TOWARD A BETTER UNDERSTANDING OF MONOCULAR DEPTH EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Monocular depth estimation is an important task with rapid progress, but how to evaluate it is not fully resolved, as evidenced by a lack of standardization in existing literature and a large selection of evaluation metrics whose trade-offs and behaviors are not fully understood. This paper contributes a novel, quantitative analysis of existing metrics in terms of their sensitivity to various types of perturbations of ground truth, emphasizing comparison to human judgment. Our analysis reveals that existing metrics are severely under-sensitive to curvature perturbation such as making smooth surfaces bumpy. To remedy this, we introduce a new metric based on relative surface normals, along with new depth visualization tools and a principled method to create composite metrics with better human alignment. All code, data, and tools will be open-sourced.

1 INTRODUCTION

Monocular depth estimation (MDE) is the task of estimating pixelwise depth from a single RGB image. It has become a standard task with rapid progress, due to its importance for many applications, such as robotics, AR/VR, and the generation of images, videos or 3D worlds.

However, how to evaluate monocular depth estimation is not fully resolved. This is reflected by a lack of standardization in existing literature, and a very large menu of evaluation metrics. It is in fact possible for two papers (e.g., Bochkovskiy et al. (2025); Wang et al. (2025a)) to report 5 to 7 different metrics each but with zero overlap[1].

Even when models are compared under the same set of metrics, it is often not clear how to interpret the results. If a model performs well under one metric (say AbsRel after affine alignment on disparity) but poorly under another (say δ^1 with no alignment), what does it mean? Often there lack clear answers, because the trade-offs between different metrics are not fully understood.

The large selection of metrics arises from a combination of decisions that need to be made when designing an evaluation metric. For example, the decisions can include how to compute error of individual depth values against ground truth (by difference in depth, log depth, or inverse depth), whether and how to discount error for faraway points, whether to binarize the error and if so with what threshold, and how to account for the unknown global scale through alignment (in depth or inverse depth), and whether to also allow an unknown global shift.

The complexity of these decisions is due to several issues inherent to monocular depth estimation that preclude straightforward comparisons of individual depth values. One is scale ambiguity. The global scale of a scene is fundamentally ambiguous: a scene can be a miniature replica of a larger identical scene. This scale ambiguity can also occur locally in a scene: it can be impossible to tell from a single image whether a drone in the sky is a big one far away or small one close by. Even without any scale ambiguity, another issue is the unbounded range of possible depth values in the same scene—consider an ocean that extends to the horizon. In such cases, some form of normalization or weighting is necessary to prevent errors on faraway objects from dominating.

To make matters worse, the predicted depth values alone do not reconstruct a 3D shape because it does not tell us the X and Y coordinates in 3D; for that we also need the camera intrinsics, which are often unknown and need to be predicted. But camera intrinsics can be ambiguous from a single

[1]Two metrics can share the same name but are different due to differences in depth alignment.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

image and the prediction can be reasonable but off. And the predicted depth and predicted camera intrinsics can together give a reasonable 3D shape but have large errors when evaluated separately.

The discussion above suggests that the evaluation of monocular depth is inherently complex. The existence of many metrics reflects best efforts to address such complexity. On the other hand, there has been limited study on the trade-offs of these metrics, and research on monocular depth could benefit from additional investigation.

In this paper we seek to improve our understanding of monocular depth evaluation. Our goal is to study various evaluation metrics and shed light on their trade-offs and behaviors, and to develop a principled method to customize or combine metrics to align with particular preferences, such as those based on human perception.

It is important to note that our goal is *not* a new "perfect" metric that is better in all aspects and supersedes all existing metrics. A single ideal metric probably does not exist and the best metric is likely application dependent. We only seek to offer additional insights on the behaviors of existing metrics and how they might be customized and combined toward particular goals.

Our Approach: Our main approach is to quantify the sensitivity of each metric to various types of perturbations of the ground truth depth. In particular, we focus on 6 types of perturbations—surface orientation, camera intrinsics, relative scale, curvature, affine transform, and boundary—which are interpretable and representative changes to ground truth depth that could shed light on the behavior of a metric. The definition of each perturbation is given in Section 3.

Given these perturbations, we can measure the sensitivity of a metric to each perturbation, and compare the sensitivity between metrics. The basic idea is to measure the numerical responses of a metric under various perturbations, and compare such responses across metrics, after proper normalization to account for the arbitrary choice of measuring units.

Human Sensitivity: We also apply our sensitivity analysis to human judgment. Human judgment can serve as a potentially useful reference for sensitivity comparisons because (1) many generative applications produce content for human consumption and (2) the human visual system remains arguably the best general-purpose depth perception system and human judgment may provide clues on what is and is not important for achieving human-level visual capabilities. Note that we do *not* claim that human judgment is the gold standard or an ideal proxy for all or some downstream applications; it is simply a potentially useful reference.

We measure the sensitivity of human judgment to various perturbations. Using a collection of synthetic test scenes and varying amounts of perturbations, we ask human annotators to judge whether a (possibly) perturbed depth map is the ground truth of a given RGB image. We then average the binary annotations to estimate the exchange rates of human judgment versus other metrics.

To help human annotators examine the reconstructed geometry, we introduce two new visualization tools: Textureless Relighting and Projected Contours. These new tools overcome the limitations of conventional visualizations such as textured point clouds, which can mask geometric artifacts because, like in video games, texture maps can create fake perceived geometry. Beyond measuring human sensitivity, the tools can also be useful in developing future depth models.

Using human judgment as a reference yields interesting findings about existing metrics. Humans are sensitive to affine transforms of depth or disparity, but many metrics perform affine alignment and are thus completely insensitive. Most notably, all widely used metrics have very poor sensitivity to curvature perturbation (e.g., making a smooth surface bumpy).

Sensitivity Aligned Composition: In addition to improving our understanding, our sensitivity analysis enables us to combine existing metrics to align with a specific sensitivity profile, such as that of humans or a downstream application. The basic idea, which we call "sensitivity aligned composition (SAC)", is to combine existing metrics through some parametric form of composition such as a weighted average, and optimize the composition parameters such that the combined metric achieves the desired sensitivities to a given set of perturbations.

Because existing metrics are overly insensitive to curvature perturbation compared to human judgment, we propose a new metric RelNormal, which is based on relative surface normals and thus sensitive to curvature perturbation. Combining RelNormal and a subset of existing metrics, we propose SAWA-H (Sensitivity Aligned Weighted Average based on Human judgment), a new, composite metric that

aligns better with human judgment than all existing metrics. SAWA-H can supplement existing metrics in cases where alignment with human judgment is desired.

Summary of Contributions: Our contributions are three fold. (1) We conduct a novel, quantitative sensitivity analysis of the commonly used metrics under various perturbations. (2) We introduce two new tools of depth visualizations and measure sensitivity of human judgment, revealing that existing metrics are overly insensitive to curvature perturbation. (3) We introduce SAWA-H, a new metric that better aligns with human judgment via optimized combination of a set of base metrics; we also introduce RelNormal, a new base metric designed to enable better human alignment. All code, data, and tools will be open-sourced.

2 Related Work

Monocular Depth Evaluation Metrics. Conventional metrics directly evaluate the difference between predicted and ground truth depth (Zoran et al., 2015; Chen et al., 2016; Eigen et al., 2014; Saxena et al., 2009). Koch et al. (2018); Chen et al. (2019); Örnek et al. (2022); Talker et al. (2024); Pham et al. (2025); Bochkovskiy et al. (2025) noted that these metrics are not sensitive to over-smooth boundaries and error at predicting planes, and proposed new metrics to evaluate boundary sharpness and difference in 3D. Despite these efforts, there lacks a principled way to systematically study behavior of different metrics. We attempt to fill this gap.

Monocular Depth Estimation Methods. There is rapid development in MDE recently. It is worth noting that a decent number of methods predict affine-invariant depth or disparity (Birkl et al., 2023; Yang et al., 2024a;b; Ke et al., 2024; Fu et al., 2024), and during evaluation, they perform affine alignment on depth or disparity before computing metrics. We find that affine-aligned metrics are misaligned with human judgment, which is sensitive to geometric distortions caused by affine perturbations of depth.

Monocular Depth Estimation Benchmarks Conventional MDE benchmarks (Chen et al., 2016; 2020; Guizilini et al., 2020; Vasiljevic et al., 2019; Schöps et al., 2019; Jung et al., 2023; Koch et al., 2020; Uhrig et al., 2017; Silberman et al., 2012; Obukhov et al., 2025) typically evaluate models using multiple metrics. This makes interpreting results hard, especially when a model performs well under one metric but poorly under another, and behaviors of metrics are not fully understood. Our analysis aims at providing better understanding of evaluation results. Li et al. (2025) directly evaluate models on downstream applications. We propose sensitivity aligned composition, which can also develop metrics aligned with preference of downstream applications, but via combining existing metrics. This may be favorable when downstream application evaluation is computationally expensive.

Human Sensitivity and Visualization Tools Humans exhibit some peculiarities when judging 3D structure. Chen et al. (2020) observed that annotators judged relative normals correctly but often made mistakes when estimating the overall orientation. Linsley et al. (2025) find that compared with neural networks, humans are much better at answering "can object A see object B?". This suggests that standard evaluations fail to capture the aspects of a scene that humans are sensitive to.

When visualizing depth, previous works typically plot heatmaps of either depth or disparity (Birkl et al., 2023; Eigen et al., 2014; Zoran et al., 2015). Other works also display novel views of textured point cloud un-projected from depth map (Wang et al., 2025a; Yin et al., 2021). We show in Sec. 4.1 that these tools can mask some geometric artifacts. Wang et al. (2025a); Cao et al. (2022) additionally display a gray mesh with shadows and specular highlights. Liu & Todd (2004) finds that such normal shading greatly improve humans' ability to discern curvature. We further expand on these in Sec. 4.1.

