Uncertainty Diffusion: Parameter-Efficient Depth Refinement via Uncertainty-Guided Diffusion Models

Jeng-Huo Tzeng
banne2266.cs07@nycu.edu.tw
Chuan-Yuan Huang
cyhuang.c@nycu.edu.tw
Kuan-Wen Chen
kuanwen@cs.nycu.edu.tw

Department of Computer Science
National Yang Ming Chiao Tung University
Hsinchu City, Taiwan

Abstract

We present Uncertainty Diffusion, a model-agnostic framework for refining monocular depth maps by integrating pixel-wise uncertainty into a diffusion-based process. Our method adaptively focuses refinement on regions with low prediction confidence by leveraging an uncertainty-guided sampling mechanism, enabling targeted and effective depth enhancement. For domain adaptation, only a lightweight refinement network is fine-tuned while keeping the base model fixed, resulting in a parameter-efficient adaptation strategy. Extensive experiments on NYU Depth V2, DIODE, and SUN RGB-D demonstrate consistent improvements across diverse baseline models, including recent large-scale approaches. Notably, our framework achieves up to a 61% reduction in log10 error for domain adaptation, all without retraining the base models. These results highlight the practicality and versatility of Uncertainty Diffusion for robust monocular depth estimation in varied environments.

1 Introduction

Monocular depth estimation (MDE), the task of inferring depth from a single RGB image, is crucial for applications such as autonomous driving, robotics, and augmented reality. Despite substantial progress driven by convolutional neural networks (CNNs) [7, 16], MDE remains challenging due to the inherent ambiguity of recovering 3D structure from 2D observations. This challenge is particularly pronounced in regions with occlusions, textureless surfaces, or novel environments, often resulting in unreliable predictions.

To address these limitations, recent works have integrated uncertainty estimation into depth prediction pipelines. Common techniques include Monte Carlo Dropout [9], Bayesian neural networks [15], and ensemble methods [17], which quantify the reliability of predictions. However, most existing approaches utilize uncertainty primarily for downstream adaptation or error analysis [26], without directly leveraging it to refine depth predictions. This overlooks the potential for uncertainty maps to highlight areas where targeted refinement could be most beneficial.

© 2025. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

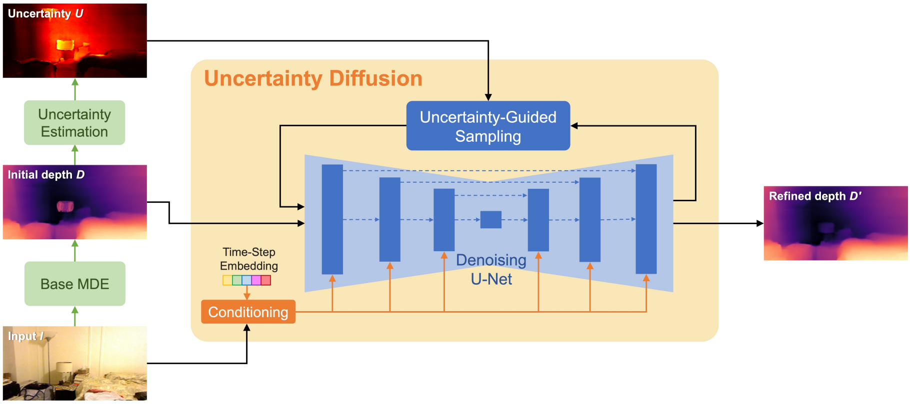

Figure 1: **Overview of the Uncertainty Diffusion architecture.** Our framework refines monocular depth predictions by leveraging pixel-wise uncertainty estimates to guide the diffusion process, where uncertainty adaptively constrains sampling during each reverse step.

Meanwhile, diffusion models have shown strong performance in various dense prediction tasks, including semantic segmentation and depth estimation [12, 33]. Their iterative denoising process offers a natural mechanism for prediction refinement. Nevertheless, current diffusion-based depth estimation methods typically treat all spatial regions equally during refinement, failing to exploit uncertainty information to focus on ambiguous or unreliable areas.

