergone a profound shift with the emergence of deep learning [107], making previous traditional methodologies [36, 75, 76] outdated. Initially, CNN-based methods relied on supervised training with depth data [22, 23, 47, 48]. Subsequently, the focus shifted to self-supervised techniques that tap into diverse sources, including stereo pairs [4,16,25,29,51,62,63,65,85,92] or video sequences [11, 14, 31, 32, 53, 84, 90, 109, 112]. Within this context, several frameworks emerged, employing multi-task approaches by incorporating data like optical flow [71, 86, 100, 113], semantic segmentation [33, 45, 103], and more. Additionally, other approaches include predicting depth uncertainty [37, 64].

Simultaneously, beyond supervised LiDAR-based techniques [9, 15, 46, 61, 93, 98, 101], recent approaches have explored techniques to mix multiple datasets [10, 21, 34, 69, 70, 96, 99], each enriched with diverse annotations from stereo or multi-view stereo followed by manual post-processing operations.

Adverse Weather. Despite significant advances [81–83], existing monocular networks struggle under adverse weather conditions. DeFeatNet [80] addressed

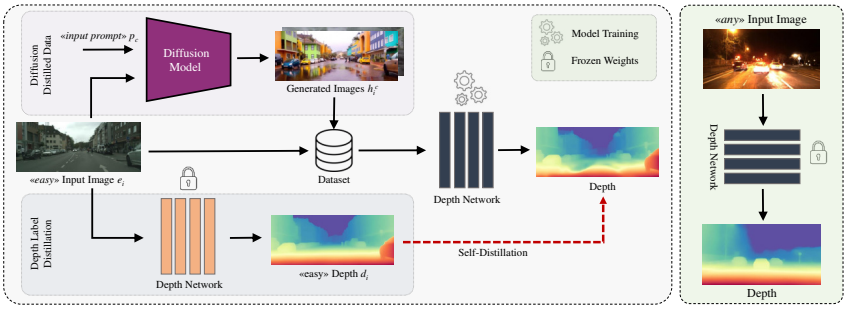


Fig. 2: Method Overview. Left: Image generation and self-distillation. *Diffusion Distilled Data* (upper): Easy image (e_i) and text prompt (p_c) input to conditional diffusion models generate adverse scenes (h_i^c). *Depth Label Distillation* (lower): Pre-trained network estimates depth (d_i) from easy image (e_i). Pairs (e_i, h_i^c) used for fine-tuning with scale-and-shift-invariant loss. **Right**: Fine-tuned network handles diverse inputs in testing, from simple to complex scenarios.

low visibility, but some had dedicated day-night branches based on GAN [87,108], used extra sensors such as radar [26], or faced daytime trade-offs [88]. Adverse weather like rain also posed problems, with few solutions needing separate encoders for each condition [108]. Recently, [27] introduced a novel GAN-based approach that addresses these issues, enabling standard models to perform robustly in diverse conditions without compromising their performance in ideal settings. Our approach, based on conditional diffusion models, relaxes the constraints of GAN-based methods as we use a single foundational model to tackle multiple challenges. This eliminates the need for separate GANs for each condition (e.g., night, rain). By combining this with text prompts, we generate potentially unlimited challenging samples from a single easy image. Furthermore, unlike other GAN-based methods that require paired easy/challenging samples, our approach needs no prior knowledge or real challenging samples beforehand.

Transparent and Specular Surfaces. Estimating depth from a single image for transparent or mirror (ToM) surfaces presents a unique and complex challenge [67, 68, 102]. To our knowledge, Costanzino et al. [18] offer the only dedicated approach to this problem. While bypassing ground truth depth, their method relies on segmentation maps or pre-trained semantic networks specialized for these materials. They generate pseudo-labels by inpainting ToM objects in images and processing them with a pre-trained monocular depth model [70]. These labels then enable fine-tuning of existing monocular or stereo networks to effectively handle challenging non-Lambertian surfaces.

2.2 Image Diffusion

Image Diffusion Models (IDMs), initially introduced by Sohl-Dickstein et al. [79], have gained widespread adoption in image generation [20, 44]. Subsequently, numerous enhancements have been proposed, improving both computational effi-

ciency [72] and generation conditioning [3, 58]. Latent Diffusion Models (LDMs) [72] have notably reduced computational costs by incorporating denoising in the latent space. In terms of scale, Stable Diffusion [1, 2] represents a large-scale implementation of LDMs. Notably, common conditioning techniques involve cross-attention [6, 12, 24, 35, 41, 43, 58, 60, 66], and the encoding of segmentation masks into tokens [5, 24]. Moreover, various conditioning schemes have been proposed to enable the generation of visual data conditioned by diverse factors such as text, images, semantic maps, sketches, and other representations [7, 8, 39, 56, 89, 106]. In addition to image generation, diffusion models have exhibited remarkable capabilities in optical flow and monocular depth estimation [42, 77, 78]. Our approach stands out for using established conditioned diffusion models to tackle diverse challenges, including rain, night scenes, and non-Lambertian surfaces.

3 Method

This section overviews our framework, illustrated in Fig. 2, to improve monocular depth estimation in adverse settings. Assuming the absence of images depicting both *easy* and *challenging* conditions in a domain, our approach converts *easy* samples to *challenging* ones using diffusion models with depth-aware control. Subsequently, we fine-tune a pre-trained monocular depth network through self-distillation and a scale-and-shift-invariant loss using the generated data.

3.1 Background: Diffusion Models

Diffusion models have significantly influenced generative modeling in computer vision. These probabilistic generative models exhibit a distinct capability to generate highly realistic images from random noise. This transformation involves two crucial phases—*forward* and *reverse* diffusion—formally outlined as follows:

Forward Diffusion. This phase consists of the progressive degradation of an image by adding scaled Gaussian noise. It is defined as $x_t = x_{t-1} + \epsilon_{t-1}$, where x_t represents the image at time step t, and ϵ_{t-1} denotes the noise increment from the preceding step. This process results in a progressively noised image, converging towards an isotropic Gaussian distribution.

Reverse Diffusion. Conversely, the reverse diffusion phase aims to restore the original image from its noised counterpart. For this purpose, the process begins with the noise x_t and generates progressively less noisy samples $x_{t-1}, x_{t-2}, \ldots$ until it reaches the original image x_0. A diffusion model is trained to generate x_{t-1} from x_t by predicting the noise component, denoted as ϵ. This prediction is performed by a neural network represented as $\mathcal{N}_\theta(x_t, t)$.