3 Sensitivity Analysis

Definition of evaluation metric Without loss of generality, we define a evaluation metric to be a function that maps a question and an answer to a non-negative score that represents the quality of the answer. For monocular depth, the question is a RGB image and the answer is a depth map. Sometimes the answer could also include predicted camera intrinsics.

This definition is general and does not assume access to a ground truth answer, because there may not be a unique correct answer. The metric can be a neural network or a human. On the other hand, an evaluation metric, like all existing monocular depth metrics, can use privileged information including a ground truth answer and return a score by comparing the answer to the ground truth.

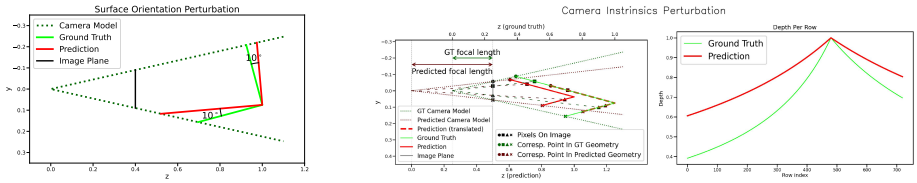

Figure 1: **Left:** Surface orientation perturbation. Orientation of prediction (red) equals rotating that of ground truth (green) around x-axis by $-10°$. **Middle & Right:** Camera intrinsics perturbation. Middle: Prediction is made under wrong focal length (2x ground truth focal length), while predicted geometry (red) has shape similar to ground truth (green). The former overlaps well with the latter through translation (red dashed). Right: Predicted depth substantially differs from ground truth depth.

Inherent Ambiguity in Monocular Depth: It is worth noting that monocular depth estimation has inherent ambiguity: there always exist multiple possible 3D shapes that project to the same 2D image, and some are considered more plausible only because of priors. For example, the input image can depict an actual castle or a printed photo of the same castle, and there can be no way to tell them apart.

In addition, the actual physical scene that generated the 2D image can be unreasonable as ground truth. Consider the Ames room (wik), which is trapezoidal-shaped but appears rectangular to the human eye. It would be unreasonable for the ground truth to be trapezoidal, because this would lead to a monocular depth system that reconstructs trapezoids even for actual rectangular rooms.

This inherent ambiguity suggests that (1) it is more fitting to use the more general definition of evaluation metric that does not assume a single known ground truth, and (2) human perception can be a reasonable reference "ground truth", because the inherent ambiguity makes it impossible to always correctly recover the actual physical ground truth.

Perturbations: We study the sensitivity of metrics to a set of interpretable perturbations.

- **Surface orientation perturbation** refers to perturbing a depth map such that the surface normal of each pixel is rotated by the same amount (or close to the same amount). The motivation is that the absolute orientation of surfaces can be ambiguous to humans (Chen et al., 2020). Fig. 1 Left shows an example.
- **Camera intrinsics perturbation** refers to perturbing both focal length and depth while keeping the 3D shape similar in terms of surface normals. The motivation is that prediction of focal length can be off but the reconstructed shape can still be reasonable. Fig. 1 Middle & Right shows an example: although the shape is similar, perturbed depth map differs with ground truth substantially.
- **Relative scale perturbation** refers to perturbing relative scale between objects, e.g. scaling down the depth of a floating foreground object while keeping the background the same. The motivation is that the scale of a (floating) foreground object can be ambiguous to determine relative to the background.
- **Curvature perturbation** refers to perturbing the curvature of back-projected 3D surfaces, e.g. making a smooth surface bumpy. We perform curvature perturbation under two spatial frequencies. The motivation is the existence of smooth surfaces in nature (e.g., sand dunes, egg shells) and built environments. Please refer to the supplement for details.
- **Affine transform perturbation** refers to perturbing ground truth depth by an affine transform of depth or disparity. The motivation is that many existing methods Yang et al. (2024a;b); Fu et al. (2024); Ke et al. (2024); Birkl et al. (2023) predict affine invariant depth or disparity.
- **Boundary perturbation** refers to blurring depth values across occlusion boundaries, resulting in distortion of geometry near boundaries. The motivation is that some depth methods produce sharper boundaries, which are considered more desirable.

Supplement A,G provide details of the perturbation algorithms and examples of perturbed depth.

Exchange rates: For each metric and perturbation type, we measure the numerical response of the metric when applying that type of perturbation to ground truth. Then, we compare the responses of different metrics across different perturbation types to examine which metric is more or less sensitive to which perturbation. To facilitate this comparison, we introduce the notion of exchange rate.

The exchange rate between metrics A and B under perturbation P, $R(A; B|P)$ is (informally) defined as the ratio of changes of metrics A and B when perturbing the ground truth depth by the same

intensity of perturbation P. Without loss of generality, we assume all metrics can be standardized such that the ground truth depth is given a score of 0 and the score increases as the depth deviates from ground truth.

For example, if after the same perturbation, the value of metric A (say RMSE) increases from 0 to 2 and value of metric B (say AbsRel) increases from 0 to 0.5, the exchange rate between metrics A and B is $2/0.5 = 4$. This rate can be interpreted as: to represent the same intensity of perturbation P, 4 units of metric A are needed, while we only need 1 unit of metric B.

This intuitive notion of exchange rate is well defined if both metric A and B respond linearly to the intensity of the perturbation P. However, this is almost never the case. For example, some metrics are bounded and will eventually saturate for large perturbations. Therefore our formal definition of the exchange rate needs to also handle nonlinear functions.

Formally, we treat metrics $A(x)$ and $B(x)$ as smooth functions of intensity x of perturbation P, and define the exchange rate $R(A; B|P)$ to be the ratio between their derivatives at zero:

$$R(A; B|P) \triangleq A'(0)/B'(0). \tag{1}$$

In other words, we approximate each metric as a linear function of perturbation intensity in a small neighborhood around zero. This approximation is justifiable because as models improve, the neighborhood around zero becomes more important.

In practice, we approximate each metric with the least squares fit of a quadratic function $ax^2 + bx$ ($b \geq 0$), and we take the derivative of the quadratic function at zero. Compared with a linear form, the quadratic form better approximates plateauing metrics and removes the need to manually select the size of the neighborhood around zero. $b \geq 0$ is considered to enforce a non-negative derivative at 0.

Comparing Sensitivity: It may appear that metric A is more sensitive than metric B to perturbation P if the exchange rate $R(A; B|P)$ is large, as metric A has a larger response. But this impression is false because units of metrics A and B can be arbitrarily chosen. For example, metric B and metric A can be identical except that A is measured in meters and B is in millimeters. Then the exchange rate is 1000, but metrics A and B are effectively identical.

This means that a single exchange rate $R(A; B|P)$ between A and B under perturbation P is not meaningful in isolation and must be interpreted in comparison with another exchange rate. Specifically, let $R(A; B|Q)$ be the exchange rate of the same metrics A and B under a new perturbation Q, using the same units of metric A and B. If the exchange rate under P (for example 1000:1) is bigger than that under Q (for example 2:1), then we can conclude that relative to B, metric A is more sensitive to perturbation P than Q. This is because 1 unit worth of perturbation P measured under metric B translates to 1000 units under metric A, whereas 1 unit worth of perturbation Q measured under metric B translates to only 2 units under metric A. In other words, compared to metric B, perturbation P causes larger numerical changes to metric A than perturbation Q.

Dataset for measuring sensitivity. To estimate the exchange rate in practice, we need a dataset of 3D scenes with ground truth depth maps. This is because the sensitivity of a metric is often scene dependent, in which case we obtain the expected sensitivity averaged over multiple scenes.

We create a dataset of synthetic scenes using Infinigen (Raistrick et al., 2023; 2024), a procedural generator of photorealistic nature and indoor scenes. We choose to use synthetic data because real-world data are limited in the diversity of scenes and availability of dense depth ground truth. For example, no pixelwise real-world ground truth is available for large natural scenes. We choose Infinigen because it covers both indoor and natural scenes and is easy to customize.

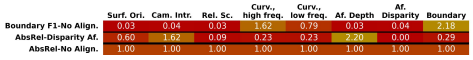

Figure 2: Exchange rate between each metric and AbsRel with no alignment. Because absolute scale of the exchange rate is not important, we normalize each row for better visualization. Perturbations from left to right: surface orientation, camera intrinsics, relative scale, curvature (high frequency), curvature (low frequency), affine transform of depth, affine transform of disparity, and boundary.

The dataset consists of 95 scenes. Among them, 56 are indoor scenes, which have short depth range and more regular shape, and 39 are nature scenes, which have long depth range and less regular shape. For each perturbation type, we perturb by at least 6 different intensities, with a total of 5320 perturbed depth in the dataset. Example scenes are included in Supplement G.

An Example of Comparison of Existing metrics We show an example of using our methodology to compare sensitivity between 3 metrics: Boundary F1-No Align (Boundary F1 (Bochkovskiy et al., 2025) with no alignment), AbsRel-Disparity Af (AbsRel (Saxena et al., 2009) after affine alignment of disparity), and AbsRel-No Align (AbsRel with no alignment). Fig. 2 shows the exchange rate with respect to AbsRel-No Align. Comparing Boundary F1-No Align and AbsRel-No Align, the exchange rate under boundary perturbation is the highest. This is expected as Boundary F1 is designed to capture boundary sharpness which is not reflected in AbsRel and other metrics. Comparing AbsRel-Disparity Af and AbsRel-No Align, exchange rate under affine transform of disparity is the lowest. This is also expected as affine alignment on disparity can perfectly align depth of this perturbation to ground truth.

4 Human Sensitivity

We are particularly interested in how the sensitivity of existing metrics compares to human judgment. This is a useful question because many generative AI applications generate content for human consumption, and for such applications human judgment is highly relevant. In addition, findings of human sensitivity can reveal what is and is not important for achieving human-level visual capabilities, which might be sufficient for robotic applications.

To measure human sensitivity, we seek to model human judgment as a smooth non-negative function $H(x)$ of the intensity x of a perturbation. This is tricky because it can be difficult for humans to produce consistent and well-calibrated numerical values, e.g., rate the depth map from 1 to 5.