To bridge this gap, we propose **Uncertainty Diffusion**, a model-agnostic framework that integrates pixel-wise uncertainty estimates into a diffusion-based refinement process for depth maps. Our approach adaptively concentrates the denoising process on less confident regions, leading to more accurate and reliable depth predictions. Importantly, this design enables parameter-efficient domain adaptation: only a lightweight refinement network is fine-tuned for new domains, while keeping the large base model fixed. This makes our method broadly applicable and practical for real-world deployment.

We extensively evaluate our method on multiple benchmarks, including NYU Depth V2 [24], DIODE [30], and SUN RGB-D [29]. Experiments demonstrate consistent performance improvements across diverse and competitive baseline models such as Monodepth2 [11], ZoeDepth [4], and Depth-Anything [33]. Notably, our method achieves up to a 61% reduction in log10 error for parameter-efficient domain adaptation, without retraining large base models.

Our main contributions are as follows:

- We propose an uncertainty-guided diffusion refinement framework for monocular depth estimation, which leverages pixel-wise uncertainty to adaptively guide the denoising process and improve prediction quality.
- We introduce a novel uncertainty-guided sampling mechanism that dynamically focuses refinement on regions with low prediction confidence, enabling targeted and effective depth improvement.
- We demonstrate that our approach is model-agnostic and enables parameter-efficient domain adaptation by fine-tuning only a lightweight refinement network, validated

across multiple strong depth estimation backbones and datasets.

2 Related Work

Monocular Depth Estimation. Monocular depth estimation (MDE) has evolved from traditional handcrafted features [13, 22, 28] to deep neural networks [6, 7, 16], which have significantly advanced prediction accuracy from a single RGB image. Reformulating depth estimation as a classification or binning task [2, 3, 19] has further improved robustness and quantization. To reduce the need for labeled data, self-supervised methods [10, 36] leverage photometric consistency in stereo or video, enabling large-scale training. Recent competitive models such as ZoeDepth [4], VPD [35], and Depth-Anything [33] push the frontier through multi-cue fusion, diffusion backbones, and teacher-student learning on massive unlabeled datasets. Other approaches, such as GeoWizard [31], DepthFM [18], and Metric3D [34], explore transformer architectures, geometric cues, or feature-matching strategies, but do not leverage diffusion-based generative modeling.

Uncertainty Estimation. Uncertainty estimation is essential for understanding model reliability and is typically categorized as either epistemic (model) or aleatoric (data) [14, 15]. Techniques such as Monte Carlo Dropout [9], bootstrap ensembles [17], and Bayesian neural networks [23] are widely used for quantifying uncertainty. In MDE, Godard et al. [10] introduced horizontal flipping as a simple yet effective estimator, Poggi et al. [26] provided a comprehensive benchmark, and Nix and Weigend [25] proposed joint output and uncertainty prediction. However, most works use uncertainty for error analysis or downstream adaptation, rather than for directly improving predictions. Our approach systematically leverages pixel-wise uncertainty to guide the refinement process, enabling targeted enhancement of depth maps.

Diffusion Models for Dense Prediction. Diffusion models [12] have recently achieved strong results in image synthesis [8, 27] and have been adapted for dense prediction tasks such as semantic segmentation [1, 32], object detection [5], and depth estimation [21, 31, 35]. In particular, Marigold [21] and Marigold E2E-FT [20] are pioneering diffusion-based models for monocular depth estimation, demonstrating the potential of generative approaches for dense prediction. While some methods produce confidence maps, few utilize uncertainty for post-prediction refinement. To our knowledge, our work is among the first to tightly integrate pixel-wise uncertainty into the diffusion process, enabling adaptive and principled refinement, especially in ambiguous regions. Furthermore, our method enables parameter-efficient domain adaptation by requiring only a lightweight refinement module to be fine-tuned for new domains, eliminating the need to retrain large base models and thus significantly reducing adaptation cost.

3 Method

We propose **Uncertainty Diffusion**, a framework that refines monocular depth predictions by leveraging uncertainty estimates within a diffusion model (Figure 1). An RGB image is first processed by a base MDE model to generate a depth map and corresponding uncertainty, which are then jointly refined using our diffusion-based approach.

3.1 Base Monocular Depth Estimation

Given an RGB image, a monocular depth estimation (MDE) model predicts a single-channel depth map of the same spatial dimensions. Our Uncertainty Diffusion framework is model-agnostic and can be flexibly applied on top of any MDE backbone. In our experiments, we select three representative models: Monodepth2 [11], ZoeDepth [4], and Depth-Anything [33].