Through iterative application, the reverse diffusion mechanism equips the model to approximate the underlying data distribution adeptly. The prevalent architecture for noise-predicting networks involves the UNet [73] framework, trained using Mean Squared Error (MSE) loss at each temporal interval. Diffusion models are characterized by their iterative approach, which provides stability in training and generation, unlike other generative models such as GANs [19].

3.2 Conditional Diffusion Models

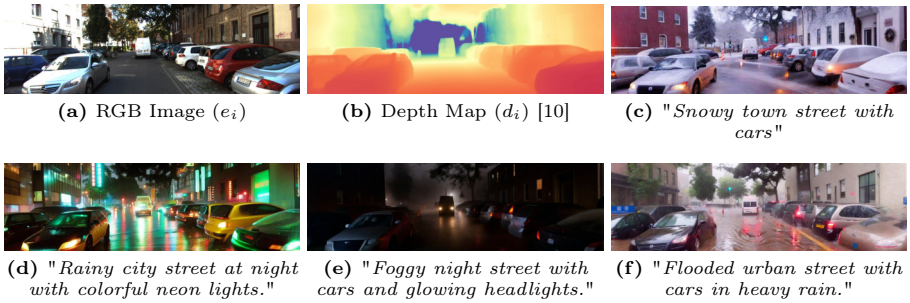
Conditional Diffusion Models (CDMs), on the other hand, transform generative modeling by incorporating various conditions for image generation, ranging from textual cues to advanced visual information such as depth maps, segmentation maps, gradients, normals, and key points. Notably, ControlNet [106], a neural network capable of learning to condition large diffusion models trained on billion samples from hundreds of times fewer data, plays a key role in enabling controllable image generation. Seamlessly integrating diverse input conditions, from traditional textual prompts to complex visual data, ControlNet comprises two sets of weights within a pretrained diffusion model: a *trainable copy* (θ_{train}) and a *locked copy* (θ_{locked}). The trainable copy adapts dynamically to task-specific datasets during learning, fine-tuning parameters based on input conditions. In contrast, the locked copy retains knowledge from generic datasets, forming a robust foundation for image generation. Central to ControlNet is the *zero convolution* layer, denoted as $\mathcal{L}_{\text{zero-conv}}$. Initialized with zero weights and biases, this layer plays a pivotal role during training. The convolution weights gradually evolve from zeros to optimized parameters, ensuring adaptability without introducing new noise to deep features.

Following the principles established by ControlNet, several other works have emerged, exploring similar approaches to enhance the controllability and flexibility of pre-trained diffusion models [38, 55, 56, 105, 110]. These studies aim to further improve the alignment between internal model knowledge and external control signals, enabling more precise and versatile image generation while maintaining the benefits of large-scale pre-training.

Leveraging these advancements in text-to-image diffusion models, we focus on generating complex scenarios from textual cues, employing 3D data information presented as depth maps derived from images devoid of challenges.

3.3 Diffusion-Distilled Data

Our core goal is to curate a training dataset that addresses the scarcity of real-world challenging data with associated depth and is explicitly designed to improve the robustness of monocular depth estimation networks in adverse scenarios. Drawing inspiration from [27], we introduce paired images, denoted as (e_i, h_i^c). In this context, e_i represents an *easy* sample belonging to the set E of images that do not pose challenges for depth estimation models. These samples depict images captured under optimal environmental conditions, making them ideal for robust training and testing of monocular networks. They may include well-lit daytime scenes with excellent visibility, as well as scenes portraying surfaces and objects with Lambertian material characteristics. On the other hand, the collection of paired *challenging* samples, denoted as $h_i^c \in H$, is meticulously designed to replicate a diverse array of adverse scenarios faithfully. Here, H represents the set of the difficult samples for a specific condition of interest ($c \in C$). The set of conditions C includes complex surfaces, non-Lambertian objects, and adverse weather conditions. Typically, training a monocular network on these



(a) RGB Image (e_i) **(b)** Depth Map (d_i) [10] **(c)** "*Snowy town street with cars*"

(d) "*Rainy city street at night with colorful neon lights.*" **(e)** "*Foggy night street with cars and glowing headlights.*" **(f)** "*Flooded urban street with cars in heavy rain.*"

Fig. 3: Generated Images – Weather Conditions. (a-b): RGB and depth maps from KITTI 2015 [54]. (c-f): images generated by a diffusion model [56], conditioned by the depth map from (b) and text prompts indicated in each subfigure.

scenarios is challenging due to their limited availability, and annotating them is very costly and often impractical, even with the use of active sensors. In response to this, we employ recently developed text-to-image diffusion models, such as ControlNet [106] and T2I-Adapter [56], guided by external textual prompts and, crucially, depth. These models systematically transform the *easy* samples e_i into *challenging* counterparts h_i^c. The transformation process involves using the e_i as input, alongside the depth map d_i from e_i, and a corresponding text prompt p_c, which essentially describes the target scenario for a specific condition. The two conditioning inputs serve distinct purposes: i) modeling tasks such as simulating day-to-night transitions or transforming opaque objects into transparent or highly reflective surfaces, and ii) preserving the inherent 3D structure of the original *easy* image e_i. This ensures that the depth d_i remains consistent between the source and generated h_i^c images. This process allows for the distillation of hard samples on which depth estimation models struggle, yet obtaining reliable depth labels for free, *i.e.* by predicting depth on *easy* images.

Playing with text-prompts, as shown in Fig. 3, allows for a coverage of a potentially infinite number of complexities, ranging from adverse weather to non-Lambertian objects and more. This process uses only diffusion model prompts to generate samples realistically depicting desired real-world conditions.

3.4 Self-Distillation Training

After generating challenging images using text-to-image diffusion models with depth-aware control, we fine-tune existing pre-trained models to improve their robustness to complex, out-of-distribution data. It is worth noting that, pre-training methods and fine-tuning models can be arbitrary, including supervised techniques using LiDAR-derived depth, photometric losses from video, stereo sequences, or other approaches. This makes our approach model agnostic and applicable to various existing and future diffusion or monocular models.

By employing distillation in a teacher-student paradigm, the pre-trained depth estimation model acts as the teacher, providing depth labels for both

easy and *challenging* images generated, which are paired. Subsequently, the student network, instantiated from the same pre-trained teacher depth estimation network, undergoes a fine-tuning process using the scale-and-shift-invariant loss [69, 70] defined as $L_{\text{ssi}}(\hat{d}, \hat{d}^*) = \frac{1}{2M} \sum_{i=1}^{M} \rho(\hat{d}_i - \hat{d}_i^*)$.

