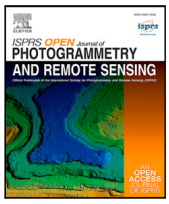
ISPRS Open Journal of Photogrammetry and Remote Sensing 21 (2026) 100146

Contents lists available at ScienceDirect

ISPRS Open Journal of Photogrammetry and Remote Sensing

journal homepage: www.journals.elsevier.com/isprs-open-journal-of-photogrammetry-and-remote-sensing



A critical synthesis of uncertainty quantification and foundation models in monocular depth estimation

Steven Landgraf [a,*], Rongjun Qin [b], Markus Ulrich [a]

[a] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany
[b] Department of Civil Environmental and Geodetic Engineering, Department of Electrical and Computer Engineering, The Ohio State University, Columbus, USA

ARTICLE INFO

Keywords:
Uncertainty quantification
Monocular depth estimation

ABSTRACT

Recent foundation models have led to major breakthroughs in monocular depth estimation (MDE), enabling dense 3D reconstruction from single-view imagery across a wide range of domains. However, safe and reliable deployment — particularly for metric depth estimation, which involves predicting absolute distances in real-world units — remains an open challenge. Even state-of-the-art foundation models are prone to critical errors, raising concerns for applications in photogrammetry, robotics, and remote sensing. Equipping these massive Vision Transformers with uncertainty quantification (UQ) is highly non-trivial; standard UQ methods often fail to scale, and it remains an open question how internet-scale pre-training and subsequent metric fine-tuning alter a model's uncertainty quality. To address these limitations, we integrate five scalable UQ methods with the DepthAnythingV2 foundation model to obtain pixel-wise uncertainty estimates alongside metric depth predictions. We evaluate these combinations across four diverse datasets, including indoor scenes, outdoor urban environments, synthetic-to-real data, and high-resolution aerial imagery. Our findings identify fine-tuning with the Gaussian Negative Log-Likelihood Loss as particularly promising, yielding reliable uncertainty estimates while preserving predictive performance and computational efficiency. By successfully adapting UQ to foundation models, this work advances toward more interpretable and trustworthy MDE pipelines, with direct implications for real-world photogrammetric workflows relying on single-view imagery.

1. Introduction

Monocular depth estimation (MDE) received significant attention in the photogrammetry and remote sensing community due to its crucial role in various downstream tasks ranging from autonomous driving (Xue et al., 2020; Xiang et al., 2022) and robotics (Dong et al., 2022; Roussel et al., 2019) to AI-generated content such as images (Zhang et al., 2023), videos (Liew et al., 2023), and 3D scene reconstruction (Xu et al., 2023; Shahbazi et al., 2024; Shriram et al., 2024). At its core, MDE aims to transform a single image into a depth map by regressing range values for each pixel, all without exploiting direct range or stereo measurements. Theoretically, MDE is a geometrically ill-posed problem that is fundamentally ambiguous and can only be solved with the help of prior knowledge about object shapes, sizes, scene layouts, and occlusion patterns. This inherent requirement for scene understanding perfectly aligns MDE with deep learning approaches, which have proven proficient in encoding potent priors (Bhat et al., 2023; Piccinelli et al., 2024; Yang et al., 2024a,b). These models benefit from extreme scaling, i.e., training on massive datasets and increasing model size, which facilitates the emergence of high-level visual scene understanding.

Based on these findings, a plethora of models have been proposed to address the challenges of MDE, with recent state-of-the-art solutions often leveraging large vision transformers trained on internet-scale data (Chen et al., 2016, 2020; Li and Snavely, 2018; Ranftl et al., 2020; Bhat et al., 2023; Piccinelli et al., 2024; Yang et al., 2024a,b), yielding foundation models capable of generalizing to a wide range of applications and scenes. A particularly challenging yet crucial application in fields such as robotics (Dong et al., 2022; Roussel et al., 2019), augmented reality (Kalia et al., 2019), and autonomous driving (Xue et al., 2020; Xiang et al., 2022) is the estimation of absolute distances in real-world units (e.g., meters), commonly referred to as metric depth estimation. This task is especially difficult due to inherent metric ambiguities caused by different camera models and scene variations. Fortunately, these foundation models can successfully be fine-tuned in the respective domain (Bhat et al., 2023; Piccinelli et al., 2024; Yang et al., 2024a,b) to determine exceptionally accurate metric depths.

However, the strong performance of foundation models on common benchmarks (Silberman et al., 2012; Geiger et al., 2012; Cordts et al.,

* Corresponding author.
E-mail address: steven.landgraf@kit.edu (S. Landgraf).

https://doi.org/10.1016/j.ophoto.2026.100146
Received 4 August 2025; Received in revised form 6 August 2026; Accepted 13 August 2026
Available online 27 August 2026

2016; Song et al., 2015) can lead to naive deployment, potentially overlooking their limitations. This is particularly detrimental in safety-critical applications where errors can have catastrophic consequences. Real-world deployment of deep learning models faces persistent challenges, including the "black box" nature of end-to-end systems (Roy et al., 2019; Gawlikowski et al., 2022), the tendency toward overconfidence (Guo et al., 2017), and the inability to flag out-of-distribution (OOD) samples (Lee et al., 2022, 2017). To mitigate these risks, recent research has advocated for uncertainty quantification (UQ) (Landgraf et al., 2024b; Leibig et al., 2017; Loquercio et al., 2020). Yet, integrating established UQ methodologies with massive MDE foundation models is a highly non-trivial endeavor, both computationally and theoretically, which explains the distinct lack of literature at this intersection.

Applying UQ to foundation models represents a substantive challenge rather than a mere incremental extension for several reasons. First, the sheer scale of Vision Transformers (ViTs) renders gold-standard UQ approaches, such as Deep Ensembles, computationally prohibitive, necessitating the adaptation of efficient approximations like Sub-Ensembles (SE) or Test-Time Augmentation (TTA). Second, the architectural idiosyncrasies of ViTs — such as the frequent absence of traditional dropout in modern pre-trained weights — complicate the direct application of methods like Monte Carlo Dropout (MCD) (Gal and Ghahramani, 2016). Most importantly, from a theoretical standpoint, it remains an open question how massive pre-training on internet-scale data alters a model's uncertainty landscape. Because these models have effectively "seen it all", traditional definitions of OOD become ambiguous. Furthermore, it is critical to understand whether the subsequent fine-tuning required for metric depth estimation distorts the latent UQ properties or induces overconfidence.

To bridge this gap and move from naive deployment to reliable application, this paper investigates the complex intersection of MDE foundation models and UQ. As illustrated in Fig. 1, even state-of-the-art foundation models are not immune to inaccuracies; however, successfully adapting UQ to correlate high uncertainties with erroneous predictions unlocks a pathway to safer deployment. We specifically focus on the state-of-the-art DepthAnythingV2 foundation model (Yang et al., 2024b) and systematically adapt five diverse UQ methods to enable pixel-wise uncertainty measures for metric depth estimation:

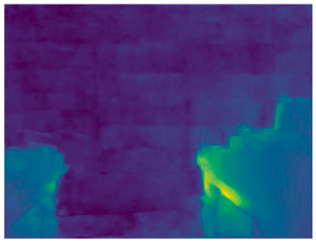
To bridge the gap between ground-breaking results in research and safe, reliable deployment in real-world applications, we investigate multiple UQ methods in combination with MDE foundation models. We specifically focus on combining the state-of-the-art DepthAnythingV2 foundation model (Yang et al., 2024b) with five different UQ methods to enable pixel-wise uncertainty measures for metric depth estimation:

1. **Learned Confidence (LC)** (Wan et al., 2018): Confidences, interpreted as uncertainties, are learned by extending the primary objective function with an additional loss term.
2. **Gaussian Negative Log-Likelihood (GNLL)** (Nix and Weigend, 1994): Predictions are treated as samples from a Gaussian distribution, with the network outputting both a predictive mean and its corresponding variance, which is learned implicitly through minimizing the Gaussian Negative Log-Likelihood.
3. **Monte Carlo Dropout (MCD)** (Gal and Ghahramani, 2016): Dropout layers remain active during inference, sampling from the posterior distribution to estimate a predictive mean and variance.
4. **Sub-Ensembles (SE)** (Valdenegro-Toro, 2023): A Deep Ensemble (Lakshminarayanan et al., 2017) is approximated by multiplying a subset of the model's layers instead of using the full model.
5. **Test-Time Augmentation (TTA)** (Ayhan and Berens, 2018): Perturbations applied to inputs during inference produce unique samples, enabling computation of a predictive mean and corresponding variance.