Monodepth2 utilizes self-supervised training based on photometric reconstruction. ZoeDepth integrates both relative and metric depth estimation in a two-stage pipeline. Depth-Anything adopts large-scale pre-training with teacher-student learning on diverse unlabeled data. Together, these models encompass a range of training strategies and represent recent advances in monocular depth estimation.

3.2 Uncertainty Estimation

Uncertainty estimation quantifies the reliability of predictions and identifies potential failure regions in depth maps. Following established taxonomy [14], we focus primarily on epistemic uncertainty (model uncertainty) as it aligns with our goal of correcting base model errors.

We investigate three alternative uncertainty estimation methods for our framework:

1) **Monte Carlo Dropout** [9]: By activating dropout during inference, we approximate Bayesian inference in deep networks. The depth estimate and uncertainty are computed as:

$$\hat{D} = \frac{1}{N}\sum_{n=1}^{N} D_n, \quad U = \frac{1}{N}\sum_{n=1}^{N} (D_n - \hat{D})^2, \tag{1}$$

where D_n is the output depth map from the n-th forward propagation in Monte Carlo Dropout, and N is the total number of forward propagations.

2) **Bootstrap Ensemble** [17]: We train multiple models on different data subsets, capturing uncertainty through prediction variance:

$$\hat{D} = \frac{1}{B}\sum_{b=1}^{B} D_b, \quad U = \frac{1}{B}\sum_{b=1}^{B} (D_b - \hat{D})^2, \tag{2}$$

where D_b is the output depth map from the b-th model in the bootstrap ensemble, and B is the total number of models in the ensemble.

3) **Horizontal Flipping** [10]: This efficient post-hoc method requires only two forward passes, leveraging the assumption that depth structure should be preserved under horizontal flipping:

$$\hat{D} = \frac{D + \mathcal{F}^{-1}(D_{\mathrm{f}})}{2}, \quad U = |D - \mathcal{F}^{-1}(D_{\mathrm{f}})|, \tag{3}$$

where D is the original depth prediction, D_{f} is the prediction from the flipped input, and $\mathcal{F}^{-1}$ represents the inverse flipping transformation.

In our framework, we select one of these methods to provide uncertainty estimates that guide the diffusion refinement process. In practice, we adopt Monte Carlo Dropout, as it demonstrates the best overall performance in our comparative analysis (see Sec. 4.6).

3.3 Uncertainty-Guided Diffusion Framework

Our model is inspired by the diffusion model [12], a powerful generative framework that has emerged in recent years. It learns to perform a reverse diffusion process to reconstruct high-definition images. Through our study of diffusion models and uncertainty, we found that the denoising capability of diffusion models aligns well with our goal of reducing prediction uncertainty. We thus develop a framework that connects uncertainty and diffusion, allowing the model to leverage uncertainty information to optimize outputs. Figure 1 illustrates our architecture. Based on the DDPM [12] architecture, we make the following modifications:

1. Since our target image is a depth map, we set the channel number of the generated image to 1. Simultaneously, we map the numeric range during the Diffusion process of [-1, 1] to the real depth range of [0m, 10m].

2. We integrate the RGB image (the input I in Algorithm 1) of the depth map to be corrected with the time embedding, which then serves as a condition input into the U-Net of the Diffusion model.

3. The original diffusion model needs to perform the set number of time steps for the reverse diffusion process when generating images. However, since the baseline output depth map is not pure noise, we only need to perform the last few reverse diffusion processes (T' in Algorithm 1).

4. **Most critically**, when the network corrects the depth map generated by the baseline, we introduce the information of uncertainty (the input U in Algorithm 1) after each reverse diffusion process during resampling, limiting the sampling range through our uncertainty guided-sampling mechanism. Normally, the sampling would be a normal distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$, but we replace its variance with the pixel-wise uncertainty map U, resulting in $\mathcal{N}(\mathbf{0},U)$.