This loss function compares scaled and shifted predictions $\hat{d}$ with corresponding inverse depth labels $\hat{d}^*$ from the teacher. Here, ρ quantifies the absolute difference between predictions and provided annotations.

4 Experimental Results

4.1 Evaluation Datasets & Protocol

Autonomous Driving Datasets. Our study draws on a diverse selection of datasets for thorough evaluation. The **nuScenes** dataset [13], known for its diverse weather conditions and integration of LiDAR data, consists of 1000 scenes. We adopt the split recommended in [26, 27], which yields 15,129 training images and 6,019 validation images categorized by night and rainy weather conditions. The **RobotCar** dataset [52], which captures over 1,000 km in central Oxford, offers a collection of 20M images, LIDAR, GPS, and INS ground truth. Following [27], we use 16,563 training samples and 1,411 test images, with 709 showing nighttime scenes. We further leverage **DrivingStereo** dataset [95], originally designed for stereo but adapted for monocular research, with a specific focus on 500 frames representing rainy weather conditions. Furthermore, in our experiments, we utilize images from **KITTI 2012** [28], **KITTI 2015** [54], **Apolloscape** [40], **Mapillary** [57], and **Cityscapes** [17] for training purposes only.

Non-Lambertian Datasets. We select datasets based on their availability of ground truth depths for non-Lambertian objects. The **Booster** [104] dataset contains 228 and 191 images for training and testing, respectively. Images feature indoor scenes with non-Lambertian objects such as mirrors or glasses. The training set provides disparity and segmentation maps employed to evaluate approaches in our experiments. Each pixel in segmentation maps is categorized into 4 classes based on the type of surface material. Following [18], we define 2 macro-categories – "ToM" (Transparent or Mirror) for classes 2-3, "Other" materials for labels 0-1. The **ClearGrasp** [74] dataset, instead, comprises a synthetic and a real-world split. We use the latter in our experiments, made of 286 RGB-D images of transparent objects and their ground truth geometries, together with binary masks for ToM or Other objects.

Evaluation Metrics. Our evaluation in driving scenarios, following [27], employs three standard metrics [23]: *AbsRel*, defined as $\frac{1}{n} \sum_{ij} \left| \frac{d_{\text{gt}} - d_{\text{p}}}{d_{\text{gt}}} \right|$; $RMSE = \sqrt{\frac{1}{n} \sum_{ij} (d_{\text{gt}} - d_{\text{p}})^2}$; $\delta < \tau$ is the percentage of pixels with $\max\left(\frac{|d_{\text{p}}|}{d_{\text{gt}}}, \frac{|d_{\text{gt}}|}{d_{\text{p}}} \right) < \tau$, where n is the total number of valid depth ground truth points, while d_{gt} and d_{p} represent ground truth and predicted depths at a given pixel, respectively. We follow the metrics used in [18] on non-Lambertian datasets. We employ AbsRel, δ_τ with τ being 1.05, 1.15, and 1.25, RMSE and the mean absolute error (MAE).

Fig. 4: Generated Images – ToM Objects. From top to bottom: *easy* scenes from Stable Diffusion [2], depth from Depth Anything [96], transformed scenes using [56].

We report results on all valid pixels (*All*) or for those belonging to either ToM or other objects to assess the impact of our strategy on diverse surfaces. As the predictions by monocular networks are up to an unknown scale factor, we rescale them according to the LSE criterion from [70]. For uniformity, our experiments use the evaluation frameworks by [27] and [18], respectively. In tables, we identify the **top** and next best results across macro-categories.

4.2 Diffusion Distilled Data

To create challenging settings from *easy* images, we employ the original code and pre-trained weights provided by T2I-Adapter [56]. To ensure fairness with [27] on the nuScenes and Robotcar datasets, we exclusively generate night and rain images from the original *easy* ones. This choice is motivated by the fact that these datasets only feature challenging scenarios involving such conditions. It is important to note, however, that the diffusion model can generate data in any challenging scenario, including snow, sun glare, fog, and more.

For data synthesis, we provide the diffusion model with different, random text prompts for night and rain scenarios to ensure diverse environmental variation. Consequently, we generate a comparable number of images to [27], totaling 30,258 images for nuScenes and 17,790 images for Robotcar, which we utilize during training.

In our exploration of challenging surface materials, we also want to highlight the flexibility of diffusion models in generating such representations without relying on real RGB images. To this end, we first use Stable Diffusion [2], guided by textual prompts, to generate approximately 20K images that resemble Lambertian surfaces such as wooden bottles, ceramic vessels and so on. Then, using [56]

along with specific textual prompts designed to transform these standard materials into highly challenging ones, we modify the images to represent a range of non-Lambertian objects. An example of this process is illustrated in Fig. 4. This underlines the adaptability of our approach, and the capability of generative processes to simulate a variety of materials, even in the absence of dedicated datasets. Refer to the **supplementary material** for additional details.

4.3 Training Details

All experiments are conducted on a single 3090 NVIDIA GPU. For fair comparison, we follow the training/testing protocols and frameworks of [27] for adverse weather conditions, and [18] for transparent and reflective objects, employing the same monocular networks (md4all for [27], and DPT-Large for [18]). We integrate their codebase, substituting their datasets with ours generated using a diffusion model. Furthermore, for extensive experiments involving other networks, such as ZoeDepth [10], MiDaS [69], and DPT [70], we fine-tune for 30K iterations, with an initial learning rate of 10^{-6}, reduced to 10^{-7} after 25K iterations. For Depth Anything [96], we fine-tune for 5K iterations, reducing the learning rate at 4.5K iterations. We use a batch size of 8 by default, except for ZoeDepth, which is set to 3. The AdamW [50] optimizer is used for all networks. mages are padded, cropped, and resized to maintain 384 pixels for either the long or short side, except for Depth-Anything which uses 518 pixels, preserving aspect ratio with square cropping. We apply data augmentation techniques including color jitter, RGB shift, and horizontal flip, among others.

4.4 Adverse Weather Conditions

Improving the Baselines. Tab. 1 evaluates monocular depth networks, highlighting their performance across diverse nuScenes [13] scenarios. As baselines, we examine four state-of-the-art methods – DPT [69], MiDaS [70], ZoeDepth [10], and Depth Anything [96] – known for their strong generalization capabilities, under varying atmospheric conditions: *day-clear*, *night*, and *day-rain*.