(a) Input image.

(b) Prediction.

(c) Binary accuracy map.



(d) Uncertainty.

Fig. 1. Qualitative example of a fine-tuned DepthAnythingV2 for metric monocular depth estimation on the NYUv2 dataset (Silberman et al., 2012), using a ViT-S encoder and Monte Carlo Dropout for an additional uncertainty estimate. The binary accuracy map is based on the δ_1 error. The strong correlation between erroneous predictions and high uncertainties highlights the potential of integrating uncertainty quantification (UQ) methods with foundation models for MDE.

We conduct extensive evaluations across four diverse datasets:

1. **NYUv2** (Silberman et al., 2012): Low-resolution indoor scenes.
2. **Cityscapes** (Cordts et al., 2016): High-resolution outdoor urban scenes.
3. **HOPE** (Tyree et al., 2022): Synthetic training images and real-world testing images of household objects.
4. **UseGeo** (Nex et al., 2024): High-resolution aerial images.

This wide selection ensures coverage of a broad spectrum of real-world applications. Additionally, we examine all publicly available pre-trained encoders of the DepthAnythingV2 model (ViT-S, ViT-B, and ViT-L), comprehensively revealing insights into the performance of models across various sizes and configurations.

We believe that our analysis will contribute to a deeper understanding of the limitations of these powerful foundation models, enhancing transparency and reliability in MDE. By highlighting often overlooked challenges, such as the lack of explainability, our work paves the way for future research that not only seeks to improve performance but also aims to teach these models to know what they do not know.

2. Related work

2.1. Monocular depth estimation & foundation models

Monocular Depth Estimation (MDE) is a fundamental, geometrically ill-posed task that has transitioned from early convolutional architectures (Eigen et al., 2014; Fu et al., 2018) to Vision Transformer (ViT)-based paradigms (Dosovitskiy et al., 2020; Ranftl et al., 2021; Piccinelli et al., 2023). By leveraging self-attention mechanisms, ViT architectures aggregate global contextual cues, significantly improving scene-level geometry understanding. Subsequent developments introduced novel formulation strategies, such as adaptive depth binning (Bhat et al., 2021; Li et al., 2024) and generative latent diffusion models (Saxena et al., 2024; Ke et al., 2024), to handle complex depth distributions.

To address the challenge of estimating depth "in the wild", where lighting, camera parameters, and scene structures vary drastically, recent efforts focus on scaling model capacity and dataset diversity (Chen et al., 2020; Li and Snavely, 2018). Prominent foundation frameworks like MiDaS (Ranftl et al., 2020) and DepthAnythingV2 (Yang et al., 2024b) utilize internet-scale pre-training to learn robust geometric priors, yielding state-of-the-art zero-shot performance for affine-invariant depth. However, safety-critical applications such as autonomous driving and robotics (Dong et al., 2022; Xue et al., 2020) require absolute metric measurements. Adapting zero-shot foundation models to these operational domains typically necessitates metric fine-tuning (Bhat et al., 2023; Piccinelli et al., 2024; Yang et al., 2024b), introducing a need to evaluate model reliability alongside predictive accuracy.

2.2. Uncertainty quantification in deep learning & MDE

Uncertainty Quantification (UQ) is essential for assessing model trustworthiness and safety under domain shifts (MacKay, 1992; Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017). Sampling-based methods, most notably Monte Carlo Dropout (MCD) (Gal and Ghahramani, 2016) and Deep Ensembles (Lakshminarayanan et al., 2017), remain benchmark standards for predictive quality, but their runtime complexity limits real-world deployment. To mitigate this overhead, efficiency-oriented approaches — such as sub-ensembles (Valdenegro-Toro, 2023) or parametric heteroscedastic loss formulations (e.g., Gaussian Negative Log-Likelihood) (Nix and Weigend, 1994; Kendall and Gal, 2017) — estimate uncertainty in a single forward pass by directly predicting distributional parameters.

In the context of MDE, UQ has been explored extensively across self-supervised frameworks (Poggi et al., 2020; Choi et al., 2021) and supervised settings. Supervised techniques range from auxiliary variance-estimation networks (Yu et al., 2021) and latent prototype analysis (Franchi et al., 2022) to test-time augmentation (TTA) and gradient-based post-hoc estimates (Hornauer and Belagiannis, 2022; Mi et al., 2022). Despite these advancements, existing MDE UQ literature focuses predominantly on custom CNNs or lightweight standalone ViTs trained from scratch, leaving the behavior of UQ techniques under large-scale pre-trained weights largely unexamined.

2.3. Research gap

While recent concurrent studies have begun exploring uncertainty in foundation models for specialized downstream tasks like floorplan localization (Wüest et al., 2025) or semantic segmentation (Landgraf et al., 2026), a systematic evaluation of scalable UQ paradigms on state-of-the-art MDE foundation models remains missing. Specifically, it remains unclear how different UQ strategies interact with frozen vs. fine-tuned geometric priors, and whether reliable uncertainty can be extracted without compromising predictive accuracy or introducing prohibitive inference latency. We address this gap by conducting a comprehensive, multi-domain evaluation of five scalable UQ strategies integrated into DepthAnythingV2 (Yang et al., 2024b) during metric fine-tuning.

3. Methodology

3.1. The DepthAnything foundation model

We select DepthAnythingV2 (Yang et al., 2024b) as our base architecture, as it currently represents the state-of-the-art in zero-shot MDE and demonstrates an exceptional capacity for metric fine-tuning. Built upon the powerful DINOv2 encoder (Oquab et al., 2023) and a DPT decoder (Ranftl et al., 2021), its training paradigm represents a significant evolution from its predecessor (Yang et al., 2024a). To eliminate the label noise and missing details inherent to real-world datasets, DepthAnythingV2 trains primarily on highly precise synthetic data. To subsequently prevent distribution shifts and ensure robust real-world scene coverage, this synthetic foundation is augmented with massive-scale pseudo-labeled real images generated by a scaled-up teacher model.

This dual-source training approach enables the network to learn potent geometric and semantic priors. Consequently, it serves as the ideal baseline to investigate whether these complex, internet-scale priors degrade or persist when subjected to uncertainty-aware metric fine-tuning.

3.2. Baseline metric fine-tuning

To enable a fair and consistent comparison across all five uncertainty-aware methods, we use the standard DepthAnythingV2 model (Yang et al., 2024b) as a baseline. For training the baseline models, we simply follow the fine-tuning recommendations of DepthAnythingV2 (Yang et al., 2024b). In the case of metric depth estimation, they suggest fine-tuning solely using the scale-invariant loss

$$\mathcal{L}_{\mathrm{SI}}(\hat{y}_i, y_i) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2 - \frac{\lambda}{N^2}\left(\sum_{i=1}^{N} d_i\right)^2}, \tag{1}$$

where N is the number of pixels with valid ground truth, $\hat{y}_i$ is the predicted depth of the ith pixel, y_i is the true depth, $d_i = \log y_i - \log \hat{y}_i$, and $\lambda = 0.2$.

3.3. Uncertainty quantification strategies

Our primary objective is to bridge the gap between theoretical foundation model performance and safe real-world deployment by extracting robust uncertainties alongside metric depth estimates. As illustrated in Fig. 2, we investigate five distinct UQ approaches. Notably, standard Deep Ensembles are excluded from this study as they require random weight initialization for the network (Lakshminarayanan et al., 2017; Fort et al., 2020), which would destroy the pre-trained weights of DepthAnythingV2 and severely degrade the baseline metric performance.