Algorithm 1: Uncertainty Diffusion

Input: Initial depth map D, Uncertainty map U, RGB image I
Output: Refined depth map D'
$\mathbf{x}_{T'} = D$
for $t = T', \ldots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(\mathbf{x}_t,t,I)\right) + \sigma_t \mathbf{z} U$
end
$D' = \mathbf{x}_0$
return D'

Algorithm 1 presents our proposed Uncertainty Diffusion method, which incorporates the aforementioned modifications to the standard DDPM [12] framework. Through this method, we introduce uncertainty into the diffusion process and leverage its denoising capability to optimize regions of the depth map with higher uncertainty values. The fourth modification is particularly significant: uncertainty can represent the plausible range where the true depth values may lie. By limiting the sampling range of the reverse diffusion process during resampling, we ensure that areas with low uncertainty—where the model is confident—have a narrower sampling range, preserving accurate predictions. In contrast, regions with higher uncertainty are assigned a broader sampling range, allowing for more flexible adjustment and improved depth estimation in ambiguous areas.

Importantly, our framework enables parameter-efficient domain adaptation: when transferring to new environments or datasets, only a lightweight refinement module (e.g., a U-Net with approximately 14 million parameters) needs to be fine-tuned, while the large base depth model remains fixed. This design greatly reduces adaptation cost and computational burden, making our approach highly practical and broadly applicable for real-world deployment.

3.4 Probabilistic Formulation and Sampling Strategy

We formalize the uncertainty-guided diffusion process by extending the standard diffusion model framework. Given a depth prediction D and its uncertainty estimation U, we model the reverse diffusion process as a sequence of de-noising steps guided by uncertainty information.

The reverse diffusion process can be formulated as a conditional probability distribution:

$$p_\theta(x_{t-1}|x_t, I) = \mathcal{N}(\mu_\theta(x_t, t, I), \Sigma_\theta(x_t, t)), \tag{4}$$

where x_t represents the depth map at timestep t, I denotes the conditional RGB image, and μ_θ and Σ_θ are the mean and covariance functions, respectively.

To incorporate uncertainty guidance into the standard diffusion framework, we define these parameters as:

$$\mu_\theta(x_t, t, I) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t, t, I)\right), \tag{5}$$

$$\Sigma_\theta(x_t, t) = \sigma_t^2(\mathbf{I} + \lambda U), \tag{6}$$

where σ_t is the base noise level at timestep t, $\mathbf{I}$ is the identity matrix, λ is a hyper-parameter controlling the uncertainty influence, and U is the estimated uncertainty map. This formulation explicitly connects our uncertainty-guided sampling mechanism with the probabilistic diffusion framework.

The training objective follows the standard diffusion loss:

$$\mathcal{L} = \mathbb{E}_t[||\varepsilon_\theta(x_t, t, I) - \varepsilon||^2], \tag{7}$$

where ε_θ is the predicted noise and ε is the ground truth noise. During inference, our approach dynamically adjusts the sampling variance according to pixel-wise uncertainty, providing a theoretically sound mechanism for targeted refinement of uncertain regions while preserving confident predictions.

4 Experiments

4.1 Datasets

We conducted our experiments on the NYU Depth V2 [24] dataset, DIODE [30] dataset, and SUN RGB-D [29] dataset. The NYU Depth V2 dataset is an indoor depth dataset that

Table 1: Comparison with recent monocular depth estimation methods on NYU Depth V2. U.D. denotes our Uncertainty Diffusion. (*) indicates retraining. Depth-Anything results are from the in-domain metric depth estimation experiment weights provided by their official repository.

	RMSE(↓)	AbsRel(↓)	log10(↓)
SC-DepthV2	0.532	0.138	0.059
BTS	0.392	0.110	0.047
AdaBins	0.364	0.103	0.044
BinsFormer	0.330	0.094	0.040
DINOv2	0.279	0.090	0.037
Monodepth2(*)	0.4554	0.1365	0.0550
Monodepth2 with U.D.	**0.4469**	**0.1285**	**0.0528**
ZoeDepth	0.2700	0.0750	0.0320
ZoeDepth with U.D.	**0.2665**	**0.0730**	**0.0310**
Depth-Anything	0.206	0.0560	0.0240
Depth-Anything with U.D.	**0.2041**	**0.0556**	**0.0238**

includes 1,449 fully labeled paired RGB images and depth maps, as well as over 400,000 unpaired images and depth maps. We used the synchronized NYU Depth V2 dataset for training, which contains nearly 50,000 paired RGB images and depth maps. For testing, we used the fully labeled paired RGB images and depth maps. The DIODE dataset is a depth dataset with dense-labeled depth data, including high-resolution RGBD images with both indoor and outdoor scenes (8,574 indoor images and 16,884 outdoor images). SUN RGB-D contains about ten thousand RGBD images for training and about two thousand for testing.