In the table, we present the improvements that our approach yields for each pre-trained model. Specifically, we fine-tune each of the four baselines using our framework and internal protocol. For this purpose, we employ [56] and the baseline depth network itself to create challenging images for the tuning phase. Crucially, the pre-computed depth pseudo-labels of each specific baseline depth network, derived from day-clear images and adopted for their challenging counterpart, remain unchanged and are used directly to minimize loss during subsequent fine-tuning. It is worth emphasizing again that our methodology relies solely on the availability of simple daytime samples to randomly generate challenging conditions, without relying on any prior knowledge of the target image characteristics of the considered condition (e.g., rain, night). Examining the experimental results reported in Tab. 1, it is clear that despite being trained on extensive datasets to generalize across scenarios, the baselines face significant

Table 1: Evaluation of monocular networks on the nuScenes [13] validation set. Original networks [10, 69, 70, 96] versus their fine-tuned versions.

Method	*day-clear*			***night***			***day-rain***		
	absRel ↓	RMSE ↓	δ_1 ↑	absRel ↓	RMSE ↓	δ_1 ↑	absRel ↓	RMSE ↓	δ_1 ↑
MiDaS [69]	0.171	7.703	76.75	0.261	9.729	54.66	0.218	8.823	69.39
MiDaS [69] ft. Ours	**0.168**	**7.563**	**76.86**	**0.254**	**9.692**	**63.47**	**0.195**	**8.278**	**72.37**
DPT [70]	0.189	8.094	75.39	0.354	12.875	60.97	0.237	8.780	66.96
DPT [70] ft. Ours	**0.184**	**7.839**	**75.50**	**0.224**	**8.375**	**67.87**	**0.199**	**8.079**	**72.85**
Depth Anything [96]	0.137	7.063	82.23	0.291	11.804	67.10	0.167	7.867	75.17
Depth Anything [96] ft. Ours	**0.134**	**6.792**	**82.53**	**0.219**	**9.140**	**70.26**	**0.157**	**7.570**	**77.42**
ZoeDepth [10]	**0.181**	**8.517**	**71.71**	0.258	9.863	54.12	0.217	9.263	65.57
ZoeDepth [10] ft. Ours	**0.181**	8.946	71.43	**0.211**	**9.551**	**65.69**	**0.199**	**9.212**	**67.65**

Table 2: Evaluation of monocular networks on the nuScenes [13] validation set. Supervisions (sup.): M: monocular videos, S: singe-view images, ∗: test-time median-scaling via LiDAR, v: weak velocity, r: weak radar. Training data (tr. data): *d: day-clear, T: Translated in, n: night* (including *night-rain*), r: *day-rain*, a: all. Target Condition (T. Cond.): target atmospheric condition images known in advance. †: depth networks exclusively used within the diffusion model and not employed in the fine-tuning phase. We highlight the **1st** and <u>2nd</u> absolute bests.

Method	T. Cond.	sup.	tr. data	*day-clear*			***night***			***day-rain***		
				absRel ↓	RMSE ↓	δ_1 ↑	absRel ↓	RMSE ↓	δ_1 ↑	absRel ↓	RMSE ↓	δ_1 ↑
In-Domain												
R4Dyn w/o r in [26]	✗	Mvr	d	0.130	6.536	85.76	0.273	12.430	52.85	0.147	7.533	80.59
R4Dyn [26] (radar)	✗	Mvr	d	**0.126**	**6.434**	**86.97**	**0.219**	**10.542**	**62.28**	**0.134**	**7.131**	**83.91**
Monodepth2 [30]	✗	M*	d	0.137	6.692	85.00	0.283	9.729	51.83	0.173	7.743	77.57
PackNet-SfM [32]	✗	Mv	d	0.157	7.230	82.64	0.262	11.063	56.64	0.165	8.288	77.07
md4all (**baseline**) [27]	✗	Mv	d	<u>0.133</u>	6.459	**85.88**	0.242	10.922	58.17	0.157	7.453	79.49
md4all-DD [27] ft. Ours	✗	Mv	dT(nr)	0.137	**6.318**	<u>85.05</u>	**0.188**	**8.432**	<u>69.94</u>	0.147	7.345	79.59
md4all-DD [27] ft. Ours (DPT†)	✗	Mv	dT(nr)	0.140	6.573	83.51	0.197	8.826	69.65	<u>0.143</u>	<u>7.317</u>	<u>80.28</u>
md4all-DD [27] ft. Ours (Depth Anything †)	✗	Mv	dT(nr)	**0.128**	<u>6.449</u>	84.03	<u>0.191</u>	<u>8.433</u>	**71.14**	**0.139**	**7.129**	**81.36**
Monodepth2 [30]	✓	M*	a: dnr	<u>0.148</u>	<u>6.771</u>	**85.25**	2.333	32.940	10.54	0.411	9.442	60.58
RNW [91]	✓	M*	dn	0.287	9.185	56.21	0.333	10.098	43.72	0.295	9.341	57.21
md4all-AD [27] ft. Gasperini et al. [27]	✓	Mv	dT(nr)	0.152	6.853	83.11	<u>0.219</u>	<u>9.003</u>	<u>68.84</u>	<u>0.160</u>	<u>7.832</u>	<u>78.97</u>
md4all-DD [27] ft. Gasperini et al. [27]	✓	Mv	dT(nr)	**0.137**	**6.452**	<u>84.61</u>	**0.192**	**8.507**	**71.07**	**0.141**	**7.228**	**80.98**

challenges in achieving optimal performance in adverse *day-rain* and *night* scenarios, while performing effectively in the simpler *day-clear* setting. Significantly, our approach consistently outperforms the baselines across all metrics and atmospheric conditions, including both simple and adverse scenarios. This highlights the effectiveness of our methodology in mitigating the complexities associated with this task on state-of-the-art monocular depth networks designed for strong generalization across domains.

Comparison with Existing Methods. In Tables 2 and 3, we compare our approach to existing methods for single-image depth estimation, particularly those addressing challenging conditions. We categorize these methods based on their reliance on prior knowledge of the target image characteristics for a specific condition (*T. Cond.* flag in the tables). Monodepth2 [30] and PackNet [32] struggle to achieve satisfactory results if trained solely on *day-clear* images or incorporating all atmospheric conditions in nuScenes. Conversely, approaches that augment the inputs with additional information, such as R4Dyn [26] using radar data, show improvements, while methods such as that of Gasperini et al. [27] – which uses a ForkGAN [111] to transform all *day-clear* training samples into rainy and nocturnal environments – show superior performance in specific con-

Table 3: Evaluation of monocular depth frameworks on the RobotCar [52] test set. We follow the same notation provided in Tab. 2.