Fortunately, we only care about the behavior near zero, which admits a simple solution. We show an RGB image and visualizations of a depth map to a human annotator for a binary response: whether the depth map is the ground truth. The depth map is perturbed with varying intensities—we probe when a perturbation starts to become noticeable, which corresponds to a response of 1. We collect data points from multiple human annotators and scenes and fit a quadratic curve to obtain the derivative at zero.

Note that unlike computer metrics, which relies on privileged access to ground truth answers, humans are not given two depth maps to compare, because two depth maps can be obviously different but indistinguishable in the terms of reconstructing the input image. Consistent with our definition of an evaluation metric, humans are shown an RGB image and visualizations of a (possibly) perturbed depth map and asked to judge whether the depth map is the correct reconstruction.

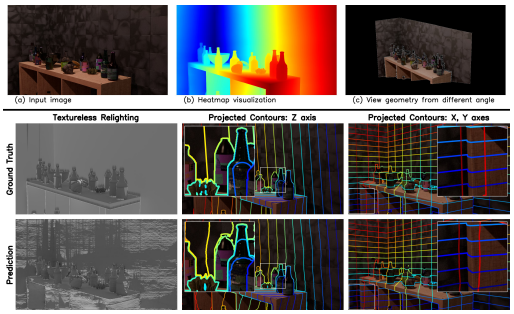

Figure 3: Top: input image and visualization of a depth using existing tools. Defects like a wavy wall are not apparent due to interference of texture. Bottom: Textureless Relighting and Projected Contours make geometric defects more apparent. Bumps on the wall are obvious.

4.1 Visualization Tools

When measuring human sensitivity, one issue we encounter is the lack of effective depth visualization tools. Widely used visualizations include (1) a heatmap and (2) textured 3D point cloud viewed in a few new angles. Though these visualizations can illustrate coarse shape and severe deformations, they can mask many other defects. While heatmaps are widely considered insufficient, the limitation of textured point clouds is perhaps less well understood. The main issue with textured point clouds is that to the human eye, textures can create illusions of geometry that differs from the actual one, in the same way videos games use textures maps to fake geometrical details on flat surfaces. Consider the depth map visualized with these standard methods in Fig. 3. Is the wall flat? From only these two visualizations, it appears flat but is in fact wavy. To help human annotators more efficiently and effectively inspect the reconstructed 3D shape, we introduce two new visualization tools:

Textureless Relighting. When casting light to a geometry, different shape results in different patterns of shadow. Motivated by this, we develop Textureless Relighting by casting directional light to depth induced mesh. To prevent texture misleading users, the mesh is made textureless. Moreover, relighting under one direction of light is not enough, as it is hard to tell the shape of pixels in shadow. So the mesh is rendered under various lighting directions. Fig. 3 Bottom Col. 1 shows an example, where bumps are easily noticeable under textureless relighting.

	Sensitivity								Similarity to Human Sensitivity	SAWA-H weights	
	Surf. Ori.	Cam. Intr.	Rel. Sc.	Curv., high freq.	Curv., low freq.	Af. Depth	Af. Disparity	Boundary		rescaled metric units	original metric units
AbsRel-Disparity Af.	0.29	0.62	0.05	0.00	0.00	2.74	0.00	0.00	0.46	0.00	0.00
δ^1-Depth Af. (Lst. Sq.)	0.65	0.37	0.00	0.00	0.00	0.00	2.73	0.04	0.47	0.00	0.00
RMSE(log)-Depth Af. (Lst. Sq.)	0.63	0.33	0.08	0.00	0.00	0.00	2.74	0.02	0.47	0.00	0.00
δ^1-Depth Af.	0.64	0.46	0.20	0.00	0.00	0.00	2.71	0.06	0.51	0.00	0.00
AbsRel-Depth Af. (Lst. Sq.)	0.67	0.38	0.29	0.00	0.01	0.00	2.71	0.05	0.51	0.00	0.00
RMSE-Depth Af. (Lst. Sq.)	0.73	0.38	0.12	0.00	0.01	0.00	2.70	0.17	0.51	0.00	0.00
RMSE-Depth Af.	0.83	0.44	0.16	0.00	0.01	0.00	2.66	0.13	0.53	0.00	0.00
δ^1-Disparity Af.	0.71	0.59	0.24	0.00	0.00	2.66	0.00	0.04	0.53	0.00	0.00
RMSE-Disparity Af.	0.61	1.04	0.03	0.00	0.00	2.56	0.00	0.00	0.53	0.00	0.00
RMSE(log)-Depth Af.	0.80	0.76	0.58	0.01	0.02	0.00	2.54	0.11	0.60	0.00	0.00
RMSE(log)-Disparity Af.	0.92	1.02	0.45	0.00	0.01	2.43	0.00	0.06	0.61	0.00	0.00
WKDR-No Align.	2.35	1.14	1.07	0.05	0.17	0.00	0.00	0.11	0.61	0.19	3.65
AbsRel-Depth Af.	1.21	0.79	0.77	0.01	0.03	0.00	2.31	0.06	0.65	0.00	0.00
δ^1-Depth Sc.	0.54	0.50	0.55	0.00	0.00	1.90	1.89	0.02	0.67	0.00	0.00
RMSE(log)-Depth Sc.	0.65	0.50	0.45	0.00	0.01	1.66	2.09	0.05	0.68	0.00	0.00
RMSE(log, scale-inv.)-No Align.	0.69	0.55	0.45	0.00	0.01	1.58	2.13	0.06	0.68	0.00	0.00
AbsRel$_p$-Pnt Af.	0.61	0.69	0.41	0.00	0.01	1.82	1.92	0.02	0.69	0.00	0.00
RMSE-Depth Sc.	0.85	0.55	0.41	0.00	0.00	1.59	2.07	0.09	0.69	0.00	0.00
AbsRel-Depth Sc.	0.78	0.65	0.48	0.00	0.01	1.78	1.90	0.02	0.70	0.00	0.00
$\delta^{0.125}$-Depth Af.	1.63	1.06	1.23	0.00	0.00	0.00	1.64	0.06	0.70	0.00	0.00
$\delta^{0.125}$-Disparity Af.	1.53	1.05	1.48	0.00	0.00	1.54	0.00	0.05	0.71	0.14	0.18
δ^1-No Align.	0.48	0.52	1.20	0.00	0.00	1.66	1.82	0.01	0.71	0.00	0.00
RMSE(log)-No Align.	0.71	0.58	0.76	0.00	0.01	1.52	2.07	0.05	0.71	0.00	0.00
RMSE-No Align.	1.18	0.53	0.50	0.00	0.00	1.38	2.04	0.09	0.72	0.00	0.00
$\delta^{0.125}$-Depth Af. (Lst. Sq.)	1.59	1.07	1.33	0.00	0.00	0.00	1.59	0.15	0.72	0.01	0.01
AbsRel-No Align.	0.72	0.57	0.83	0.00	0.01	1.87	1.73	0.02	0.72	0.00	0.00
AbsRel$_p$-Pnt sc.	0.73	1.19	0.44	0.00	0.01	1.65	1.77	0.02	0.73	0.00	0.00
$\delta^{0.125}$-Depth Sc.	1.26	1.01	0.69	0.00	0.00	1.49	1.64	0.03	0.77	0.00	0.00
$\delta^{0.125}$-No Align.	1.21	0.93	1.25	0.00	0.00	1.35	1.51	0.03	0.78	0.00	0.00
Boundary F1-No Align.	0.58	0.63	0.61	0.13	0.20	1.41	1.97	0.99	0.81	0.19	0.20
New Metric: RelNormal	0.52	0.55	0.03	1.70	1.28	1.03	1.21	0.61	0.87	0.48	1.94
New Metric: SAWA-H (w/o RelNormal)	1.40	0.96	0.95	0.12	0.21	1.16	1.47	0.79	0.88	NA	NA
New Metric: SAWA-H	1.27	0.93	0.68	1.05	0.85	1.21	1.18	0.63	0.97	NA	NA
Human	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	NA	NA

Figure 4: **Sensitivity:** Each row in red is a vector of exchange rates between a metric and human judgment, reflecting the sensitivity of a metric to different perturbations, using human judgment as a reference. Within a vector, a higher value means higher sensitivity of the metric to a perturbation, relative to humans. Human judgment (last row) against itself results in a vector of all 1s. Each vector can be arbitrarily scaled, equivalent to changing the unit of the metric; here the unit of the metric is rescaled such that the L2 norm of the vector is the same as a vector of all 1s, to facilitate comparisons across rows. All rows except the last four are existing metrics. **Similarity to human sensitivity:** cosine similarity between each sensitivity vector and the human sensitivity vector (all 1s). **SAWA-H Weights:** SAWA-H is a new, composite metric that is a weighted average of base metrics, with weights optimized to align with human sensitivity. RelNormal is a new base metric we introduce to allow better human alignment. The first column lists weights for averaging metrics under rescaled units, reflecting relative contributions invariant to the original choices of metric units; the second column lists equivalent weights for averaging metrics under their original units.

Projected Contours. This visualization is made by projecting contour lines to the 3D geometry un-projected from depth. Contour lines are projected along the direction of X,Y,Z axes. The shape of the contour lines can reveal the curvature of surfaces. For example, in Fig. 3 Bottom Col. 2,3, contour lines on flat planes should be straight (first row) and wavy on bumpy surface (second row).

4.2 Sensitivity of Existing Metrics Relative to Human Judgment

Metrics: Using the dataset described in Sec. 3, we compare the sensitivity of 9 widely used metrics against human judgment: AbsRel, AbsRel$_p$, $\delta^{0.125}$, δ^1, RMSE, RMSE (log), RMSE (log, scale invariant), WKDR, and boundary F1. Because it is common to align prediction with ground truth before computing metrics, we compute these metrics with no alignment, or with one of the following alignments if applicable: scale alignment of depth, affine alignment of depth (L1 or least square), affine alignment of disparity, and scale or affine alignment of point map (Ke et al., 2024; Wang et al., 2025a). More details are in Supplement C.