In our experiments, we used DIODE and SUN RGB-D datasets to evaluate the parameter-efficient domain adaptation ability of our framework for different base models, where only the lightweight refinement module is fine-tuned on the target domain data while keeping the base model fixed.

4.2 Implementation Details

Our implementation adopts the DDPM [12] architecture with a linear alpha scheduler (10^{-4} to 0.02). The denoising U-Net progressively downsamples inputs to 1/8 of the original size, with feature dimensions [64, 128, 256, 512]. RGB images are provided as conditional input alongside time steps encoded via sinusoidal embeddings. Training is performed with the Adam optimizer (learning rate 1×10^{-4}) using an L1 loss, and target time steps are sampled uniformly.

For uncertainty estimation, we employ Monte Carlo Dropout (MCD) [9] with 8 forward passes (dropout $p = 0.1$), Bootstrap Ensemble [17] with 6 differently-initialized models, and horizontal flipping (H-Flipping) [10]. We utilize official weights for ZoeDepth [4] and Depth-Anything [33], while Monodepth2 [11] is retrained with supervision.

4.3 Comparison with Recent Methods on NYU Depth V2

Our proposed Uncertainty Diffusion consistently reduces prediction errors across all evaluated base models. As shown in Table 1, integrating Uncertainty Diffusion leads to improvements for each of the three representative depth estimation models. The most notable gain is observed for Monodepth2, which achieves an improvement of over 5.1% in AbsRel compared to its baseline. When applied to the strong ZoeDepth model, our method further yields

Table 2: Parameter-efficient domain adaptation results for base models on DIODE and SUN RGB-D. U.D. denotes our Uncertainty Diffusion. Percentages in parentheses indicate improvement over the base model without Uncertainty Diffusion.

	DIODE			SUN RGB-D		
	RMSE(↓)	AbsRel(↓)	log10(↓)	RMSE(↓)	AbsRel(↓)	log10(↓)
Monodepth2	3.817	0.911	1.142	2.009	0.876	0.924
Monodepth2 with U.D.	**2.975 (-22.06%)**	**0.583 (-36.00%)**	**0.444 (-61.12%)**	**1.533 (-23.69%)**	**0.626 (-28.53%)**	**0.431 (-53.35%)**
ZoeDepth	2.092	0.394	0.241	0.750	0.291	0.120
ZoeDepth with U.D.	**1.923 (-8.07%)**	**0.371 (-5.83%)**	**0.202 (-16.18%)**	**0.685 (-8.67%)**	**0.286 (-1.71%)**	**0.115 (-4.17%)**
Depth-Anything	1.943	0.391	0.230	0.491	0.223	0.068
Depth-Anything with U.D.	**1.729 (-11.01%)**	**0.329 (-15.85%)**	**0.176 (-23.47%)**	**0.455 (-7.33%)**	**0.219 (-1.79%)**	**0.066 (-2.94%)**

Table 3: Ablation on diffusion hyperparameters using Depth-Anything as the base model: (a) number of diffusion iterations; (b) number of diffusion timesteps. Negative values indicate improvement over baseline.

(a)

Iter.	RMSE↓	AbsRel↓	log10↓
0	1.943	0.391	0.230
4	1.861	0.361	0.208
8	1.807	0.342	0.193
12	1.784	0.339	0.187
16	**1.776**	**0.337**	**0.185**
20	**1.776**	**0.337**	**0.185**

(b)

Steps	RMSE↓	AbsRel↓	log10↓
50	**-0.176**	-0.026	**-0.046**
200	<u>-0.166</u>	**-0.051**	**-0.045**
500	-0.138	<u>-0.049</u>	<u>-0.039</u>
1000	-0.120	-0.045	-0.034

enhancements of 2.6% in AbsRel and 3.1% in log10 error. Even with Depth-Anything, a recent large-scale model, we observe additional improvements of 0.9% in RMSE, 0.6% in AbsRel, and 0.8% in log10 error.

These consistent improvements across diverse baselines demonstrate the generalization capability and broad applicability of our approach. The results indicate that Uncertainty Diffusion can effectively enhance depth estimation performance for a wide range of existing models.