Method	T. Cond	Source	Sup.	Tr. Data	*day-clear*				***night***			
					absRel ↓	sqRel ↓	RMSE ↓	δ_1 ↑	absRel ↓	sqRel ↓	RMSE ↓	δ_1 ↑
DeFeatNet [80]	✓	[80]	M*	a: dn	0.247	2.980	7.884	65.00	0.334	4.589	8.606	58.60
ADIDS [49]	✓	[88]	M*	a: dn	0.239	2.089	6.743	61.40	0.287	2.569	7.985	49.00
RNW [91]	✓	[88]	M*	a: dn	0.297	2.608	7.996	43.10	0.185	1.710	6.549	73.30
WSGD [88]	✓	[88]	M*	a: dn	0.176	1.603	6.036	75.00	0.174	1.637	6.302	75.40
md4all-DD [27] ft. Gasperini et al. [27]	✓	[27]	Mv	dT(n)	**0.113**	**0.648**	**3.206**	**87.13**	**0.122**	**0.739**	**3.604**	**84.86**
Monodepth2 [30]	✗	[27]	M*	d	**0.112**	**0.670**	**3.164**	86.38	0.303	1.724	5.038	45.88
md4all (**baseline**) [27]	✗	[27]	Mv	d	0.121	0.723	3.335	86.61	0.391	3.547	8.227	22.51
md4all-DD [27] ft. Ours	✗	[27]	Mv	dT(n)	0.119	0.676	3.239	**87.20**	0.139	**0.739**	3.700	82.46
md4all-DD [27] ft. Ours (DPT †)	✗	[27]	Mv	dT(n)	0.123	0.724	3.333	86.62	0.133	0.824	3.712	**83.95**
md4all-DD [27] ft. Ours (Depth Anything †)	✗	[27]	Mv	dT(n)	0.119	0.728	3.287	87.17	**0.129**	0.751	**3.661**	83.68

figurations (md4all-AD and md4all-DD) over the self-supervised md4all baseline model [27]. Nevertheless, it is important to emphasize that Gasperini et al. [27], in order to achieve such results, rely on prior knowledge of specific image properties (such as noise, luminosity, etc.) for the considered atmospheric conditions in the target dataset. This underscores the stringent requirement for their approach to have images in the desired adverse environment. Our method, instead, significantly enhances the effectiveness of the self-supervised md4all baseline architecture, attaining comparable or even superior performance to Gasperini et al. [27], relying exclusively on the availability of easy samples. Importantly, our approach allows challenging images to be generated for fine-tuning by T2I-Adapter [56] using user text prompts and depth maps from various sources: md4all, advanced networks like DPT, or Depth Anything applied to *day-clear* images. This choice can be exploited based on T2I-Adapter's built-in compatibility with depth maps derived from these models. Nevertheless, the depth pseudo-labels are consistently derived from the md4all-baseline network during the fine-tuning process, and thus do not take advantage of other external depth annotations. We argue using DPT or Depth Anything for the conditional diffusion model does not yield an unfair advantage to our method: according to Tab. 1, these networks, at best, achieve (in generalization) comparable results to those trained specifically on nuScenes, like md4all-baseline. Notably, our results show improvements across challenging conditions regardless of the depth maps used for image generation. This trend is consistent in Tab. 3: our approach, without prior knowledge of the challenging images, outperforms others specialized in day-to-night translation and is comparable to md4all-DD [27].

Method Applicability with Daytime-only Datasets. In this experiment, we intend to demonstrate our distinct capability of generating adverse images from datasets that provide *easy* images only. To prove this, in Tab. 4, we present the results obtained by fine-tuning DPT and Depth Anything networks using five different datasets (Mapillary [57], Cityscapes [17], KITTI 2012 [28], KITTI 2015 [54], and Apolloscapes [40], totaling approximately 33K images). These datasets contain only *easy* conditions and are employed specifically for testing the ability to generalize across *challenging* conditions on the DrivingStereo [95], nuScenes [13], and RobotCar [52] datasets, without prior exposure to them. Importantly, other methodologies, such as [27], cannot be applied in

Table 4: Performance comparison of DPT [70] and Depth Anything [96] pre- and post-fine-tuning. Challenging images generated via [56] from samples in datasets with only *easy* conditions: Mapillary, Cityscapes, KITTI, and Apolloscapes.

Method	DrivingStereo [95] *day-rain*		nuScenes [13] *night*		nuScenes [13] *day-rain*		RobotCar [52] *night*	
	absRel ↓	δ_1 ↑	absRel ↓	δ_1 ↑	absRel ↓	δ_1 ↑	absRel ↓	δ_1 ↑
DPT (**baseline**) [70]	0.188	0.700	0.354	60.97	0.237	66.96	0.154	83.40
ft. Gasperini et al. [27]	✗	✗	✗	✗	✗	✗	✗	✗
ft. Ours	**0.124**	**0.836**	**0.263**	**67.39**	**0.202**	**70.38**	**0.130**	**86.60**
Depth Anything (**baseline**) [96]	0.112	0.854	0.291	67.10	0.167	75.17	0.125	87.15
ft. Gasperini et al. [27]	✗	✗	✗	✗	✗	✗	✗	✗
ft. Ours	**0.110**	**0.868**	**0.250**	**70.38**	**0.154**	**78.86**	**0.117**	**88.18**

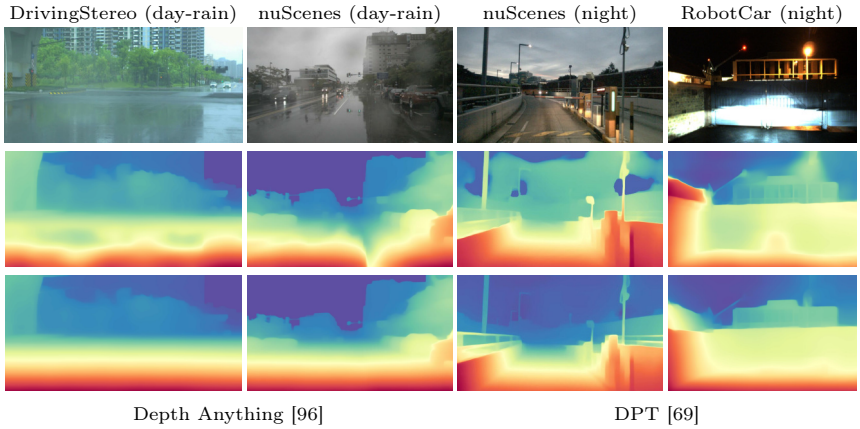


Fig. 5: Qualitative Results. From top to bottom: RGB images, depth maps predicted by the original models and the fine-tuned versions using our method.

this scenario due to the absence of both the *easy* and *challenging* images. Our methodology significantly enhances baseline results, as also evident in Fig. 5.