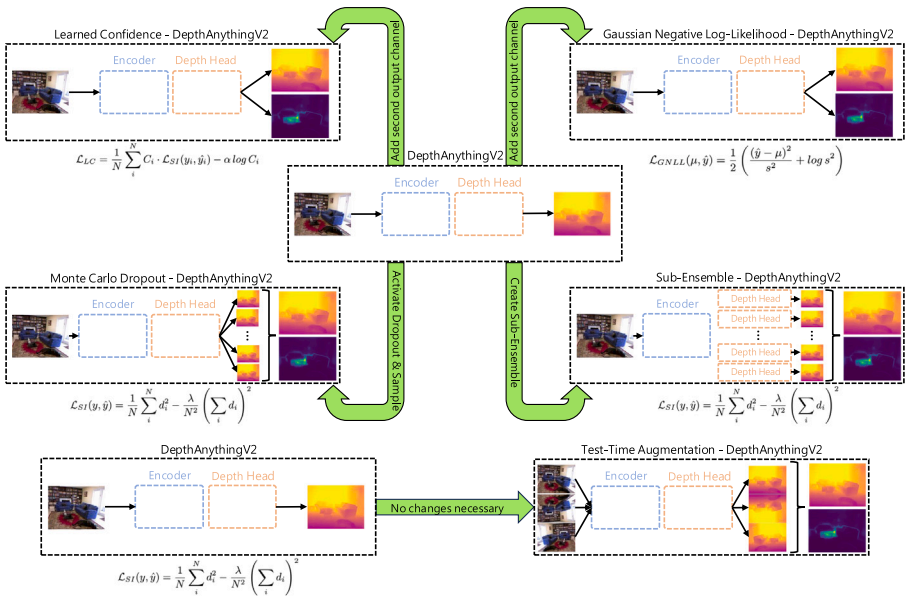


Fig. 2. A schematic overview of how to fuse the five different uncertainty quantification approaches with the DepthAnythingV2 (Yang et al., 2024b) foundation model.

3.3.1. Learned Confidence (LC)

Originally proposed for classification (Wan et al., 2018) and recently adapted for dense regression (Wang et al., 2024), LC requires adding a second output channel to the depth head. This channel predicts a confidence score, which we interpret as uncertainty. These confidences are learned by extending the primary objective function:

$$\mathcal{L}_{\mathrm{LC}} = \frac{1}{N}\sum_{i}^{N} C_i \cdot \mathcal{L}_{\mathrm{SI}}(y_i, \hat{y}_i) - \alpha \log C_i \ , \tag{2}$$

where C_i denotes the confidence score of the ith valid pixel generated by the model, and α is set to 0.2, following (Wang et al., 2024).

3.3.2. Gaussian Negative Log-Likelihood (GNLL)

Similar to LC, GNLL requires a secondary output channel. By treating the neural network prediction as a sample from a Gaussian distribution, the network outputs both a predictive mean μ and a corresponding variance s^2 (representing the uncertainty) (Nix and Weigend, 1994). Because there is no ground truth for uncertainty, s^2 is learned implicitly by minimizing the Gaussian Negative Log-Likelihood loss:

$$\mathcal{L}_{\mathrm{GNLL}}(\mu, \hat{y}) = \frac{1}{2}\left(\frac{(\hat{y}-\mu)^2}{s^2} + \log s^2\right) \ . \tag{3}$$

3.3.3. MC Dropout (MCD)

MCD (Gal and Ghahramani, 2016) enables sampling from the network's posterior distribution without architectural changes. Because DepthAnythingV2 already contains dropout layers, we simply keep them active during both fine-tuning (using the baseline $\mathcal{L}_{\mathrm{SI}}$ loss) and inference. We extract multiple stochastic depth predictions y_t and compute the predictive mean μ:

$$\mu = \frac{1}{T}\sum_{t}^{T} y_t \ , \tag{4}$$

where T is the number of samples and y_t is the tth depth prediction of the network. The uncertainty is quantified as the variance across these samples:

$$s^2 = \frac{1}{T-1}\sum_{t}^{T} \left(y_t - \mu\right)^2 \ . \tag{5}$$

Besides that, we follow the fine-tuning recommendations of DepthAnythingV2 (Yang et al., 2024b), using the scale-invariant loss $\mathcal{L}_{SI}(y, \hat{y})$ from Eq. (1) as the objective function.

3.3.4. Sub-Ensembles (SE)

To approximate a Deep Ensemble without the prohibitive cost of training multiple massive ViTs, we utilize Sub-Ensembles (Valdenegro-Toro, 2023). As shown in Fig. 2, we retain a single shared, pre-trained encoder but attach a sub-ensemble of randomly initialized depth heads. To maximize diversity and optimize training time, only one randomly selected head is optimized per training batch using $\mathcal{L}_{\mathrm{SI}}$. During inference, each head predicts a unique sample y_t based on the shared latent features, allowing us to compute the mean μ and variance s^2 using Eqs. (4) and (5).

3.3.5. Test-Time Augmentation (TTA)

Unlike the previous methods, TTA (Ayhan and Berens, 2018) requires no architectural modifications or specialized fine-tuning. The baseline DepthAnythingV2 model is trained using $\mathcal{L}_{\mathrm{SI}}$, as noted before. During test time, we apply horizontal and vertical flipping to the input image, generating three distinct spatial inputs (including the original). Inference is performed on all three, and the outputs are inversely transformed to the original coordinate space to compute the mean μ and variance s^2 using Eqs. (4) and (5).

3.3.6. Synthesis of UQ methods

To summarize the theoretical and practical distinctions among the evaluated UQ methods, Table 1 provides a comparative overview. These methods can be broadly categorized into two paradigms: parametric approaches (LC, GNLL) and sampling-based approaches (MCD, SE, TTA). From a theoretical standpoint, these paradigms capture fundamentally different sources of uncertainty. GNLL natively models aleatoric uncertainty, capturing the inherent noise, occlusion patterns, and metric ambiguities in the input data. Conversely, MCD and SE primarily capture epistemic uncertainty, reflecting the model's internal confidence in its learned geometric priors.

Furthermore, the methods diverge significantly in their computational requirements. While GNLL and LC introduce negligible inference overhead — making them highly attractive for real-time applications like autonomous driving — sampling-based methods like MCD and TTA require multiple forward passes, linearly scaling the computational cost. Sub-Ensembles present a strategic middle ground, requiring multiple passes only through the task-specific heads rather than the massive shared ViT encoder. By evaluating this diverse spectrum of UQ paradigms, we ensure a comprehensive analysis of how different

Table 1
Theoretical and practical synthesis of the evaluated UQ methods when integrated with the DepthAnythingV2 foundation model.

Method	Paradigm	Primary uncertainty type	Output quantity	Variance (s^2) estimation	Inference overhead
LC	Parametric	Predictive	Confidence score C_i	No (Heuristic Proxy)	1× (Minimal)
GNLL	Parametric	Aleatoric	Mean μ, Variance s^2	Yes (Directly Learned)	1× (Minimal)
MCD	Sampling-based	Epistemic	T stochastic samples	Yes (Computed via Samples)	T×
SE	Ensemble-based	Epistemic	K head samples	Yes (Computed via Samples)	K× (Partial)
TTA	Sampling-based	Predictive	M augmented samples	Yes (Computed via Samples)	M×

Table 2
Overview of metric depth datasets that we used for evaluation.

Dataset	Scene/Application	Resolution (W × H)	Training images	Test images
Cityscapes (Cordts et al., 2016)	Outdoor	2048 × 1024	2975	500
NYUv2 (Silberman et al., 2012)	Indoor	640 × 480	795	654
UseGeo (Nex et al., 2024)	Aerial	1989 × 1320	551	277
HOPE (Tyree et al., 2022)	Robotics	1920 × 1080	49 450 (Synth.)	457 (Real)

theoretical uncertainty types behave when integrated with massive pre-trained foundation models.

4. Experiments

In this Section, we conduct an extensive set of experiments to answer the question on how to fuse UQ with metric MDE. Firstly, we describe the experimental setup and explain the evaluation metrics. Secondly, we provide quantitative results comparing all five UQ methods applied to the DepthAnythingV2 foundation model on four diverse datasets with varying encoder sizes. Lastly, we offer qualitative examples of the UQ methods for all four datasets, showcasing the potential for foundation model uncertainty in metric MDE.

4.1. Experimental setup

Training. For all training processes, we follow the default settings of DepthAnythingV2 for metric depth fine-tuning (Yang et al., 2024b), using an AdamW optimizer (Loshchilov, 2017) with a base learning rate of $6 \cdot 10^{-5}$, a weight decay of 0.01, and a polynomial learning rate scheduler:

$$lr = lr_{\text{base}} \cdot (1 - \frac{\text{iteration}}{\text{total iterations}})^{0.9} \quad , \tag{6}$$

where lr is the current learning rate and lr_{base} is the initial base learning rate. Every model is trained for 25 epochs with an effective batch size of 16, using four NVIDIA A100 GPUs. We do not employ any early stopping techniques and hence only evaluate the final model checkpoints.