Collecting Data of Human Judgment: Human judgment is collected on a randomly selected subset of the dataset, which contains 5 or more scenes for each nonzero perturbation intensity of each perturbation type. To control annotation quality, we randomly select 25 depth maps as gold standard. We also mix 13 ground truth depth maps in the annotation dataset to detect annotators who rate everything as "not ground truth". Annotations from annotators who do not reach the accuracy of 70% on the gold standard or who label more than 30% of ground truth depth maps as "not ground truth" are discarded. 6 experts in computer vision participated in the annotation. 5 passed the quality test, and annotation on a total of 315 depth maps are collected. More details are in Supplement B.

Results: Fig. 4 shows the exchange rates between every metric and human judgment under various perturbations, revealing two main findings: (1) relative to humans, existing metrics (all rows except the last four) are severely under-sensitive to the curvature perturbation; (2) humans are sensitive to affine

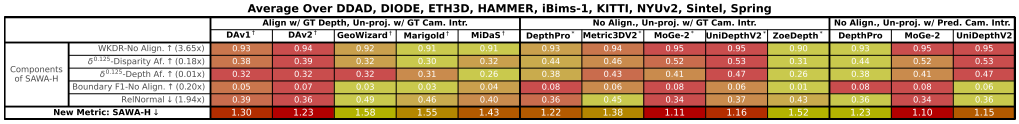

Figure 5: Performance of state-of-the-art monocular depth estimation methods under SAWA-H, averaged over 9 datasets. For component metrics of SAWA-H, the weight used to combine them to obtain SAWA-H under original metric units are listed next to metric names. *Align w/ GT Depth, Un-proj. w/ GT Cam. Intr.* refers to methods whose output is first aligned with ground truth depth then un-projected using ground truth camera intrinsics; *No Align., Un-proj. w/ GT Cam. Intr.* refers to methods whose output is directly un-projected using ground truth camera intrinsics; *No Align., Un-proj. w/ Pred. Cam. Intr.* refers to methods whose output is directly un-projected using predicted camera intrinsics. Note that results are not directly comparable if one method is aligned with ground truth or un-projected using ground truth camera intrinsics while the other is not.

transforms of depth or disparity, which is ignored by metrics that perform the respective affine alignment. We use simple shapes to illustrate the theoretical causes of these observations in Supplement D.

5 Sensitivity Aligned Composition

Our sensitivity analysis of existing metrics reveals that no existing metrics align well with human judgment. How do we create a metric that does align well, if one is desired for some reason? We introduce Sensitivity Aligned Composition (SAC), a method to compose a set of existing metrics such that the new metric achieves a desired sensitivity profile, such as that of humans or a specific application. The idea is to combine existing metrics to form a new metric, with the combination parameters optimized such that new metric is maximally similar to a target metric (which can be a black-box oracle like human judgment) in terms of sensitivity.

Given perturbations $P_1, \ldots, P_M$, we define the sensitivity vector of a metric A relative to a reference metric Z as a vector of exchange rates under the perturbations $\mathbf{R}(A; Z|\mathbf{P}) \triangleq (R(A; Z|P_1), \ldots, R(A; Z|P_M)) \in \mathbb{R}^M$. We then define the similarity of two sensitivity vectors using cosine similarity. This is because the magnitude of the sensitivity vector does not matter and is due to the choice of the unit of the metric. Let $\mathbf{T} \in \mathbb{R}^M$ be a target sensitivity vector, we can optimize the parameters w of a composite metric $C(w)$ to maximize the cosine similarity between $\mathbf{R}(C(w); Z|\mathbf{P})$ and $\mathbf{T}$.

Sensitivity Aligned Weighted Average (SAWA): The simplest way to compose metrics is a non-negatively weighted average $C(w) \triangleq \sum_i w_i A_i$ of metrics A_i. Due to the linearity of derivative, it is easy to verify that the sensitivity vector of metric $C(w)$ relative to a reference metric Z is a linear combination of the sensitivity vectors of metrics A_i, that is, $\mathbf{R}(C(w); Z|\mathbf{P}) = \sum_i w_i \mathbf{R}(A_i; Z|\mathbf{P})$.

Setting the reference metric to human judgment H and the target sensitivity vector to $\mathbf{1}_M$ (a vector of all 1s) leads to the following optimization problem that maximizes human alignment:

$$\max_{\mathbf{w} \geq 0} \frac{\langle \sum_i w_i \mathbf{R}(A_i; H|\mathbf{P}), \mathbf{1}_M \rangle}{\| \sum_i w_i \mathbf{R}(A_i; H|\mathbf{P}) \|_2 \| \mathbf{1}_M \|_2} \tag{2}$$

This is equivalent to a convex problem which can be solved optimally and efficiently. We refer to the resulting composite metric as SAWA-H (Sensitivity Aligned Weighted Average aligned to Human judgment). Intuitively, this optimization finds the best non-negative weights to linearly combine the red rows in Fig. 4 to maximize cosine similarity with a target vector of all 1s.

New metric: RelNormal. Because all existing metrics are insensitive to curvature error, this limits how much we can align with human judgment. Intuitively, the rows of existing metrics in Fig. 4 cannot linearly combine to approximate well a vector of all 1s. To remedy this we introduce RelNormal, a new metric based on relative surface normals.

RelNormal considers the relation between surface normals described by patches (sets of pixels) within the depth map. Given patches p and q, it first calculates the normal of the surface described by each patch $\hat{n}_p, \hat{n}_q, n_p, n_q$ for both ground truth (denoted by $\hat{n}_p, \hat{n}_q$) and prediction (denoted by n_p, n_q), and the angle between these normals $\angle(\hat{n}_p, \hat{n}_q), \angle(n_p, n_q)$. It then takes the mean angle difference over many such patches, normalized by $\frac{1}{\pi}$,

$$\text{RelNormal} = \frac{1}{\pi} \cdot \frac{1}{|\mathcal{C}|} \cdot \sum_{p,q \in \mathcal{C}} \left| \angle(\hat{n}_p, \hat{n}_q) - \angle(n_p, n_q) \right|, \tag{3}$$

If p and q lie on a continuous surface connected by a circular arc, this measures error in the degree of curvature. For p and q on separate but adjacent surfaces, this metric captures the local geometry (e.g. the angle between two walls in a room). Details of the RelNormal metric are in Supplement E.

SAWA-H (with RelNormal): With RelNormal, we are able to achieve substantially better human alignment, improving the cosine similarity to 0.97 (with RelNormal) from 0.88 (without RelNormal). Fig. 4 shows the SAWA-H weights and similarity after including RelNormal.

Note that we do not claim that SAWA-H is a better metric, or a substitute for any existing metric. We only claim that SAWA-H aligns better with human judgment, which may or may not be desirable depending on the use case. Also, our approach is not restricted to human preferences. One can create composite metrics to align with arbitrary preferences based on the desired sensitivity to perturbations.

Example of SAWA-H on Real Predictions: In some cases, the SAWA-H (with RelNormal) metric provides different rankings among depth predictions than the AbsRel metric. In Fig. 6 we visualize one such difference for the predictions of Depth Pro and Metric3DV2 using error maps.Though Metric3D achieves a smaller AbsRel error, its SAWA-H error is substantially larger. This is because the Depth Pro model made a mistake judging the scale of the iceberg and orientation of the camera relative to ocean. Hence, the iceberg and ocean cannot both be properly aligned, so the global reconstruction is inaccurate and the AbsRel error is high. As seen in the projected contour plot, Depth Pro better approximates the local geometry and has a lower SAWA-H error score. Supplement F provides more examples. It is important to note that this example does *not* show that SAWA-H is a better metric than AbsRel because that depends on the application; it simply illustrates their different behaviors.

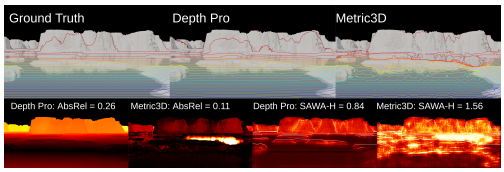

Figure 6: Disagreement between SAWA-H and AbsRel. Top row: projected contour plots of the ground truth and predicted depths. Bottom row: heatmaps for model predictions calculated with AbsRel and SAWA-H.

Performance of MDE Methods under SAWA-H: We use SAWA-H (with RelNormal) to evaluate 10 state-of-the-art MDE methods: DepthAnything(Yang et al., 2024a), DepthAnythingV2(Yang et al., 2024b), GeoWizard(Fu et al., 2024), Marigold(Ke et al., 2024), MiDaS(Birkl et al., 2023), DepthPro(Bochkovskiy et al., 2025), Metric3DV2(Hu et al., 2024), MoGe-2(Wang et al., 2025b), UniDepthV2(Piccinelli et al., 2026), and ZoeDepth(Bhat et al., 2023). Evaluation is conducted on 9 commonly used benchmarks: DDAD(Guizilini et al., 2020), DIODE(Vasiljevic et al., 2019), ETH3D(Schöps et al., 2019), HAMMER(Jung et al., 2023), iBims-1(Koch et al., 2020), KITTI(Uhrig et al., 2017), NYUv2(Silberman et al., 2012), Sintel(Butler et al., 2012), and Spring(Mehl et al., 2023).

Like some existing metrics, computing SAWA-H requires not just the depth map prediction but also the camera intrinsics. We use either ground truth or predicted (if applicable) camera intrinsics. If the method does not predict metric depth (e.g. predicts affine invariant depth/disparity), we first align prediction with ground truth depth using the conventional alignment strategy (e.g. perform affine alignment of disparity for affine invariant disparity prediction) and then un-project the aligned depth. Note that evaluation results that use ground truth depth for alignment are not directly comparable to those that do not. The same holds for using ground truth camera intrinsics for back-projection.

Fig. 5 presents metric values of each method averaged over 9 datasets. Performance of methods on each dataset are reported in Supplement H. Note that Boundary F1 is one component of SAWA-H but some datasets do not provide valid depth value on pixels near boundaries, preventing SAWA-H from properly penalizing blurry boundaries on those datasets. We additionally report performance of methods averaged on datasets with accurate depth across boundaries in Supplement H.