4.5 Challenging Materials

We now assess the effectiveness of our approach in handling non-Lambertian materials. Collecting many *easy* samples in this scenario would be complex, requiring the manual setup of scenes with only Lambertian objects. As such, we recall that [27], based on a style-transfer network, is unable to handle this scenario. To this aim, as described in Sec. 4.2, we first generate *easy* samples using Stable Diffusion [2] from text prompts, and then convert them to challenging ones using [56]. In Tab. 5, we report the performance of DPT [70] with official weights (Baseline) with DPT fine-tuned according to [18] (Depth4Tom) or with our approach (Ours). We fine-tune the networks using the framework from [18]. While this latter is trained on real data featuring transparent or mirror objects [94, 97] with ground truth segmentation maps, our method relies solely on images generated from text prompts. We test the *generalization* performance of all methods on the Booster [104] and ClearGrasp [74] datasets, following [18].

Table 5: Fine-tuning for ToM objects. Results on Booster [104] train set and ClearGrasp [74], at quarter and full resolution respectively. R. ToM: a-priori availability of a real dataset for non-Lambertian objects, GT Seg.: use of ground truth segmentation masks for ToM objects during training. All models start from the official weights [70].

Category	Method	R. ToM	GT Seg.	Booster [104]						ClearGrasp [74]					
				$\delta < 1.25$ ↑ (%)	$\delta < 1.15$ ↑ (%)	$\delta < 1.05$ ↑ (%)	MAE ↓ (mm)	absRel ↓	RMSE ↓ (mm)	$\delta < 1.25$ ↑ (%)	$\delta < 1.15$ ↑ (%)	$\delta < 1.05$ ↑ (%)	MAE ↓ (mm)	absRel ↓	RMSE ↓ (mm)
DPT-Large [69]															
All	DPT [70] (**baseline**)			96.79	89.71	56.26	75.35	0.06	100.68	98.71	94.68	64.95	32.77	0.05	45.31
	ft. Ours			**98.23**	**93.66**	**60.90**	65.29	**0.05**	**85.48**	98.61	95.67	66.23	30.64	0.05	**41.71**
	ft. Costanzino et al. [18]	✓	✓	97.99	93.55	60.46	**64.93**	**0.05**	85.93	**98.97**	**96.29**	**66.30**	**30.59**	**0.04**	41.75
ToM	DPT [70] (**baseline**)			92.77	80.98	37.70	113.14	0.10	136.28	96.50	87.39	45.63	41.04	0.07	47.85
	ft. Ours			96.17	**92.54**	52.88	79.64	0.07	92.56	**98.38**	**93.69**	57.15	**31.32**	0.06	**37.41**
	ft. Costanzino et al. [18]	✓	✓	**96.68**	92.23	**54.67**	**70.68**	**0.06**	**83.06**	97.46	92.85	**58.53**	31.55	**0.05**	37.45
Other	DPT [70] (**baseline**)			97.10	90.08	57.31	73.19	0.06	95.63	98.86	95.17	66.10	32.25	0.05	44.42
	ft. Ours			**98.35**	**96.79**	61.14	65.17	**0.05**	**85.08**	98.65	95.83	66.79	30.58	0.05	**41.42**
	ft. Costanzino et al. [18]	✓	✓	98.07	93.52	**61.19**	**64.70**	**0.05**	85.57	**99.05**	**96.50**	**66.81**	**30.54**	**0.04**	41.62
Depth Anything [96]															
All	Depth Anything [96] (**baseline**)			97.87	93.69	69.47	59.43	0.05	84.63	98.75	96.76	78.23	24.15	0.04	36.27
	ft. Ours			**99.44**	**97.18**	**76.44**	**41.50**	**0.03**	**56.78**	**99.89**	**99.16**	**79.13**	**19.73**	**0.03**	**26.53**
ToM	Depth Anything [96] (**baseline**)			84.23	71.10	39.94	137.96	0.13	162.62	83.46	59.26	15.84	82.22	0.15	91.88
	ft. Ours			**98.91**	**93.47**	**63.04**	**54.31**	**0.05**	**71.51**	**99.23**	**94.32**	**50.65**	**33.88**	**0.06**	**39.71**
Other	Depth Anything [96] (**baseline**)			99.05	95.48	71.90	52.13	0.04	70.14	99.74	98.92	81.34	21.08	**0.03**	29.57
	ft. Ours			**99.44**	**97.53**	**77.33**	**40.57**	**0.03**	**54.46**	**99.93**	**99.41**	**80.72**	**19.03**	**0.03**	**25.17**

Additionally, we conducted experiments with Depth Anything, one of the latest and most accurate state-of-the-art monocular depth estimation models, to verify the effectiveness of our method. As shown in Tab. 5, our approach enhances the baseline networks' performance on ToM surfaces across both datasets. Our method achieves results comparable to [18] on both the Booster and ClearGrasp datasets, with only slight variations in performance. Notably, while [18] relies on real-world images and manually annotated masks for ToM surfaces, our approach utilizes only text prompts. We argue that collecting a curated dataset with annotations for each setup would be costly and time-consuming. In contrast, our text-prompt-based image generation offers an efficient, cost-effective alternative. Moreover, we believe that this approach sets a precedent for wider future adoption and scalability in various scenarios. Lastly, our technique demonstrates superior versatility, adapting to any challenging setting without modifications, whereas Depth4Tom [18] is confined to mirrors and glasses.

5 Conclusion

In this work, we have introduced a pioneering training paradigm for monocular depth estimation that leverages diffusion models to address out-of-distribution scenarios. By transforming easy samples into complex ones, we generate diverse data that captures real-world challenges. Our fine-tuning protocol enhances the robustness and generalization capabilities of existing depth networks, enabling them to handle adverse weather and non-Lambertian surfaces without domain-specific data. Extensive experiments across multiple datasets and state-of-the-art architectures demonstrate the effectiveness and versatility of our approach.

Acknowledgements.