Datasets. We conduct our experiments on four highly different datasets, simulating a broad range of real-world applications, as shown by Table 2. Cityscapes (Cordts et al., 2016) provides an urban street scene benchmark dataset with high-resolution images. In contrast, NYUv2 (Silberman et al., 2012) presents indoor scenes with a very low image resolution. UseGeo (Nex et al., 2024) covers high-resolution aerial images, which are often neglected in the computer vision community despite their significance in many real-world applications. Finally, the HOPE (Tyree et al., 2022) dataset offers a variety of household objects, originally designed for pose estimation. The main reason why the HOPE dataset is so interesting is that the training dataset is based on almost 50,000 synthetic images, whereas the test dataset consists of just 457 real images. A unique challenge that is often overlooked but fairly common, especially in robotics (Hodaň et al., 2020; Sundermeyer et al., 2023; Hodan et al., 2024).

Data Augmentations. Regardless of the trained model, we apply random cropping with a crop size of 756 × 756 pixels on all datasets except NYUv2, which uses 630 × 476 pixels, and random horizontal flipping with a flip chance of 50%. For testing purposes, we use the original image resolutions as shown by Table 2.

Uncertainty Quantification. Unless stated otherwise, for MCD and SE, the final depth map and corresponding uncertainty are computed using ten samples or ten depth heads, respectively, in accordance with previous research (Lakshminarayanan et al., 2017; Fort et al., 2019; Landgraf et al., 2024b,a).

4.2. Evaluation metrics

Predictive Metrics. For quantitative evaluations of the metric depth estimation, we report all standard metrics widely used in MDE research (Masoumian et al., 2022; Arampatzakis et al., 2023):

1. RMSE, which measures overall depth prediction accuracy,
2. Absolute Relative Error (AbsRel), which captures relative depth deviations,
3. Logarithmic Mean Absolute Error (log10), which emphasizes errors on a logarithmic scale,
4. and three threshold-based accuracies (δ_1, δ_2, δ_3), which assess the proportion of predictions within specific error thresholds.

In contrast to classification tasks like semantic segmentation, prediction errors are continuous in MDE. As a result, there is no natural binary notion of correctness, prohibiting the straightforward utilization of standard calibration metrics like the Expected Calibration Error (ECE) (Naeini et al., 2015), which rely on categorical outcomes.

Threshold-Based Uncertainty Metrics. To evaluate the uncertainty quality in this continuous setting, we first adopt the metrics proposed by (Mukhoti and Gal, 2018):

1. **p(accurate|certain)**: The probability that the model is accurate on its output given that the uncertainty is below a specified threshold.
2. **p(uncertain|inaccurate)**: The probability that the uncertainty of the model exceeds a specified threshold given that the prediction is inaccurate.
3. **PAvPU**: The combined measure that captures both accurate pixels among certain predictions and uncertain pixels among inaccurate predictions.

The conditional probabilities p(accurate|certain) and p(uncertain|inaccurate) can be calculated as:

$$\begin{aligned} p(\text{accurate}|\text{certain}) &= \frac{n_{ac}}{(n_{ac} + n_{ic})} \quad , \\ p(\text{uncertain}|\text{inaccurate}) &= \frac{n_{iu}}{(n_{iu} + n_{ic})} \quad , \end{aligned} \tag{7}$$

where n_{ac} represents the number of pixels that are accurate and certain, n_{ic} the number of pixels that are inaccurate and certain, and n_{iu} the number of pixels that are inaccurate and uncertain. Finally, we combine both desired cases to compute PAvPU as:

$$\text{PAvPU} = \frac{(n_{ac} + n_{iu})}{n_{ac} + n_{au} + n_{ic} + n_{iu}} \quad , \tag{8}$$

Table 3
Efficiency analysis of five uncertainty quantification methods (TTA, LC, GNLL, SE, MCD) across three ViT encoder sizes (ViT-S, ViT-B, ViT-L) on the NYUv2 dataset. We compare parameter count, FLOPs, training time per epoch (on a single A100), and inference performance (latency and FPS). Latency statistics are averaged over 1000 forward passes.

NYUv2 (Indoor)		Parameters [M] ↓	FLOPs [G] ↓	Train time [mm:ss] ↓	Inference [ms] ↓	FPS ↑
ViT-S	Baseline	24.8	66.82	00:27	12.9 ± 0.1	77.5
	TTA	24.8	66.82	00:27	40.0 ± 0.9	25.0
	LC	24.8	66.83	00:27	12.9 ± 0.2	77.5
	GNLL	24.8	66.83	00:27	12.9 ± 0.2	77.5
	SE	49.3	177.29	00:29	36.5 ± 0.5	27.4
	MCD	24.8	66.82	00:30	128.4 ± 0.5	7.8
ViT-B	Baseline	97.5	217.49	00:48	24.9 ± 0.5	40.2
	TTA	97.5	217.49	00:48	75.5 ± 1.2	13.2
	LC	97.5	217.50	00:49	25.1 ± 0.5	39.8
	GNLL	97.5	217.50	00:49	25.0 ± 0.7	40.0
	SE	195.5	608.09	00:52	61.9 ± 0.8	16.2
	MCD	97.5	217.49	00:54	248.4 ± 2.7	4.0
ViT-L	Baseline	335.3	741.40	02:11	57.1 ± 3.5	17.5
	TTA	335.3	741.40	02:11	172.6 ± 11.6	5.8
	LC	335.3	741.41	02:14	57.0 ± 3.9	17.5
	GNLL	335.3	741.41	02:13	57.1 ± 3.3	17.5
	SE	613.8	2204.07	02:17	131.6 ± 5.2	7.6
	MCD	335.3	741.40	02:26	569.8 ± 24.3	1.8

where n_{au} is the number of pixels that are accurate and uncertain. Clearly, these metrics depend heavily on the choice of the accuracy and uncertainty thresholds.

To determine whether a depth prediction is accurate, we use the strictest threshold-based accuracy δ_1, defined as:

$$\delta_1 = \max\left(\frac{y}{\hat{y}}, \frac{\hat{y}}{y}\right) \quad . \tag{9}$$

In this case, a pixel is deemed accurate if $\delta_1 < 1.25$, indicating that the predicted depth is within 25% of the ground truth.

To determine if a pixel is certain or uncertain, we threshold the predicted uncertainty at the median uncertainty of the given image. In real-world deployment scenarios — such as autonomous navigation or robotic grasping — downstream control systems often require binary "trust or reject" signals to safely fuse sensor data or trigger fallback protocols. Unlike a static absolute threshold, an image-specific dynamic threshold like the median mimics this boolean logic while providing robustness against varying overall confidence levels (Landgraf et al., 2025).

Threshold-Free Uncertainty Metrics. While threshold-based metrics provide actionable insights for deployment, they remain inherently sensitive to the chosen cutoff. To guarantee a comprehensive, threshold-agnostic evaluation of the uncertainty calibration, we additionally report two threshold-free metrics:

1. **Area Under the Sparsification Error (AUSE):** AUSE evaluates how well the estimated uncertainty ranks the pixel-wise errors. A sparsification curve is generated by progressively removing fractions of pixels with the highest predicted uncertainty and computing the error (e.g., RMSE) on the remaining pixels. The Sparsification Error is the area between this estimated curve and the "oracle" curve (where pixels are ideally removed based on their true error). A lower AUSE indicates that the model's high uncertainties perfectly correlate with its highest errors.
2. **Spearman's Rank Correlation Coefficient** (ρ): We compute the Spearman correlation between the predicted pixel-wise uncertainties and the actual absolute errors. This metric assesses the monotonic relationship between the two variables, with a higher positive ρ ($\in [-1, 1]$) indicating that as the predictive error increases, the predicted uncertainty reliably increases as well.

4.3. Quantitative evaluation

Efficiency. As shown by Table 3, training times are comparable to the baseline for all methods, with LC and GNLL matching it exactly in both training and inference time. SE and MCD increase training times by around 5%–10%. While SE nearly triples inference time and roughly doubles the trainable parameters, MCD requires roughly 10 times the inference time due to the costly sampling process that roughly scales linearly with the amount of samples. TTA also triples inference time as it requires three forward passes. Overall, LC and GNLL are the most efficient approaches as they do not require any additional computational overhead compared to the baseline. These experiments were conducted on the NYUv2 dataset only, so exact numbers may vary across datasets, but the general findings should be representative.

NYUv2. Table 4 shows a quantitative comparison for the NYUv2 dataset (Silberman et al., 2012). LC, GNLL, and SE maintain depth quality similar to the baseline, with SE and GNLL even surpassing it for the ViT-S and ViT-B encoders, respectively. TTA and MCD exhibit a slight degradation but remain competitive.