6 Limitations and Conclusions

One limitation is that our list of perturbations, although representative, may not be exhaustive. Our dataset for sensitivity measurement may have room for improvement in diversity and coverage. These may affect our numerical results. Still, most of our contributions are on methodology and tools, and we expect them to stay useful and increasingly so, as more perturbations and better data become available.

AI USE STATEMENT

In this work, we used generative AI tools for implementing the textureless relighting visualization tool, where all usage is asking LLMs about blender python APIs. We have reviewed all AI-assisted work by inspecting LLM output and verifying it through looking up the official blender python API documentation. We have not used generative AI tools for any other tasks, including but not limited to: generate synthetic data sets, help develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments, implement other parts of methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. Moreover, the following are not applicable to this work: formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide open access to our code in our openreview submission. Moreover, all code, data, and tools will be open-sourced. We describe the details of how each metric is computed in Supp. C and E, and describe the details of collecting human judgment in Sec. 4 and Supp. B.

REFERENCES

Ames room. https://en.wikipedia.org/wiki/Ames_room.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *CoRR*, abs/2302.12288, 2023. doi: 10.48550/ARXIV.2302.12288. URL https://doi.org/10.48550/arXiv.2302.12288.

Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 - A model zoo for robust monocular relative depth estimation. *CoRR*, abs/2307.14460, 2023. doi: 10.48550/ARXIV.2307.14460. URL https://doi.org/10.48550/arXiv.2307.14460.

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=aueXfY0Clv.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (eds.), *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI*, volume 7577 of *Lecture Notes in Computer Science*, pp. 611–625. Springer, 2012. doi: 10.1007/978-3-642-33783-3_44. URL https://doi.org/10.1007/978-3-642-33783-3_44.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. URL https://arxiv.org/abs/2001.10773.

Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Bilateral normal integration. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I*, volume 13661 of *Lecture Notes in Computer Science*, pp. 552–567. Springer, 2022. doi: 10.1007/978-3-031-19769-7_32. URL https://doi.org/10.1007/978-3-031-19769-7_32.

Chuangrong Chen, Xiaozhi Chen, and Hui Cheng. On the over-smoothing problem of CNN based disparity estimation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 8996–9004. IEEE, 2019. doi: 10.1109/ICCV.2019.00909. URL https://doi.org/10.1109/ICCV.2019.00909.

Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 730–738, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/0deb1c54814305ca9ad266f53bc82511-Abstract.html.

Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. OASIS: A large-scale dataset for single image 3d in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 676–685. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00076. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chen_OASIS_A_Large-Scale_Dataset_for_Single_Image_3D_in_the_CVPR_2020_paper.html.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2366–2374, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/7bccfde7714a1ebadf06c5f4cea752c1-Abstract.html.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXII*, volume 15080 of *Lecture Notes in Computer Science*, pp. 241–258. Springer, 2024. doi: 10.1007/978-3-031-72670-5_14. URL https://doi.org/10.1007/978-3-031-72670-5_14.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 2482–2491. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00256. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Guizilini_3D_Packing_for_Self-Supervised_Monocular_Depth_Estimation_CVPR_2020_paper.html.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10579–10596, 2024. doi: 10.1109/TPAMI.2024.3444912. URL https://doi.org/10.1109/TPAMI.2024.3444912.

HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, Ales Leonardis, Nassir Navab, and Benjamin Busam. On the importance of accurate geometry data for dense 3d vision tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 780–791. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00082. URL https://doi.org/10.1109/CVPR52729.2023.00082.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9492–9502. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00907. URL https://doi.org/10.1109/CVPR52733.2024.00907.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In Laura Leal-Taixé and Stefan Roth (eds.), *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part III*, volume 11131 of *Lecture Notes in Computer Science*, pp. 331–348. Springer, 2018. doi: 10.1007/978-3-030-11015-4_25. URL https://doi.org/10.1007/978-3-030-11015-4_25.

Tobias Koch, Lukas Liebel, Marco Körner, and Friedrich Fraundorfer. Comparison of monocular depth estimation methods using geometrically relevant metrics on the ibims-1 dataset. *Comput. Vis. Image Underst.*, 191:102877, 2020. doi: 10.1016/J.CVIU.2019.102877. URL https://doi.org/10.1016/j.cviu.2019.102877.

Zhenyu Li, Haotong Lin, Jiashi Feng, Peter Wonka, and Bingyi Kang. Benchdepth: Are we on the right way to evaluate depth foundation models? *CoRR*, abs/2507.15321, 2025. doi: 10.48550/ARXIV.2507.15321. URL https://doi.org/10.48550/arXiv.2507.15321.

Drew Linsley, Peisen Zhou, Alekh Karkada Ashok, Akash Nagaraj, Gaurav Gaonkar, Francis E. Lewis, Zygmunt Pizlo, and Thomas Serre. The 3d-pc: a benchmark for visual perspective taking in humans and machines. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=UIFAJZ22ZF.

Baoxia Liu and James T. Todd. Perceptual biases in the interpretation of 3d shape from shading. *Vision Research*, 44(18):2135–2145, 2004. doi: 10.1016/j.visres.2004.03.024. URL https://doi.org/10.1016/j.visres.2004.03.024.

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 4981–4991. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00482. URL https://doi.org/10.1109/CVPR52729.2023.00482.

Anton Obukhov, Matteo Poggi, Fabio Tosi, Ripudaman Singh Arora, Jaime Spencer, Chris Russell, Simon Hadfield, Richard Bowden, Shuaihang Wang, Zhenxin Ma, Weijie Chen, Baobei Xu, Fengyu Sun, Di Xie, Jiang Zhu, Mykola Lavreniuk, Haining Guan, Qun Wu, Yupei Zeng, Chao Lu, Huanran Wang, GuangYuan Zhou, Haotian Zhang, Jianxiong Wang, Qiang Rao, Chunjie Wang, Xiao Liu, Zhiqiang Lou, Hualie Jiang, Yihao Chen, Rui Xu, Minglang Tan, Zihan Qin, Yifan Mao, Jiayang Liu, Jialei Xu, Yifan Yang, Wenbo Zhao, Junjun Jiang, Xianming Liu, Mingshuai Zhao, Anlong Ming, Wu Chen, Feng Xue, Mengying Yu, Shida Gao, Xiangfeng Wang, Gbenga Omotara, Ramy Farag, Jacket Demby's, Seyed Mohamad Ali Tousi, Guilherme N. DeSouza, Tuan-Anh Yang, Minh-Quang Nguyen, Thien-Phuc Tran, Albert Luginov, and Muhammad Shahzad. The fourth monocular depth estimation challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 6182–6195. Computer Vision Foundation / IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/MDEC/html/Obukhov_The_Fourth_Monocular_Depth_Estimation_Challenge_CVPRW_2025_paper.html.

Evin Pinar Örnek, Shristi Mudgal, Johanna Wald, Yida Wang, Nassir Navab, and Federico Tombari. From 2d to 3d: Re-thinking benchmarking of monocular depth prediction. *CoRR*, abs/2203.08122, 2022. doi: 10.48550/ARXIV.2203.08122. URL https://doi.org/10.48550/arXiv.2203.08122.

Duc-Hai Pham, Tung Do, Phong Nguyen, Binh-Son Hua, Khoi Nguyen, and Rang Nguyen. Sharpdepth: Sharpening metric depth predictions using diffusion distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 17060–17069. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01590. URL https://openaccess.thecvf.com/content/CVPR2025/html/Pham_SharpDepth_Sharpening_Metric_Depth_Predictions_Using_Diffusion_Distillation_CVPR_2025_paper.html.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler, 2025. URL https://arxiv.org/abs/2502.20110.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segù, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(3):2354–2367, 2026. doi: 10.1109/TPAMI.2025.3628473. URL https://doi.org/10.1109/TPAMI.2025.3628473.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 12630–12641. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01215. URL https://doi.org/10.1109/CVPR52729.2023.01215.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 21783–21794. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02058. URL https://doi.org/10.1109/CVPR52733.2024.02058.

Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(5):824–840, 2009. doi: 10.1109/TPAMI.2008.132. URL https://doi.org/10.1109/TPAMI.2008.132.

Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: bundle adjusted direct RGB-D SLAM. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 134–144. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00022. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Schops_BAD_SLAM_Bundle_Adjusted_Direct_RGB-D_SLAM_CVPR_2019_paper.html.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (eds.), *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V*, volume 7576 of *Lecture Notes in Computer Science*, pp. 746–760. Springer, 2012. doi: 10.1007/978-3-642-33715-4_54. URL https://doi.org/10.1007/978-3-642-33715-4_54.

Lior Talker, Aviad Cohen, Erez Yosef, Alexandra Dana, and Michael Dinerstein. Mind the edge: Refining depth edges in sparsely-supervised monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 10606–10616. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01009. URL https://doi.org/10.1109/CVPR52733.2024.01009.

Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pp. 11–20. IEEE Computer Society, 2017. doi: 10.1109/3DV.2017.00012. URL https://doi.org/10.1109/3DV.2017.00012.

Igor Vasiljevic, Nicholas I. 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.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 5261–5271. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.00496. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_MoGe_Unlocking_Accurate_Monocular_Geometry_Estimation_for_Open-Domain_Images_with_CVPR_2025_paper.html.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/336572db3e99930814d6b328d4220cb6-Abstract-Conference.html.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 10371–10381. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.00987. URL https://doi.org/10.1109/CVPR52733.2024.00987.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html.

Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 204–213. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00027. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yin_Learning_To_Recover_3D_Scene_Shape_From_a_Single_Image_CVPR_2021_paper.html.

Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T. Freeman. Learning ordinal relationships for mid-level vision. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 388–396. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.52. URL https://doi.org/10.1109/ICCV.2015.52.

A Perturbation Algorithm Of Controllable Intensity

Here we provide details of algorithms to generate each type of perturbation.