This study was funded by the European Union – Next Generation EU within the framework of the National Recovery and Resilience Plan NRRP – Mission 4 "Education and Research" – Component 2 - Investment 1.1 "National Research Program and Projects of Significant National Interest Fund (PRIN)" (Call D.D. MUR n. 104/2022) – PRIN2022 – Project reference: "RiverWatch: a citizen-science approach to river pollution monitoring" (ID: 2022MMBA8X, CUP: J53D23002260006).

It was carried out also within the MOST – Sustainable Mobility National Research Center and received funding from the European Union Next Generation EU within the framework of the National Recovery and Resilience Plan NRRP – Mission 4 "Education and Research" – Component 2 - Investment 1.4 (D.D. 1033 17/06/2022, CN00000023). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

1. Stable diffusion v1.5 model card (2022), `https://huggingface.co/runwayml/stable-diffusion-v1-5`
2. Stable diffusion xl - sdxl 1.0 model card (2023), `https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0`
3. Alembics: Disco diffusion (2022), `https://github.com/alembics/disco-diffusion`
4. Aleotti, F., Tosi, F., Poggi, M., Mattoccia, S.: Generative adversarial networks for unsupervised monocular depth prediction. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
5. Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: Spatext: Spatio-textual representation for controllable image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18370–18380 (2023)
6. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022)
7. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
8. Bashkirova, D., Lezama, J., Sohn, K., Saenko, K., Essa, I.: Masksketch: Unpaired structure-guided masked image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1879–1889 (2023)
9. 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)
10. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
11. Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.M., Reid, I.: Unsupervised scale-consistent depth and ego-motion learning from monocular video. Advances in neural information processing systems **32** (2019)
12. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
13. 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)

14. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8001–8008 (2019)
15. Chen, Y., Zhao, H., Hu, Z., Peng, J.: Attention-based context aggregation network for monocular depth estimation. International Journal of Machine Learning and Cybernetics **12**, 1583–1596 (2021)
16. Choi, H., Lee, H., Kim, S., Kim, S., Kim, S., Sohn, K., Min, D.: Adaptive confidence thresholding for monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12808–12818 (2021)
17. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
18. Costanzino, A., Zama Ramirez, P., Poggi, M., Tosi, F., Mattoccia, S., Di Stefano, L.: Learning depth estimation for transparent and mirror surfaces. In: The IEEE International Conference on Computer Vision (2023), iCCV
19. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE signal processing magazine **35**(1), 53–65 (2018)
20. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
21. 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)
22. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
23. 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)
24. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: European Conference on Computer Vision. pp. 89–106. Springer (2022)
25. Garg, R., Bg, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. pp. 740–756. Springer (2016)
26. Gasperini, S., Koch, P., Dallabetta, V., Navab, N., Busam, B., Tombari, F.: R4dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes. In: 2021 International Conference on 3D Vision (3DV). pp. 751–760. IEEE (2021)
27. 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 (2023)
28. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)

29. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
30. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth prediction. In: The International Conference on Computer Vision (ICCV) (October 2019)
31. Gordon, A., Li, H., Jonschkowski, R., Angelova, A.: Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8977–8986 (2019)
32. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2485–2494 (2020)
33. Guizilini, V., Hou, R., Li, J., Ambrus, R., Gaidon, A.: Semantically-guided representation learning for self-supervised monocular depth. arXiv preprint arXiv:2002.12319 (2020)
34. Guizilini, V., Vasiljevic, I., Chen, D., Ambruș, R., Gaidon, A.: Towards zero-shot scale-aware monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9233–9243 (2023)
35. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
36. Hoiem, D., Efros, A.A., Hebert, M.: Automatic photo pop-up. In: ACM SIGGRAPH 2005 Papers, pp. 577–584 (2005)
37. Hornauer, J., Belagiannis, V.: Gradient-based uncertainty for monocular depth estimation. In: European Conference on Computer Vision. pp. 613–630. Springer (2022)
38. Hu, M., Zheng, J., Liu, D., Zheng, C., Wang, C., Tao, D., Cham, T.J.: Cocktail: Mixing multi-modality control for text-conditional image generation. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
39. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)
40. Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., Yang, R.: The apolloscape open dataset for autonomous driving and its application. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2702–2719 (2019)
41. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
42. 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)
43. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2426–2435 (2022)
44. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. Advances in neural information processing systems **34**, 21696–21707 (2021)
45. Klingner, M., Termöhlen, J.A., Mikolajczyk, J., Fingscheidt, T.: Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic

guidance. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 582–600. Springer (2020)
46. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019)
47. Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1119–1127 (2015)
48. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE transactions on pattern analysis and machine intelligence **38**(10), 2024–2039 (2015)
49. 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)
50. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
51. Luo, Y., Ren, J., Lin, M., Pang, J., Sun, W., Li, H., Lin, L.: Single view stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 155–163 (2018)
52. 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)
53. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5667–5675 (2018)
54. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
55. 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)
56. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
57. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE international conference on computer vision. pp. 4990–4999 (2017)
58. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
59. OpenAI: Dall-e 2 (2023), `https://openai.com/product/dall-e-2`
60. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
61. Patil, V., Sakaridis, C., Liniger, A., Van Gool, L.: P3depth: Monocular depth estimation with a piecewise planarity prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1610–1621 (2022)

62. Peng, R., Wang, R., Lai, Y., Tang, L., Cai, Y.: Excavating the potential capacity of self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15560–15569 (2021)
63. Pilzer, A., Xu, D., Puscas, M., Ricci, E., Sebe, N.: Unsupervised adversarial depth estimation using cycled generative networks. In: 2018 international conference on 3D vision (3DV). pp. 587–595. IEEE (2018)
64. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3227–3237 (2020)
65. Poggi, M., Tosi, F., Mattoccia, S.: Learning monocular depth estimation with unsupervised trinocular assumptions. In: 2018 International conference on 3d vision (3DV). pp. 324–333. IEEE (2018)
66. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
67. Ramirez, P.Z., Costanzino, A., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Di Stefano, L.: Booster: a benchmark for depth from images of specular and transparent surfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
68. Ramirez, P.Z., Tosi, F., Di Stefano, L., Timofte, R., Costanzino, A., Poggi, M., Salti, S., Mattoccia, S., Zhang, Y., Wu, C., He, Z., Yin, S., Dong, J., Liu, Y., Jiang, H., Shi, J., A, Y., Jin, Y., Li, D., Ke, B., Obukhov, A., Wang, T., Metzger, N., Huang, S., Schindler, K., Huang, Y., Li, J., Zhang, J., Wang, Y., Huang, Z., Liu, T., Cao, Z., Li, P., Wang, J.L., Zhu, W., Geng, H., Zhang, Y., Lan, L., Xu, K., Sun, T., Xu, Q., Saini, S., Gupta, A., Mistry, S.K., Shukla, A., Jakhetiya, V., Jaiswal, S., Sun, Y., Zheng, Z., Ning, Y., Cheng, J.H., Liu, H.I., Huang, H.W., Yang, C.Y., Jiang, Z., Peng, Y.H., Huang, A., Hwang, J.N.: Ntire 2024 challenge on hr depth from images of specular and transparent surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 6499–6512 (June 2024)
69. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ICCV (2021)
70. 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) (2022)
71. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12240–12249 (2019)
72. 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)
73. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
74. Sajjan, S., Moore, M., Pan, M., Nagaraja, G., Lee, J., Zeng, A., Song, S.: Clear grasp: 3d shape estimation of transparent objects for manipulation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 3634–3642. IEEE (2020)
75. Saxena, A., Chung, S., Ng, A.: Learning depth from single monocular images. Advances in neural information processing systems **18** (2005)

76. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. IEEE transactions on pattern analysis and machine intelligence **31**(5), 824–840 (2008)
77. Saxena, S., Herrmann, C., Hur, J., Kar, A., Norouzi, M., Sun, D., Fleet, D.J.: The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. arXiv preprint arXiv:2306.01923 (2023)
78. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular depth estimation using diffusion models. arXiv preprint arXiv:2302.14816 (2023)
79. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
80. 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)
81. Spencer, J., Qian, C.S., Russell, C., Hadfield, S., Graf, E., Adams, W., Schofield, A.J., Elder, J.H., Bowden, R., Cong, H., et al.: The monocular depth estimation challenge. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 623–632 (2023)
82. Spencer, J., Qian, C.S., Trescakova, M., Russell, C., Hadfield, S., Graf, E.W., Adams, W.J., Schofield, A.J., Elder, J., Bowden, R., et al.: The second monocular depth estimation challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3064–3076 (2023)
83. Spencer, J., Tosi, F., Poggi, M., Arora, R.S., Russell, C., Hadfield, S., Bowden, R., Zhou, G., Li, Z., Rao, Q., et al.: The third monocular depth estimation challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1–14 (2024)
84. Sun, Q., Tang, Y., Zhang, C., Zhao, C., Qian, F., Kurths, J.: Unsupervised estimation of monocular depth and vo in dynamic environments via hybrid masks. IEEE Transactions on Neural Networks and Learning Systems **33**(5), 2023–2033 (2021)
85. Tosi, F., Aleotti, F., Poggi, M., Mattoccia, S.: Learning monocular depth estimation infusing traditional stereo knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9799–9809 (2019)
86. Tosi, F., Aleotti, F., Ramirez, P.Z., Poggi, M., Salti, S., Stefano, L.D., Mattoccia, S.: Distilled semantics for comprehensive scene understanding from videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4654–4665 (2020)
87. Vankadari, M., Garg, S., Majumder, A., Kumar, S., Behera, A.: Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16. pp. 443–459. Springer (2020)
88. Vankadari, M., Golodetz, S., Garg, S., Shin, S., Markham, A., Trigoni, N.: When the sun goes down: Repairing photometric losses for all-day depth estimation. In: Conference on Robot Learning. pp. 1992–2003. PMLR (2023)
89. Voynov, A., Aberman, K., Cohen-Or, D.: Sketch-guided text-to-image diffusion models. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
90. Wang, C., Buenaposada, J.M., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2022–2030 (2018)

91. 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)
92. Watson, J., Firman, M., Brostow, G.J., Turmukhambetov, D.: Self-supervised monocular depth hints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2162–2171 (2019)
93. Wu, C.Y., Wang, J., Hall, M., Neumann, U., Su, S.: Toward practical monocular indoor depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3814–3824 (2022)
94. Xie, E., Wang, W., Wang, W., Ding, M., Shen, C., Luo, P.: Segmenting transparent objects in the wild. arXiv preprint arXiv:2003.13948 (2020)
95. Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B.: Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 899–908 (2019)
96. 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)
97. Yang, X., Mei, H., Xu, K., Wei, X., Yin, B., Lau, R.W.: Where is my mirror? In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
98. Yin, W., Liu, Y., Shen, C., Yan, Y.: Enforcing geometric constraints of virtual normal for depth prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5684–5693 (2019)
99. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9043–9053 (2023)
100. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1983–1992 (2018)
101. Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: Neural window fully-connected crfs for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3916–3925 (2022)
102. Zama Ramirez, P., Fabio, T., Di Stefano, L., Timofte, R., Costanzino, A., Poggi, M., Salti, S., Mattoccia, S., Shi, J., Zhang, D., A, Y., Jin, Y., Li, D., Li, C., Liu, Z., Zhang, Q., Wang, Y., Yin, S.: NTIRE 2023 challenge on HR depth from images of specular and transparent surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2023)
103. Zama Ramirez, P., Poggi, M., Tosi, F., Mattoccia, S., Di Stefano, L.: Geometry meets semantics for semi-supervised monocular depth estimation. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 298–313. Springer (2019)
104. Zama Ramirez, P., Tosi, F., Poggi, M., Salti, S., Di Stefano, L., Mattoccia, S.: Open challenges in deep stereo: the booster dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2022), cVPR
105. Zavadski, D., Feiden, J.F., Rother, C.: Controlnet-xs: Designing an efficient and effective architecture for controlling text-to-image diffusion models. arXiv preprint arXiv:2312.06573 (2023)

106. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
107. Zhao, C., Sun, Q., Zhang, C., Tang, Y., Qian, F.: Monocular depth estimation based on deep learning: An overview. Science China Technological Sciences **63**(9), 1612–1627 (2020)
108. Zhao, C., Tang, Y., Sun, Q.: Unsupervised monocular depth estimation in highly complex environments. IEEE Transactions on Emerging Topics in Computational Intelligence **6**(5), 1237–1246 (2022)
109. 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. arXiv preprint arXiv:2208.03543 (2022)
110. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems **36** (2024)
111. Zheng, Z., Wu, Y., Han, X., Shi, J.: Forkgan: Seeing into the rainy night. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 155–170. Springer (2020)
112. 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)
113. Zou, Y., Luo, Z., Huang, J.B.: Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 36–53 (2018)