Regarding uncertainty quality, GNLL emerges as the most consistent top-performer. It dominates the threshold-based metrics, yielding p(acc|cer) values up to 98.0% and p(unc|ina) scores reaching 91.2% for the ViT-L encoder. The threshold-free metrics further substantiate these findings: GNLL consistently achieves the lowest AUSE, indicating that its uncertainty estimates provide the most accurate ranking of pixel-wise errors. While MCD yields strong results for Spearman's correlation (ρ) and PAvPU, it remains computationally expensive. In contrast, LC provides a less reliable uncertainty landscape, particularly evidenced by its poor p(unc|ina) and negative ρ values, suggesting that its learned confidence score fails to monotonically track predictive error. Overall, GNLL provides the most robust uncertainty calibration across both threshold-based decision-making and threshold-agnostic rank correlation.

Cityscapes. For the Cityscapes dataset (Cordts et al., 2016), as shown by Table 5, TTA stands out by significantly outperforming the baseline across all three encoder sizes in terms of depth quality. In contrast, LC, GNLL, and SE exhibit noticeable degradation, while MCD performs the worst.

Regarding uncertainty quality, GNLL is the standout performer, decisively surpassing all other methods across both threshold-based and threshold-free metrics for all encoder sizes. It achieves high reliability scores, including p(acc|cer) up to 69.2% and PAvPU reaching 69.5%.

Table 4
Quantitative comparison on NYUv2 across five uncertainty quantification methods (TTA, LC, GNLL, SE, MCD) and three ViT encoder sizes (ViT-S, ViT-B, ViT-L). We report 6 predictive metrics (RMSE, AbsRel, log10, and $\delta_{1,2,3}$), 3 threshold-based uncertainty metrics (p(acc|cer), p(unc|ina), PAvPU), and 2 threshold-free uncertainty metrics (AUSE, ρ). Best results are marked in **bold**.

NYUv2 (Indoor)		RMSE ↓	AbsRel ↓	log10 ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	p(acc|cer) ↑	p(unc|ina) ↑	PAvPU ↑	AUSE ↓	ρ ↑
ViT-S	Baseline	**0.340**	0.093	0.039	0.928	0.988	0.997	–	–	–	–	–
	TTA	0.399	0.111	0.049	0.881	0.984	0.997	0.903	0.602	0.522	0.091	0.336
	LC	0.343	0.090	0.039	0.930	0.988	0.997	0.920	0.369	0.490	0.170	−0.054
	GNLL	0.342	0.094	0.040	0.924	0.987	0.997	**0.953**	**0.846**	0.529	**0.068**	0.368
	SE	**0.340**	0.092	0.039	0.926	0.988	0.997	0.937	0.704	0.511	0.091	0.229
	MCD (10%)	0.422	0.121	0.050	0.867	0.973	0.992	0.900	0.692	**0.533**	0.082	**0.372**
ViT-B	Baseline	0.307	0.080	0.034	0.948	0.991	0.998	–	–	–	–	–
	TTA	0.359	0.099	0.435	0.910	0.988	0.998	0.924	0.599	0.514	0.081	0.331
	LC	0.314	0.085	0.036	0.943	0.991	0.998	0.926	0.292	0.483	0.156	−0.047
	GNLL	**0.305**	0.079	0.034	0.949	0.991	0.998	**0.966**	**0.826**	0.517	**0.057**	**0.373**
	SE	0.309	0.080	0.034	0.947	0.991	0.998	0.955	0.698	0.507	0.079	0.212
	MCD (10%)	0.339	0.091	0.039	0.925	0.986	0.997	0.952	0.774	**0.527**	0.065	0.349
ViT-L	Baseline	**0.270**	0.068	0.030	0.964	0.993	0.998	–	–	–	–	–
	TTA	0.324	0.087	0.039	0.938	0.992	0.998	0.944	0.598	0.507	0.069	**0.363**
	LC	0.275	0.069	0.030	0.963	0.993	0.998	0.946	0.268	0.483	0.133	−0.012
	GNLL	0.285	0.072	0.031	0.959	0.992	0.998	**0.980**	**0.912**	0.521	**0.049**	0.354
	SE	0.280	0.070	0.030	0.961	0.993	0.998	0.969	0.734	0.508	0.064	0.261
	MCD (10%)	0.339	0.091	0.039	0.925	0.986	0.997	0.967	0.798	**0.522**	0.057	0.335

Table 5
Quantitative comparison on Cityscapes across five uncertainty quantification methods (TTA, LC, GNLL, SE, MCD) and three ViT encoder sizes (ViT-S, ViT-B, ViT-L). We report 6 predictive metrics (RMSE, AbsRel, log10, and $\delta_{1,2,3}$), 3 threshold-based uncertainty metrics (p(acc|cer), p(unc|ina), PAvPU), and 2 threshold-free uncertainty metrics (AUSE, ρ). Best results are marked in **bold**.

Cityscapes (Outdoor)		RMSE ↓	AbsRel ↓	log10 ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	p(acc|cer) ↑	p(unc|ina) ↑	PAvPU ↑	AUSE ↓	ρ ↑
ViT-S	Baseline	7.138	0.219	0.084	0.704	0.964	0.991	–	–	–	–	–
	TTA	**6.474**	0.223	0.089	0.576	0.965	0.991	0.335	0.421	0.410	4.410	0.360
	LC	7.492	0.206	0.078	0.733	0.958	0.991	0.658	0.619	0.611	5.999	0.457
	GNLL	7.739	0.236	0.089	0.669	0.949	0.987	**0.692**	**0.706**	**0.695**	**2.188**	**0.715**
	SE	7.714	0.234	0.088	0.681	0.954	0.990	0.658	0.654	0.649	2.855	0.667
	MCD (10%)	8.527	0.307	0.113	0.417	0.920	0.984	0.328	0.525	0.516	3.644	0.661
ViT-B	Baseline	6.884	0.252	0.095	0.599	0.965	0.992	–	–	–	–	–
	TTA	**5.757**	0.247	0.095	0.526	0.967	0.993	0.288	0.420	0.396	4.911	0.364
	LC	7.711	0.285	0.107	0.434	0.958	0.991	0.329	0.506	0.502	8.065	0.355
	GNLL	7.092	0.244	0.094	0.613	0.966	0.991	**0.594**	**0.635**	**0.629**	2.581	0.611
	SE	7.824	0.271	0.102	0.534	0.957	0.991	0.490	0.587	0.588	**2.558**	**0.728**
	MCD (10%)	8.268	0.288	0.107	0.488	0.939	0.989	0.449	0.579	0.583	3.206	0.718
ViT-L	Baseline	6.655	0.256	0.097	0.558	0.972	0.993	–	–	–	–	–
	TTA	**5.298**	0.227	0.088	0.608	0.979	0.994	0.371	0.431	0.416	4.699	0.391
	LC	7.392	0.280	0.105	0.416	0.969	0.993	0.292	0.488	0.488	9.626	0.322
	GNLL	6.562	0.234	0.092	0.628	0.970	0.990	**0.581**	**0.620**	**0.607**	2.928	0.525
	SE	7.522	0.272	0.103	0.500	0.966	0.993	0.446	0.568	0.570	**2.624**	0.703
	MCD (10%)	7.684	0.272	0.102	0.526	0.956	0.991	0.480	0.584	0.584	3.258	**0.706**

These threshold-based observations are strongly corroborated by the threshold-free metrics: GNLL consistently achieves the lowest AUSE and highest Spearman correlation (ρ) in most configurations, demonstrating a robust monotonic relationship between its predicted uncertainty and absolute predictive error. While SE also demonstrates strong calibration (e.g., $\rho = 0.728$ for ViT-B), GNLL remains the most consistent across the encoder variants. Conversely, other methods exhibit significant reliability fluctuations; for instance, TTA consistently yields the poorest uncertainty scores, and LC struggles to maintain uncertainty alignment as model capacity increases.

UseGeo. For the UseGeo dataset (Nex et al., 2024), as presented by Table 6, the depth quality results are inconsistent, with methods sometimes surpassing and at other times falling short of the baseline. TTA is the only approach that consistently outperforms the baseline across all three encoder sizes.