Surface Orientation Perturbation. We first un-project ground truth depth to 3D, and compute the ground truth surface normal. Then we rotate every ground truth surface normal by the same amount. Perturbation intensity is defined as the magnitude of rotation (i.e. rotated by s degrees). Using the rotated surface normal, we compute the log depth gradient such that un-projection of neighboring pixels in 3D following the log depth gradient is perpendicular to the rotated surface normal. Then we solve the optimization problem of minimizing L2 distance between this target log depth gradient and the log depth gradient of perturbed depth, where the perturbed depth is treated as unknown variable. The optimization problem is solved by the conjugate gradient algorithm, with ground truth depth as initialization. We ignore constraints across occlusion boundaries to avoid them dominating the optimization objective, which leads to poor shape of perturbed depth.

Camera Intrinsics Perturbation. In this perturbation, focal length is perturbed by s times, $(s \geq 1)$, i.e. $f' = sf^*$, where f' and f^* are predicted and ground truth focal length. s controls perturbation intensity. Similar to surface orientation perturbation, we compute the ground truth surface normal. From the ground truth surface normals, we again compute the log depth gradient, but under the *perturbed* focal length. Then, we solve the optimization of minimizing L2 distance between this target log depth gradient computed from ground truth surface normal and the log depth gradient of perturbed depth similarly.

Relative Scale Perturbation. We partition pixels to 3 sets, S_{near}, S_{between}, and S_{far}. Depth of S_{near} remains unchanged. Depth of S_{far} is scaled by $s(s \geq 1)$, where s is the parameter to control perturbation magnitude. Depth of $(i, j) \in S_{\text{between}}$ is scaled by $1 + (s-1) \cdot \frac{D^*_{i,j} - d_l}{d_r - d_l}$, where $D^*_{i,j}$ denotes the ground truth depth, and d_l and d_r denotes the depth of farthest pixel in S_{near} and the nearest pixel in S_{far} in ground truth. To partition pixels, if the image can be split into two regions by occlusion boundaries, we set S_{near} to be the region closer to the camera, S_{far} to be the region farther to the camera, and S_{between} to be empty. For images that cannot be split into such two regions, we find $d_l, d_r (d_l < d_r)$ that minimizes d_r/d_l, such that there are 5% pixels with depth value in between, at least 30% pixels with depth value $< d_l$, and at least 30% pixels with depth value $> d_r$. $S_{\text{near}}, S_{\text{between}}, S_{\text{far}}$ are then set to be pixels with depth value $\leq d_l, \in [d_l, d_r]$, and $\geq d_r$.

Curvature Perturbation. In this perturbation, we first generate a HxW noise map $K \in \mathbb{R}^{H \times W}$, where $K_{i,j}$ are independently uniformly sampled from $[1-s, 1+s]$. Here $s(s \geq 0)$ controls perturbation magnitude. Then the perturbed depth, D', is generated by $D' \equiv D^* \otimes \text{clip}(\text{Gaussian Smooth}(K, \sigma), \min = 0.1)$, where D^* is ground truth depth, $\otimes$ denotes element-wise product, and σ controls perturbation frequency. We choose $\sigma = 1$ for high frequency perturbation, and $\sigma = 10$ for low frequency.

Affine Transform Perturbation. We study affine transform of depth and disparity. For affine transform of depth, perturbed depth D' follows $D' \equiv \frac{1}{s} D^* + \text{median}(D^*) - \text{median}(\frac{1}{s} D^*)$, where $s \geq 1$ controls perturbation magnitude. And for affine transform of disparity, perturbed depth D' follows $\frac{1}{D'} \equiv \frac{1}{s} \frac{1}{D^*} + \text{median}(\frac{1}{D^*}) - \text{median}(\frac{1}{s} \frac{1}{D^*})$, where $s \geq 1$ controls perturbation magnitude.

Boundary Perturbation. We apply mean filter of patch size $2s+1$, where $s \geq 0$ controls perturbation magnitude. For each pixel, the perturbed depth is further clipped by 0.7x and 1.3x of ground truth depth.

B More on Collecting Data of Human Judgment

Quality Control: Besides filtering annotation by accuracy on gold standard and the ratio of "not ground truth" annotation on ground truth depth maps (Sec. 4), we additionally control annotation quality by presenting annotators with multiple visualizations of a depth map. More specifically, we provide visualizations of viewing geometry generated from different angles (Fig. 3(c)), Textureless Relighting (Sec. 4.1), and Projected Contours (Sec. 4.1). Presenting multiple visualizations is important because it may be difficult to notice some artifacts of depth map from a single type of visualization, e.g. Fig. 3 Top.

To ensure that annotators examine all types of visualization, each time, we only show one image and one type of visualization of depth. A depth map is annotated as the ground truth depth if it is annotated as the ground truth depth under every type of visualization. Each visualization is examined independently and we do not reveal whether two visualizations are from the same depth map. We measure each annotator's accuracy on gold standard and the ratio of "not ground truth" annotation on ground truth depth maps using annotations on individual visualization.

Details and Statistics: We collect annotations from 6 annotators. They are all researchers with expertise in computer vision. We instruct them how visualizations are generated and show them 10 examples before they annotate. Each person annotates 49-107 depth maps (more specifically, 49, 77, 77, 77, 107, 77), with disjoint assignments except for gold standard, which are assigned to everyone. There is no time limit. For each gold standard depth map, we average all annotators' annotation after aggregating each annotator's annotation among different visualizations, and then use the averaged annotation when fitting quadratic curves.

Annotators' annotation accuracies on gold standard are 67%, 84%, 79%, 75%, 74%, and 84%. All annotations from the annotator whose accuracy on gold standard is 67% (who annotated 49 depth maps) are discarded, and all other annotations are retained. Std across different annotators' annotations on each gold standard depth map has a mean of 0.206. The ratio of "not ground truth" annotation on ground truth depth maps for each annotator is 0%, 0%, 0%, 0%, 17%, and 0%.

C Metric Definitions

Here we provide definitions for the metrics used in Fig. 4. These are all well-known or variants of well-known metrics.

For a ground truth depth map z and a predicted $\hat{z}$, AbsRel is defined as the average of $|z_i - \hat{z}_i|/z_i$. Following Wang et al. (2025a), we define AbsRel_p on point maps $p \in \mathbb{R}^{H\times W\times 3}$ by the average of $\|p_i - \hat{p}_i\| / \|p_i\|$.

δ^1 is defined as the fraction of pixels such that $\max\left\{d_i/\hat{d}_i, \hat{d}_i/d_i\right\} < 1.25$. We define a stricter metric $\delta^{0.125}$ by the fraction of pixels such that $\max\left\{d_i/\hat{d}_i, \hat{d}_i/d_i\right\} < 1.25^{0.125}$.

RMSE and RMSE (log) follow standard definitions. RMSE (log, scale invariant) is defined as

$$\text{Log RMSE SI} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log z_i - \log \hat{z}_i + \alpha)^2}$$

where $\alpha = \frac{1}{n}\sum_i(\log \hat{z}_i - \log z_i)$.

This follows the definitions in Ke et al. (2024); Eigen et al. (2014).

Finally, WKDR follows the definition given in Yin et al. (2021) based on relative depth relationships between pairs of pixels. Boundary F1 score follows the definition in Bochkovskiy et al. (2025) to compute an F1 score over the edges in the predicted and ground truth depth maps.

The affine and scale alignment procedure used throughout is identical to MoGe (Wang et al. (2025a)). Notably, affine alignment for the point cloud metric AbsRel_p is not the same as affine alignment of the depth. "Lst. Sq." denotes least squares alignment on depth, as performed by Marigold (Ke et al. (2024)).

D Causes of Different Behaviors Between Human Judgment and Existing Metrics

Fig. 7 shows instances of relative scale, surface orientation, camera intrinsics, and curvature perturbations. On each scene, instance of curvature perturbation has much lower (better) AbsRel score, but its shape is very different from ground truth.

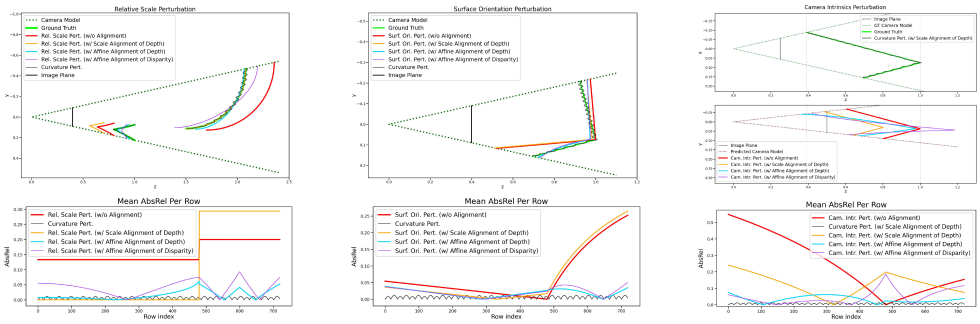

Figure 7: Top and Mid: side views of ground truth and different perturbations. Relative scale, surface orientation, and camera intrinsics perturbations (red) have good shape, but comparing with curvature perturbations (black), they have poor AbsRel, even under alignments (orange, blue, and purple). Bottom: mean AbsRel of each row. Curvature perturbation has much lower AbsRel.

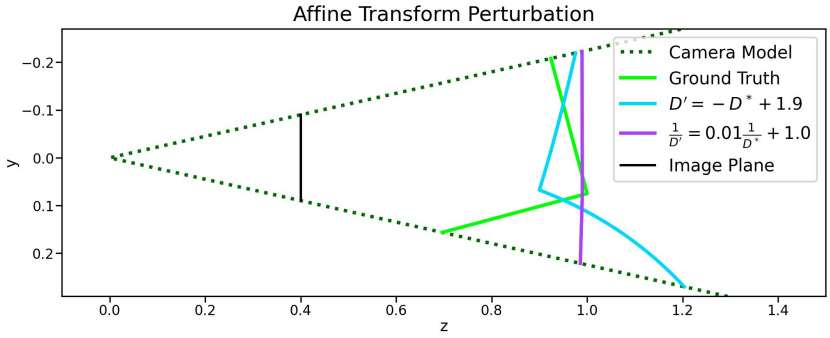

Figure 8: Side view of ground truth (green) and affine transform perturbations. Extreme affine transform parameters flatten geometry (purple) or change relative depth (blue).