4.4 Parameter-Efficient Domain Adaptation for Base Models on Different Datasets

Parameter-Efficient Domain Adaptation Protocol: For all domain adaptation experiments across different datasets (e.g., NYU Depth V2, DIODE), we keep the base depth estimation model fixed and fine-tune only the lightweight refinement network on the target domain. This protocol highlights the parameter-efficient nature of our approach and ensures that adaptation to new environments is both practical and resource-efficient.

We evaluated the parameter-efficient domain adaptation capability of our Uncertainty Diffusion framework on the DIODE [30] and SUN RGB-D [29] datasets. Base models were initialized with weights trained on NYU Depth V2 and only the lightweight U-Net refinement module within Uncertainty Diffusion (approximately 14M parameters) was fine-tuned, while the main backbone remained fixed. This setup demonstrates that our framework enables effective adaptation to new domains with minimal computational cost, eliminating

Table 4: Ablation on uncertainty estimation methods. Results on the DIODE dataset using Depth-Anything as the baseline. All methods improve over the baseline. Best results are in bold.

Method	RMSE(↓)	AbsRel(↓)	log10(↓)
Depth-Anything (baseline)	1.943	0.391	0.230
Ours (MCD)	**1.729**	**0.329**	**0.176**
Ours (Bootstrap)	1.738	0.343	0.181
Ours (H-Flipping)	1.757	0.333	0.182

the need for full model retraining.

As shown in Table 2, Uncertainty Diffusion consistently improves the adaptation performance of Monodepth2 [11], ZoeDepth [4], and Depth-Anything [33] when transferring to new datasets. The largest relative gain is observed with Monodepth2, achieving over 61% improvement in log10 error. For more recent models, such as ZoeDepth and Depth-Anything, improvements in the range of 10%–20% are still observed. These results highlight the generalization ability and practical efficiency of our approach for domain adaptation across diverse datasets and baseline models.

4.5 Ablation on Diffusion Hyperparameters

We conduct ablation experiments to analyze the effect of two key diffusion hyperparameters: the number of diffusion iterations and the number of timesteps in the diffusion model. As shown in Table 3(a), increasing the number of diffusion iterations consistently reduces RMSE, with performance gains saturating beyond 16 iterations. This suggests that a moderate number of iterations is sufficient to achieve most of the improvement, effectively balancing accuracy and computational cost.

Similarly, Table 3(b) shows that increasing the number of timesteps generally leads to better final performance, but with diminishing returns and substantially higher computational overhead beyond 200 timesteps. Based on these results, we adopt 200 timesteps as the default setting for subsequent experiments, as it provides a favorable trade-off between accuracy and efficiency for practical deployment.

4.6 Ablation on Uncertainty Estimation Methods

Table 4 summarizes the performance of three uncertainty estimation methods integrated with our Uncertainty Diffusion framework. All methods are evaluated on the DIODE dataset, and lower values indicate better performance.

We observe that all three uncertainty estimation strategies are compatible with our Uncertainty Diffusion framework and consistently enhance model outputs. Specifically, Monte Carlo Dropout (MCD) and horizontal flipping (H-Flipping) yield comparable improvements in RMSE and log10 error, while Bootstrap Ensemble also provides a notable gain. For the AbsRel metric, MCD achieves the best results, with the other two methods trailing slightly. These findings demonstrate the versatility and robustness of Uncertainty Diffusion across different uncertainty estimation strategies.

In summary, these results confirm that our Uncertainty Diffusion framework is robust to the choice of uncertainty estimation method, further supporting its practicality and broad

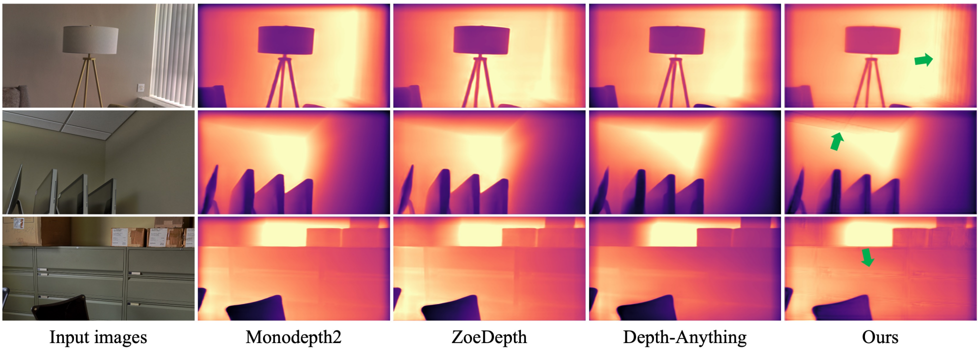

Figure 2: Qualitative comparison of monocular depth estimation methods. Green arrows highlight improvements. Our method achieves superior depth coherence in complex regions.

applicability in real-world scenarios.