In terms of uncertainty quality, all methods deliver high reliability for p(acc|cer) (≥96.3%), indicating a strong capability to identify accurate pixels. For p(unc|ina) and PAvPU, MCD generally performs best, achieving values up to 67.2% and 50.4%, respectively. In contrast to the other datasets, GNLL significantly lags in calibration on UseGeo. This is primarily attributed to the large absolute depth values characteristic of aerial imagery, which inflate the absolute magnitude of the GNLL loss. The logarithmic term within the GNLL objective function penalizes high uncertainty estimates, and in the context of large-scale aerial distances, these uncertainties — and the resulting loss — are naturally magnified. While no hyperparameter adjustments were made to ensure strict comparability, these results indicate that GNLL's sensitivity to absolute scale requires careful normalization for geospatial applications.

HOPE. Table 7 shows the final quantitative comparison for the HOPE dataset (Tyree et al., 2022). In terms of depth quality, there are only marginal differences between the baseline and all uncertainty approaches, with RMSE values ranging between 0.215 to 0.277.

Table 6
Quantitative comparison on UseGeo across five uncertainty quantification methods (TTA, LC, GNLL, SE, MCD) and three ViT encoder sizes (ViT-S, ViT-B, ViT-L). We report 6 predictive metrics (RMSE, AbsRel, log10, and $\delta_{1,2,3}$), 3 threshold-based uncertainty metrics (p(acc|cer), p(unc|ina), PAvPU), and 2 threshold-free uncertainty metrics (AUSE, ρ). Best results are marked in **bold**.

UseGeo (Aerial)		RMSE ↓	AbsRel ↓	log10 ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	p(acc|cer) ↑	p(unc|ina) ↑	PAvPU ↑	AUSE ↓	ρ ↑
ViT-S	Baseline	7.366	0.077	0.032	0.973	0.994	0.999	–	–	–	–	–
	TTA	7.078	0.074	0.031	0.977	0.995	0.999	0.976	0.647	0.500	**2.889**	0.045
	LC	8.213	0.086	0.036	0.958	0.995	1.000	0.963	0.481	**0.506**	3.085	0.058
	GNLL	7.467	0.076	0.032	0.979	0.993	0.998	0.976	0.237	0.500	3.167	**0.088**
	SE	7.259	0.076	0.032	0.976	0.994	0.999	0.971	0.467	0.495	3.383	−0.037
	MCD (10%)	**6.682**	0.068	0.029	0.982	0.994	0.998	**0.982**	**0.672**	0.501	2.978	−0.007
ViT-B	Baseline	6.386	0.067	0.028	0.981	0.995	0.999	–	–	–	–	–
	TTA	**6.060**	0.063	0.027	0.985	0.995	0.999	**0.984**	0.609	0.499	**2.444**	0.047
	LC	6.612	0.068	0.029	0.975	0.995	1.000	0.967	0.454	0.492	2.694	0.077
	GNLL	7.810	0.080	0.035	0.972	0.990	0.997	0.971	0.293	0.499	3.450	0.009
	SE	6.491	0.068	0.028	0.980	0.994	0.999	0.981	0.559	0.502	2.754	0.017
	MCD (10%)	6.641	0.070	0.029	0.978	0.994	0.999	0.982	**0.657**	**0.504**	2.543	**0.088**
ViT-L	Baseline	6.173	0.065	0.027	0.981	0.995	0.999	–	–	–	–	–
	TTA	5.898	0.063	0.026	0.982	0.995	0.999	0.980	0.596	0.499	2.417	0.017
	LC	**5.406**	0.056	0.024	0.986	0.995	1.000	**0.985**	0.477	0.500	**2.240**	0.030
	GNLL	7.082	0.073	0.031	0.980	0.991	0.998	0.980	0.294	0.498	2.897	0.055
	SE	6.260	0.067	0.028	0.980	0.995	1.000	0.977	0.577	0.497	2.529	0.066
	MCD (10%)	6.697	0.071	0.029	0.981	0.995	0.999	0.982	**0.669**	**0.501**	2.357	**0.133**

Table 7
Quantitative comparison on HOPE across five uncertainty quantification methods (TTA, LC, GNLL, SE, MCD) and three ViT encoder sizes (ViT-S, ViT-B, ViT-L). We report 6 predictive metrics (RMSE, AbsRel, log10, and $\delta_{1,2,3}$), 3 threshold-based uncertainty metrics (p(acc|cer), p(unc|ina), PAvPU), and 2 threshold-free uncertainty metrics (AUSE, ρ). Best results are marked in **bold**.

HOPE (Robotics)		RMSE ↓	AbsRel ↓	log10 ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	p(acc|cer) ↑	p(unc|ina) ↑	PAvPU ↑	AUSE ↓	ρ ↑
ViT-S	Baseline	0.263	0.265	0.115	0.537	0.821	0.942	–	–	–	–	–
	TTA	0.264	0.262	0.114	0.539	0.818	0.943	**0.421**	0.564	0.552	0.200	0.251
	LC	**0.259**	0.262	0.114	0.537	0.822	0.945	0.339	0.476	0.474	0.424	−0.162
	GNLL	0.277	0.287	0.124	0.492	0.795	0.929	0.404	**0.575**	**0.567**	**0.136**	**0.386**
	SE	0.262	0.263	0.117	0.544	0.809	0.933	0.393	0.528	0.522	0.235	0.079
	MCD (10%)	0.274	0.286	0.118	0.514	0.816	0.937	0.398	0.562	0.553	0.185	0.243
ViT-B	Baseline	0.222	0.232	0.094	0.616	0.892	0.969	–	–	–	–	–
	TTA	0.221	0.230	0.093	0.621	0.891	0.968	0.462	0.573	0.558	0.177	0.293
	LC	0.224	0.227	0.095	0.616	0.883	0.966	0.330	0.419	0.427	0.497	−0.283
	GNLL	0.223	0.225	0.094	0.619	0.892	0.972	**0.497**	**0.622**	**0.596**	**0.120**	**0.368**
	SE	**0.218**	0.230	0.094	0.625	0.889	0.967	0.448	0.546	0.543	0.243	0.143
	MCD (10%)	0.245	0.269	0.106	0.560	0.851	0.955	0.405	0.555	0.543	0.192	0.229
ViT-L	Baseline	0.223	0.238	0.096	0.588	0.906	0.980	–	–	–	–	–
	TTA	0.217	0.232	0.094	0.599	0.904	0.980	0.436	0.576	0.557	0.174	0.316
	LC	**0.215**	0.226	0.092	0.604	0.911	0.980	0.325	0.426	0.441	0.536	−0.307
	GNLL	0.229	0.244	0.099	0.588	0.888	0.973	**0.460**	**0.608**	**0.586**	**0.124**	**0.375**
	SE	0.235	0.252	0.098	0.594	0.893	0.974	0.442	0.574	0.569	0.187	0.341
	MCD (10%)	0.249	0.272	0.107	0.557	0.859	0.966	0.416	0.580	0.563	0.192	0.304

Regarding uncertainty quality, GNLL once again asserts itself as the most reliable methods. It consistently outperforms all other methods across the threshold-based metrics for all encoder sizes, achieving values up to 49.7% for p(acc|cer), 62.2% for p(unc|ina), and 59.6% for PAvPU. The threshold-free metrics confirm these findings, with GNLL yielding the lowest AUSE and highest Spearman correlation (ρ) in almost every configuration. While the other methods — particularly SE and MCD — remain competitive, LC consistently lags across all uncertainty metrics, exhibiting negative ρ values that underscore its instability as a stand-alone confidence measure in complex robotic scenes. These results further solidify GNLL as a robust and scalable UQ choice for foundation model MDE in diverse, real-world deployment scenarios.

4.4. Qualitative evaluation

To gain deeper insights into the predictive behavior and uncertainty calibration of the different UQ methods, we provide qualitative comparisons in Figs. 3 and 4. Overall, GNLL provides the most actionable uncertainty estimates for safety-critical deployment, corroborating our quantitative findings.

Visualizing Uncertainty and Error Correlation. Across all domains, we observe that the most robust UQ methods successfully align high uncertainty with regions of high predictive error. In Fig. 3, we observe that while all methods produce plausible metric depth maps, their uncertainty landscapes differ significantly. Overall, GNLL produces the most coherent uncertainty estimates, with high uncertainty for distant regions.