Fig. 8 shows instances of affine transform perturbations. The shape of these instances is distorted a lot, but they achieve a perfect score when the respective type of affine alignment is applied during evaluation.

E Details of RelNormal Metric

E.1 Patch Selection

Our selection of patches is motivated by the following procedure: first uniformly select a pixel I within the image, then select a pixel J uniformly within a neighborhood of I, finally compute the error. To approximate this expectation over a multivariate uniform distribution, we use the first m elements of the Sobol sequence. This provides a deterministic algorithm for computing the metric. In Supplement E.3 we show that 1 million samples is sufficient when computing this metric on standard RGB-D datasets. We take the average over multiple scales by downsampling the image by a factor k and computing the relative normals. All results are computed using a neighborhood radius of 32 and scale factors $[1, 2, 4, 8]$.

E.2 Normal Calculation

We compute normals of surface described by each patch using the cross product of vectors from the bottom and right pixels to the central pixel in the patch. For datasets with noisy ground truth, it may be preferable to compute normals with a plane of best fit.

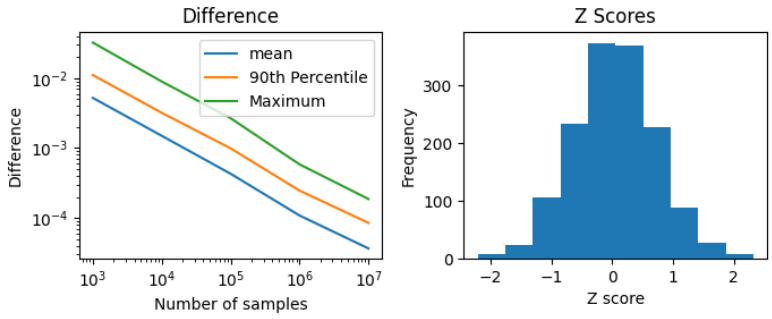

Figure 9: Comparison between deterministic and random sampling algorithm. The left plot shows the difference between the deterministic algorithm with n samples and the random algorithm with 10^8 samples. The maximum difference with 1M samples is 5.84×10^{-4} while the mean difference is 1.08×10^{-4}. The right plot shows the z scores of the deterministic algorithm with 1M samples, compared to a distribution of the randomized algorithm with 1M samples computed with 30 different seeds.

E.3 Patch Sampling Procedure

When computing the relative normal metric, we wish to ensure that our deterministic sampling technique accurately approximates the true mean of the distribution. In Fig. 9 we display the difference between the metric computed with the deterministic algorithm and the metric computed with 10^8 random samples. We perform the error computations using predictions from Depth Pro (Bochkovskiy et al. (2025)), UniDepthV2 (Piccinelli et al. (2025)), MoGeV1 (Wang et al. (2025a)), Metric3DV2 (Hu et al. (2024)), and MoGeV2 (Wang et al. (2025b)) on 100 images from the iBims dataset and 100 images from the Virtual KITTI dataset Koch et al. (2020); Cabon et al. (2020). The maximum error 5.84×10^{-4} and average error of 1.08×10^{-4} is sufficiently small to justify using 10^6 samples. We also analyze the z-score of the deterministic computation compared to random sampling. This suggests that using the Sobol sequence does not introduce unexpected irregularities.

F SAWA-H Comparison Examples

Fig. 10 shows more qualitative examples of predictions evaluated by SAWA-H (with RelNormal) and AbsRel.

G Dataset Gallery

Fig. 11 shows images of 20 scenes randomly selected from the dataset.

Fig. 12,13,14,15,16,17,18,19 show examples of every perturbation type. To make it easier to see how the geometry is perturbed, we present visualization of ground truth depth next to visualization of perturbed depth.

H Performance of MDE Methods under SAWA-H

Fig. 20 displays performance of state-of-the-art monocular depth estimation methods evaluated by SAWA-H on each dataset. Fig. 21 displays performance of state-of-the-art monocular depth estimation methods evaluated by SAWA-H averaged on datasets with accurate depth values near boundaries.

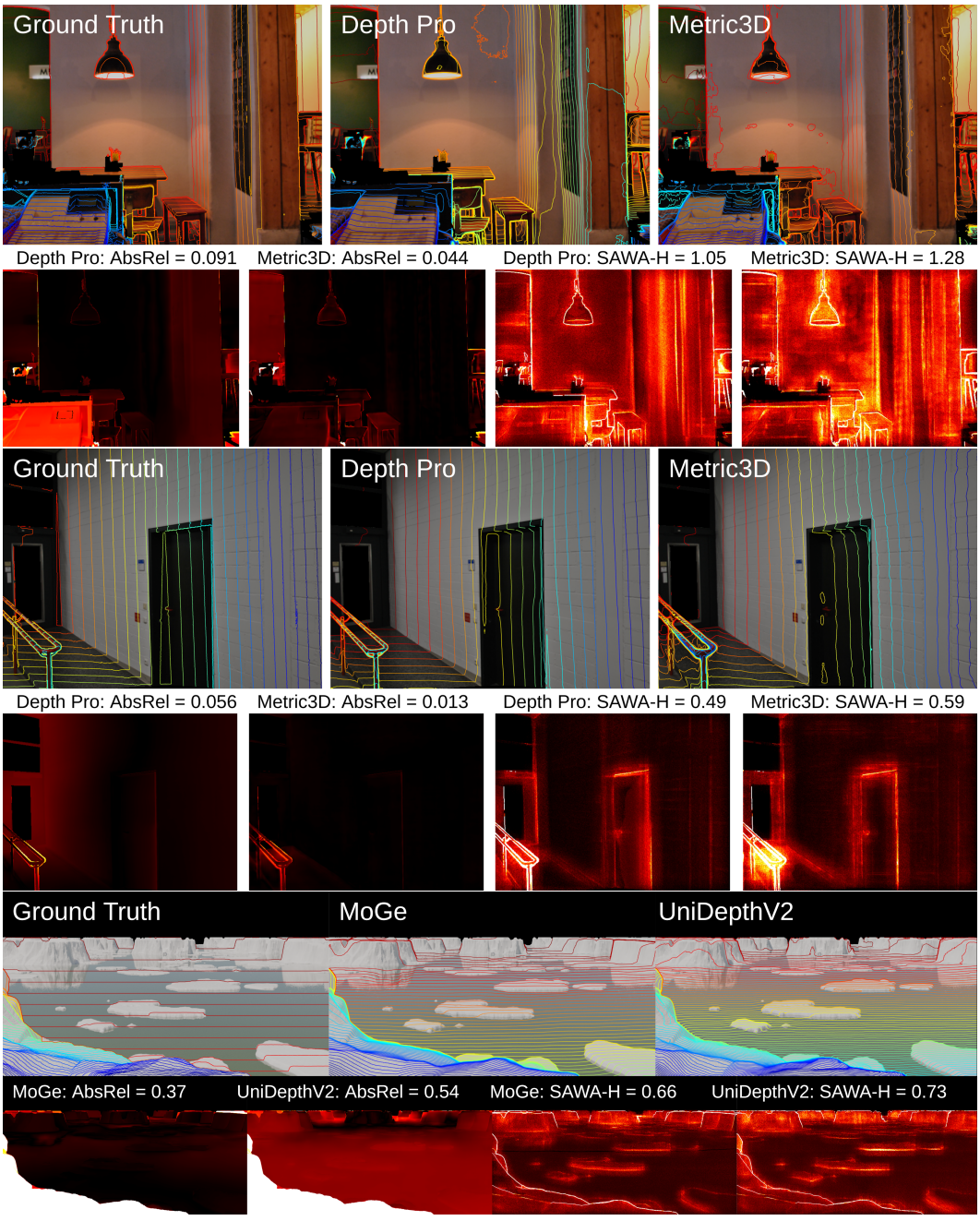

Figure 10: Additional comparisons between SAWA-H and AbsRel. The top row displays contour plots of the ground truth and predicted depths. The bottom row displays heatmaps for model predictions calculated with AbsRel and SAWA-H. AbsRel is computed by aligning the scale with the ground truth.