4.7 Qualitative Results

We further conducted qualitative evaluations on the DIODE dataset to demonstrate the parameter-efficient adaptation capability of our framework. Figure 2 presents visual comparisons across different methods. Our Uncertainty Diffusion framework consistently produces more accurate and coherent depth maps. The green arrows highlight key improvements: (1) finer curtain surface details, (2) better preservation of ceiling structures, and (3) more accurate geometric relationships in cabinet gaps.

These visual results are consistent with the quantitative gains reported in Sec. 4.4, further validating the effectiveness of our uncertainty-guided sampling. Notably, even when refining recent strong models such as Depth-Anything, our approach yields visible enhancements in regions with complex geometry and depth discontinuities, underscoring the practical value of targeted refinement.

5 Conclusions

We have presented Uncertainty Diffusion, a novel framework that leverages uncertainty estimation to guide the diffusion-based refinement of monocular depth predictions. By dynamically adjusting the denoising process according to pixel-wise confidence, our method produces more accurate and reliable depth maps across diverse scenarios. Through extensive experiments on multiple benchmarks and representative baseline models, we demonstrate consistent improvements in depth estimation performance. Notably, our approach enables highly efficient domain adaptation by requiring only a lightweight refinement network to be fine-tuned, achieving substantial gains without retraining large base models. These results underscore the generalization capability and practical value of Uncertainty Diffusion for robust monocular depth estimation in real-world applications.

Acknowledgements

This work was supported in part by the National Science and Technology Council, Taiwan (113-2221-E-A49-164-MY3, 114-2221-E-A49 -054 -MY3 and 114-2218-E-A49-025-).

References

[1] Tomer Amit, Eliya Nachmani, Tal Shaharbany, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4009–4018, 2021.

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *European Conference on Computer Vision*, pages 480–496. Springer, 2022.

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. URL https://arxiv.org/abs/2302.12288.

[5] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022.

[6] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[8] Aditya Ramesh et al. Hierarchical text-conditional image generation with clip latents, 2022.

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[11] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *The International Conference on Computer Vision (ICCV)*, October 2019.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[13] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75:151–172, 2007.

[14] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3): 457–506, 2021.

[15] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[16] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[18] Yucheng Li, Yuzhe Wang, Jinyang Guo, Yifan Sun, and Yuchao Dai. Depthfm: Diffusion-based metric depth estimation with feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15145–15155, 2024.

[19] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022.

[20] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Marigold e2e-ft: End-to-end finetuning for diffusion-based monocular depth estimation. In *arXiv preprint arXiv:2404.03517*, 2024.

[21] Zhenyu Li, Junjie Yan, Xuyang Wang, Xianming Liu, and Junjun Jiang. Marigold: Hallucinating metric depth with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15134–15144, 2024.

[22] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. Sift flow: Dense correspondence across different scenes. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10*, pages 28–42. Springer, 2008.

[23] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[24] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[25] David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pages 55–60. IEEE, 1994.

[26] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[28] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008.

[29] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[30] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR*, abs/1908.00463, 2019. URL http://arxiv.org/abs/1908.00463.

[31] Xuyang Wang, Zhenyu Li, Xianming Liu, and Junjun Jiang. Geowizard: Hallucinating dense metric depth from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15093–15103, 2024.

[32] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *Proc. Int. Conf. Computer Vision (ICCV 2023)*, 2023.

[33] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

[34] Jiawei Zhang, Yifan Sun, Yuchao Dai, Xuyang Wang, Zhenyu Li, Xianming Liu, and Junjun Jiang. Metric3d: Hallucinating metric-scale depth from a single image via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15156–15166, 2024.

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

[36] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.