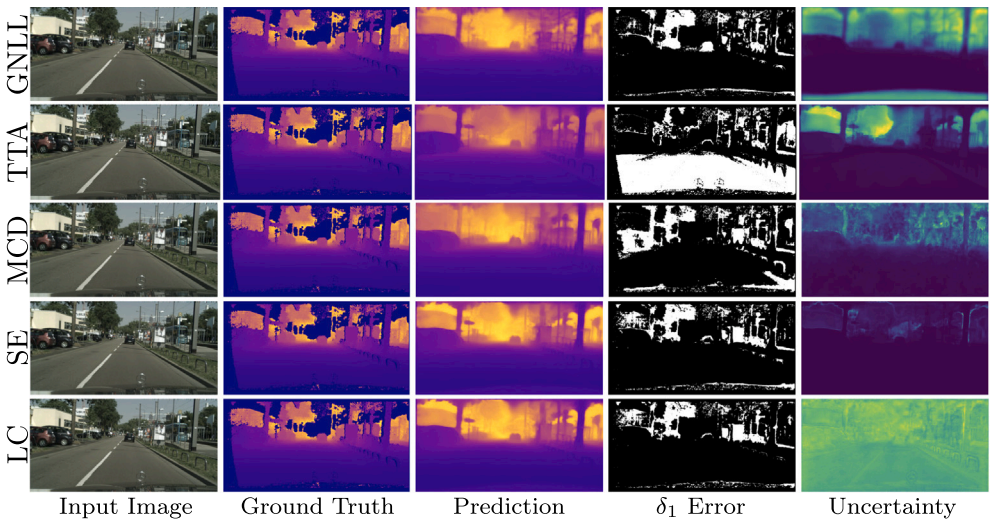


Fig. 3. Qualitative comparison of all uncertainty quantification methods (GNLL, TTA, MCD, SE, LC) on Cityscapes using a ViT-S encoder. Best viewed in color.

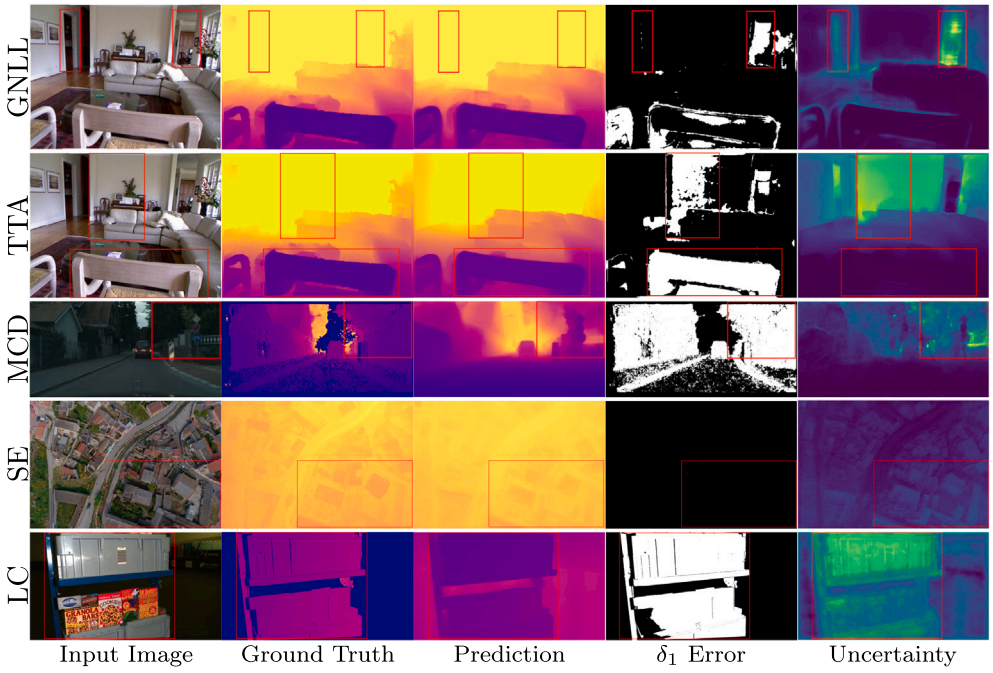


Fig. 4. Qualitative comparison of uncertainty quantification methods across indoor (Silberman et al., 2012), outdoor (Cordts et al., 2016), aerial (Nex et al., 2024), and robotics (Tyree et al., 2022) scenarios using a ViT-B encoder. Red boxes highlight regions of interest; best viewed in color.

Failure Mode Analysis. Fig. 4 illustrates how models behave under specific domain challenges:

- **Indoor (NYUv2):** GNLL excels at capturing geometric ambiguity. In the top row, uncertainty is noticeably elevated around object boundaries and open doors. While TTA provides reasonable background awareness, it fails to capture the significant error on the foreground chair backrest.
- **Outdoor (Cityscapes):** The MCD predictions (third row) exhibit localized errors in the top-right corner of the image. Notably, the uncertainty map highlights these exact areas, reinforcing the model's self-awareness of its limitations in difficult zones.
- **Aerial (UseGeo):** The SE results (fourth row) reveal that while the foundation model maintains plausible relative depth, it struggles with the absolute scale of architectural features. The uncertainty map effectively highlights these structures, suggesting that the model flags high-complexity, high-error regions within the aerial scene.
- **Robotics (HOPE):** The LC results (bottom row) highlight a critical failure: the model misestimates the absolute depth of the foreground object. However, the accompanying uncertainty map correctly flags the entire object, suggesting that the model's internal uncertainty is often more reliable than its metric regression output.

5. Scope and limitations

While our experiments demonstrate consistent performance across encoder capacities and application domains within the DepthAnythingV2 family, they do not automatically guarantee cross-model generalization. Switching foundation backbones can alter the baseline quality of both depth and uncertainty estimates. From an implementation

perspective, required effort is predominantly governed by the UQ method: following our proposed designs renders TTA, GNLL, LC, and SE largely architecture-agnostic, whereas MCD requires suitable dropout layers with deliberate placement. Consequently, transferring our findings to other foundation models — particularly architecturally similar discriminative MDE networks — is plausible but demands empirical verification. Generative, diffusion-based, or camera-conditioned architectures may exhibit fundamentally different uncertainty characteristics, requiring method-specific adaptations and dedicated controlled evaluations.

6. Conclusion

We evaluated five scalable UQ strategies for metric fine-tuning of DepthAnythingV2 across three encoder capacities and four diverse datasets, spanning indoor, urban, synthetic-to-real, and aerial imagery.

Among the evaluated methods, GNLL achieved the most favorable overall balance, consistently delivering superior uncertainty-error alignment while preserving predictive depth accuracy at negligible inference overhead. While MCD and SE showed competitive performance in specific settings, both introduced substantial computational overhead. Conversely, LC exhibited lower consistency across domains.

Overall, this study demonstrates that equipping large-scale foundation models with reliable uncertainty estimates is both feasible and practical, without compromising the potent geometric priors acquired during internet-scale pre-training. By establishing a practical blueprint for making MDE foundation models more interpretable and trustworthy, we hope to encourage future research to treat reliability as a primary objective alongside predictive accuracy. Ultimately, these insights lay the groundwork for safer AI deployment across safety-critical spatial understanding tasks, including robust 3D reconstruction and downstream visual perception.

CRediT authorship contribution statement

Steven Landgraf: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rongjun Qin:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Markus Ulrich:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Gemini in order to correct linguistic errors and improve readability. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge support by the state of Baden-Württemberg, Germany through bwHPC.

This work is supported by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition, Germany.

The work is partially supported by the Office of Naval Research, USA (N00014-23-1-2670).

References

Arampatzakis, V., Pavlidis, G., Mitianoudis, N., Papamarkos, N., 2023. Monocular depth estimation: A thorough review. IEEE Trans. Pattern Anal. Mach. Intell.

Ayhan, M.S., Berens, P., 2018. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In: Medical Imaging with Deep Learning.

Bhat, S.F., Alhashim, I., Wonka, P., 2021. Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4009–4018.

Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M., 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288.

Chen, W., Fu, Z., Yang, D., Deng, J., 2016. Single-image depth perception in the wild. Adv. Neural Inf. Process. Syst. 29.

Chen, W., Qian, S., Fan, D., Kojima, N., Hamilton, M., Deng, J., 2020. Oasis: A large-scale dataset for single image 3d in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 679–688.

Choi, H., Lee, H., Kim, S., Kim, S., Kim, S., Sohn, K., Min, D., 2021. Adaptive confidence thresholding for monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12808–12818.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3213–3223.

Dong, X., Garratt, M.A., Anavatti, S.G., Abbass, H.A., 2022. Towards real-time monocular depth estimation for robotics: A survey. IEEE Trans. Intell. Transp. Syst. 23 (10), 16940–16961.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. Adv. Neural Inf. Process. Syst. 27.

Fort, S., Hu, H., Lakshminarayanan, B., 2019. Deep ensembles: A loss landscape perspective. arXiv preprint arXiv:1912.02757.

Fort, S., Hu, H., Lakshminarayanan, B., 2020. Deep Ensembles: A loss landscape perspective. arXiv:1912.02757.

Franchi, G., Yu, X., Bursuc, A., Aldea, E., Dubuisson, S., Filliat, D., 2022. Latent discriminant deterministic uncertainty. In: European Conference on Computer Vision. Springer, pp. 243–260.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D., 2018. Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2002–2011.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (Eds.), Proceedings of the 33rd International Conference on Machine Learning. In: Proceedings of Machine Learning Research, Vol. 48, PMLR, New York, New York, USA, pp. 1050–1059, URL https://proceedings.mlr.press/v48/gal16.html.

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X., 2022. A Survey of Uncertainty in Deep Neural Networks. arXiv:2107.03342.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3354–3361.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: International Conference on Machine Learning. PMLR, pp. 1321–1330.

Hodaň, T., Sundermeyer, M., Drost, B., Labbé, Y., Brachmann, E., Michel, F., Rother, C., Matas, J., 2020. BOP challenge 2020 on 6D object localization. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, pp. 577–594.

Hodan, T., Sundermeyer, M., Labbe, Y., Nguyen, V.N., Wang, G., Brachmann, E., Drost, B., Lepetit, V., Rother, C., Matas, J., 2024. BOP challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5610–5619.

Hornauer, J., Belagiannis, V., 2022. Gradient-based uncertainty for monocular depth estimation. In: European Conference on Computer Vision. Springer, pp. 613–630.

Kalia, M., Navab, N., Salcudean, T., 2019. A real-time interactive augmented reality depth estimation technique for surgical robotics. In: 2019 International Conference on Robotics and Automation. ICRA, IEEE, pp. 8291–8297.

Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K., 2024. 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.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? Adv. Neural Inf. Process. Syst. 30.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 30, Curran Associates, Inc..

Landgraf, S., Hillemann, M., Kapler, T., Ulrich, M., 2025. A comparative study on multi-task uncertainty quantification in semantic segmentation and monocular depth estimation. Tm-Technisches Mess. 92 (7–8), 298–310.

Landgraf, S., Hillemann, M., Wursthorn, K., Ulrich, M., 2024a. Uncertainty-aware Cross-Entropy for semantic segmentation. ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci. 10, 129–136.

Landgraf, S., Hinz, J., Ulrich, M., 2026. A critical synthesis of uncertainty quantification and foundation models for semantic segmentation. ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci. 11, 673–680.

Landgraf, S., Wursthorn, K., Hillemann, M., Ulrich, M., 2024b. Dudes: Deep uncertainty distillation using ensembles for semantic segmentation. PFG–J. Photogramm. Remote. Sens. Geoinf. Sci. 92 (2), 101–114.

Lee, J., Feng, J., Humt, M., Müller, M.G., Triebel, R., 2022. Trust your robots! predictive uncertainty estimation of neural networks with sparse gaussian processes. In: Conference on Robot Learning. PMLR, pp. 1168–1179.

Lee, K., Lee, H., Lee, K., Shin, J., 2017. Training confidence-calibrated classifiers for detecting out-of-distribution samples. arXiv preprint arXiv:1711.09325.

Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S., 2017. Leveraging uncertainty information from deep neural networks for disease detection. Sci. Rep. 7 (1), 17816. http://dx.doi.org/10.1038/s41598-017-17876-z.

Li, Z., Snavely, N., 2018. Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2041–2050.

Li, Z., Wang, X., Liu, X., Jiang, J., 2024. Binsformer: Revisiting adaptive bins for monocular depth estimation. IEEE Trans. Image Process..

Liew, J.H., Yan, H., Zhang, J., Xu, Z., Feng, J., 2023. Magicedit: High-fidelity and temporally coherent video editing. arXiv preprint arXiv:2308.14749.

Loquercio, A., Segu, M., Scaramuzza, D., 2020. A general framework for uncertainty estimation in deep learning. IEEE Robot. Autom. Lett. 5 (2), 3153–3160.

Loshchilov, I., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

MacKay, D.J.C., 1992. A Practical Bayesian Framework for Backpropagation Networks. Neural Comput. 4 (3), 448–472. http://dx.doi.org/10.1162/neco.1992.4.3.448.

Masoumian, A., Rashwan, H.A., Cristiano, J., Asif, M.S., Puig, D., 2022. Monocular depth estimation using deep learning: A review. Sensors 22 (14), 5353.

Mi, L., Wang, H., Tian, Y., He, H., Shavit, N.N., 2022. Training-free uncertainty estimation for dense regression: Sensitivity as a surrogate. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 10042–10050.

Mukhoti, J., Gal, Y., 2018. Evaluating bayesian deep learning methods for semantic segmentation. arXiv preprint arXiv:1811.12709.

Naeini, M.P., Cooper, G., Hauskrecht, M., 2015. Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 29.

Nex, F., Stathopoulou, E., Remondino, F., Yang, M., Madhuanand, L., Yogender, Y., Alsadik, B., Weinmann, M., Jutzi, B., Qin, R., 2024. UseGeo-A UAV-based multi-sensor dataset for geospatial research. ISPRS Open J. Photogramm. Remote. Sens. 100070.

Nix, D.A., Weigend, A.S., 1994. Estimating the mean and variance of the target probability distribution. In: Proceedings of 1994 Ieee International Conference on Neural Networks. ICNN'94, Vol. 1, IEEE, pp. 55–60.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

Piccinelli, L., Sakaridis, C., Yu, F., 2023. Idisc: Internal discretization for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21477–21487.

Piccinelli, L., Yang, Y.-H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F., 2024. UniDepth: Universal monocular metric depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 10106–10116.

Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S., 2020. 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.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell. 44 (3), 1623–1637.

Roussel, T., Van Eycken, L., Tuytelaars, T., 2019. Monocular depth estimation in new environments with absolute scale. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 1735–1741.

Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A.D.N., et al., 2019. Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. NeuroImage 195, 11–22.

Saxena, S., Herrmann, C., Hur, J., Kar, A., Norouzi, M., Sun, D., Fleet, D.J., 2024. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. Adv. Neural Inf. Process. Syst. 36.

Shahbazi, M., Claessens, L., Niemeyer, M., Collins, E., Tonioni, A., Van Gool, L., Tombari, F., 2024. InseRF: Text-Driven generative object insertion in neural 3D scenes. arXiv preprint arXiv:2401.05335.

Shriram, J., Trevithick, A., Liu, L., Ramamoorthi, R., 2024. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. arXiv preprint arXiv:2404.07199.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from rgbd images. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. Springer, pp. 746–760.

Song, S., Lichtenberg, S.P., Xiao, J., 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 567–576.

Sundermeyer, M., Hodaň, T., Labbe, Y., Wang, G., Brachmann, E., Drost, B., Rother, C., Matas, J., 2023. Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2785–2794.

Tyree, S., Tremblay, J., To, T., Cheng, J., Mosier, T., Smith, J., Birchfield, S., 2022. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In: International Conference on Intelligent Robots and Systems. IROS.

Valdenegro-Toro, M., 2023. Sub-ensembles for fast uncertainty estimation in neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4119–4127.

Wan, S., Wu, T.-Y., Wong, W.H., Lee, C.-Y., 2018. Confnet: predict with confidence. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 2921–2925.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J., 2024. Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709.

Wüest, M., Engelmann, F., Miksik, O., Pollefeys, M., Barath, D., 2025. UnLoc: Leveraging depth uncertainties for floorplan localization. arXiv preprint arXiv:2509.11301.

Xiang, J., Wang, Y., An, L., Liu, H., Wang, Z., Liu, J., 2022. Visual attention-based self-supervised absolute depth estimation using geometric priors in autonomous driving. IEEE Robot. Autom. Lett. 7 (4), 11998–12005.

Xu, D., Jiang, Y., Wang, P., Fan, Z., Wang, Y., Wang, Z., 2023. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4479–4489.

Xue, F., Zhuo, G., Huang, Z., Fu, W., Wu, Z., Ang, M.H., 2020. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 2330–2337.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H., 2024a. 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.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024b. Depth anything V2. arXiv preprint arXiv:2406.09414.

Yu, X., Franchi, G., Aldea, E., 2021. Slurp: Side learning uncertainty for regression problems. arXiv preprint arXiv:2110.11182.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847.