Figure 11: Sample of 20 Images from the Sensitivity Dataset

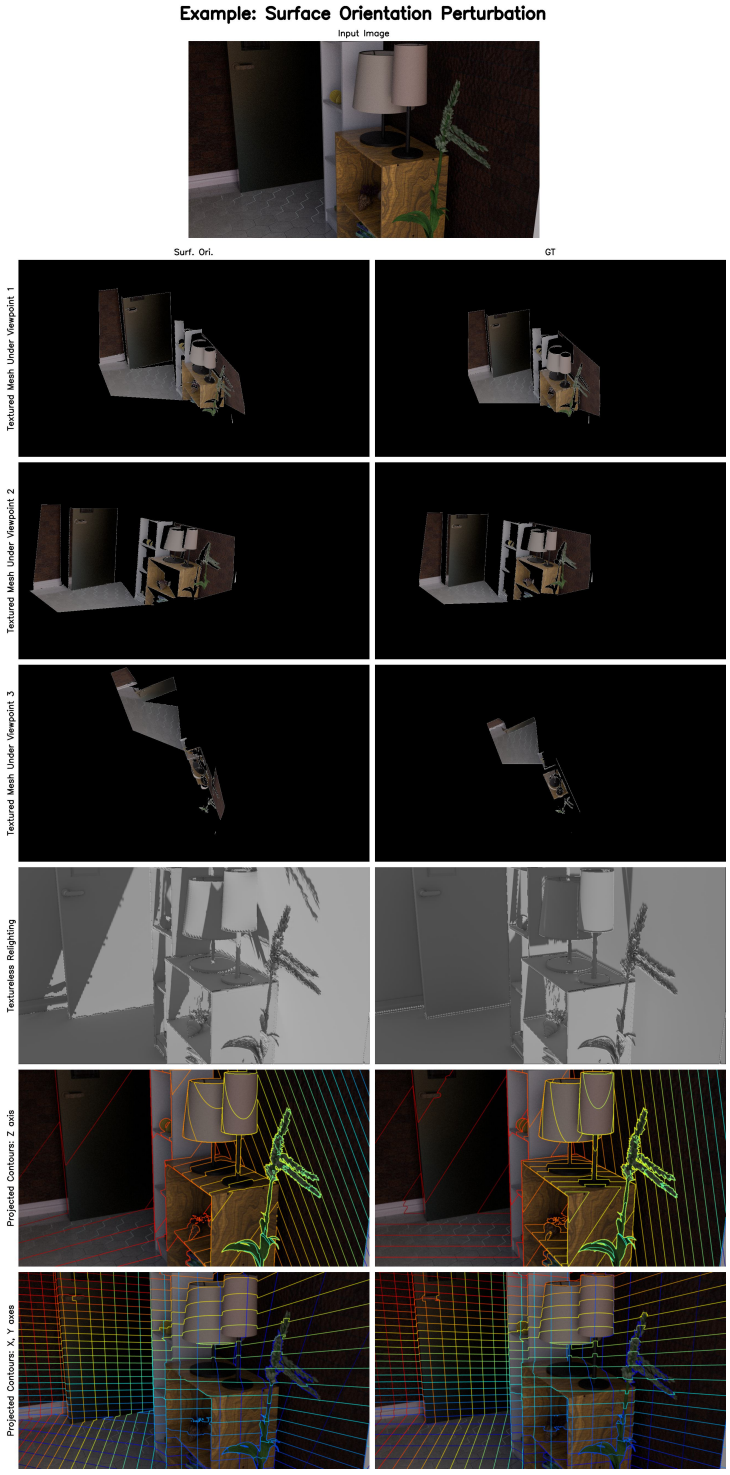

Figure 12: Example of Surface Orientation Perturbation. In visualization of textured mesh, images on the same row are captured under the same camera pose and using ground truth camera intrinsics. The geometry is perturbed by rotating 15° along the axis of $(0.25, 0.51, -0.82)$. The orientation difference is noticeable in Projected Contours: Z axis.

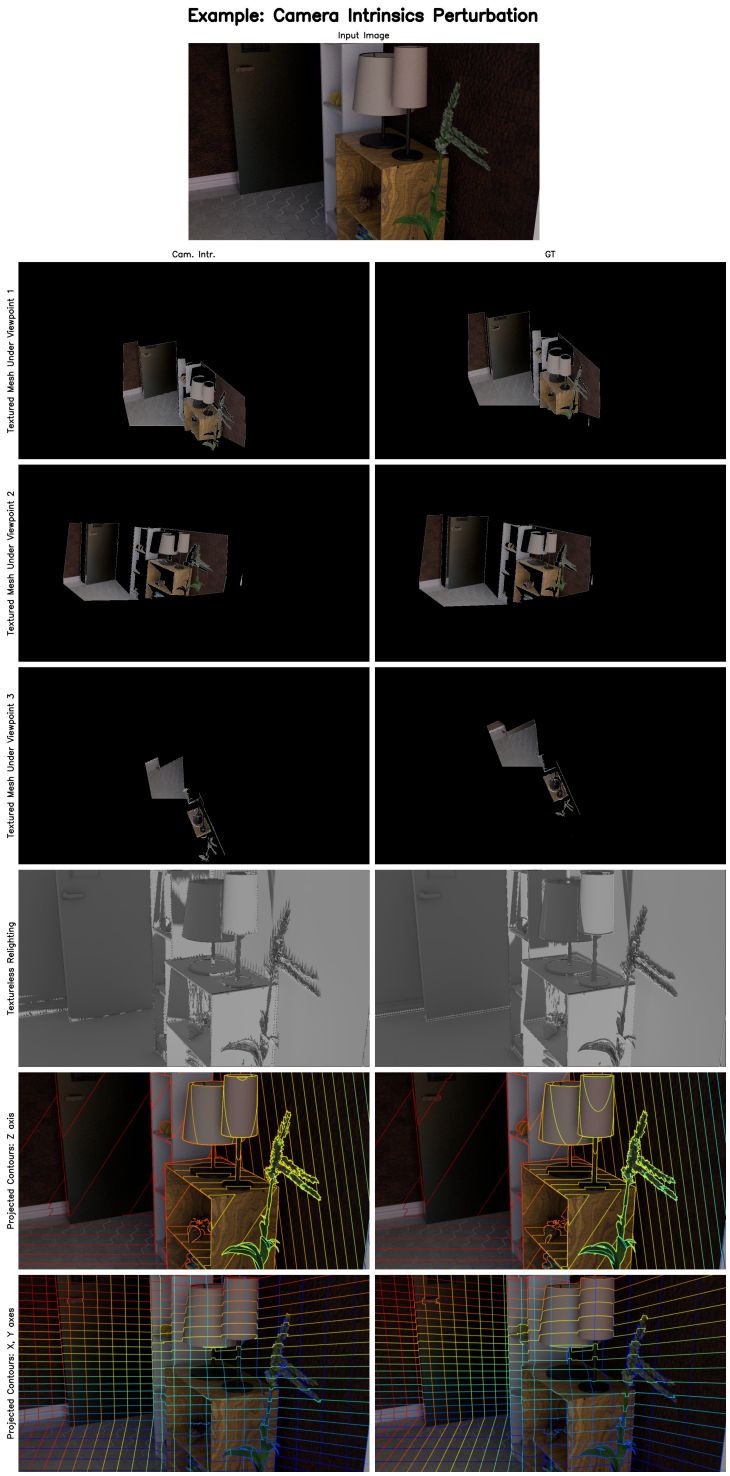

Figure 13: Example of Camera Intrinsics Perturbation. In visualization of textured mesh, images on the same row are captured under the same camera pose. Perturbed camera intrinsics are used to visualize perturbed geometry, and ground truth camera intrinsics are used for ground truth geometry. Focal length is perturbed to be 2x of ground truth focal length. To maintain similar geometry, depth range increases. This can be noticed in the textured mesh under different viewpoints.

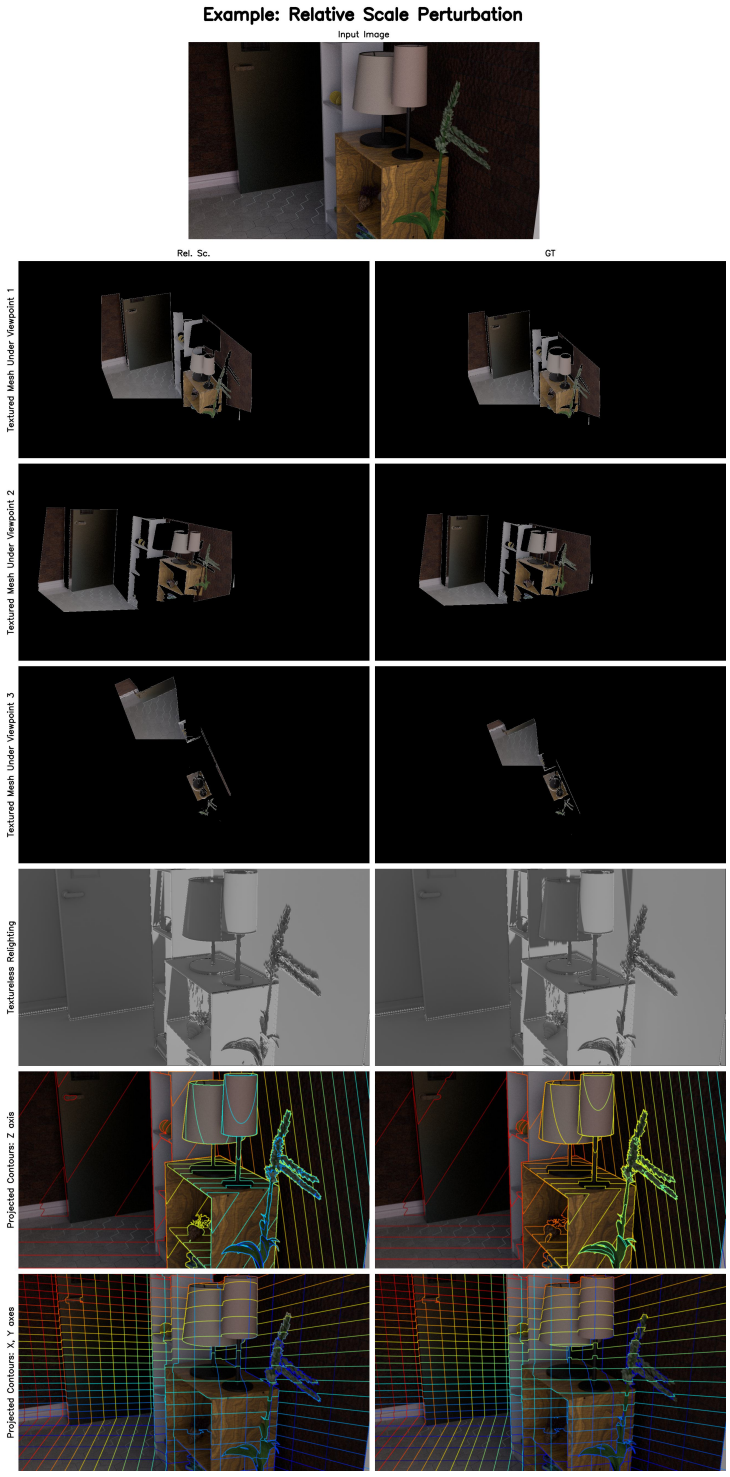

Figure 14: Example of Relative Scale Perturbation. In visualization of the textured mesh, images on the same row are captured under the same camera pose using ground truth camera intrinsics. Relative scale between the cabinet and the wall is perturbed.

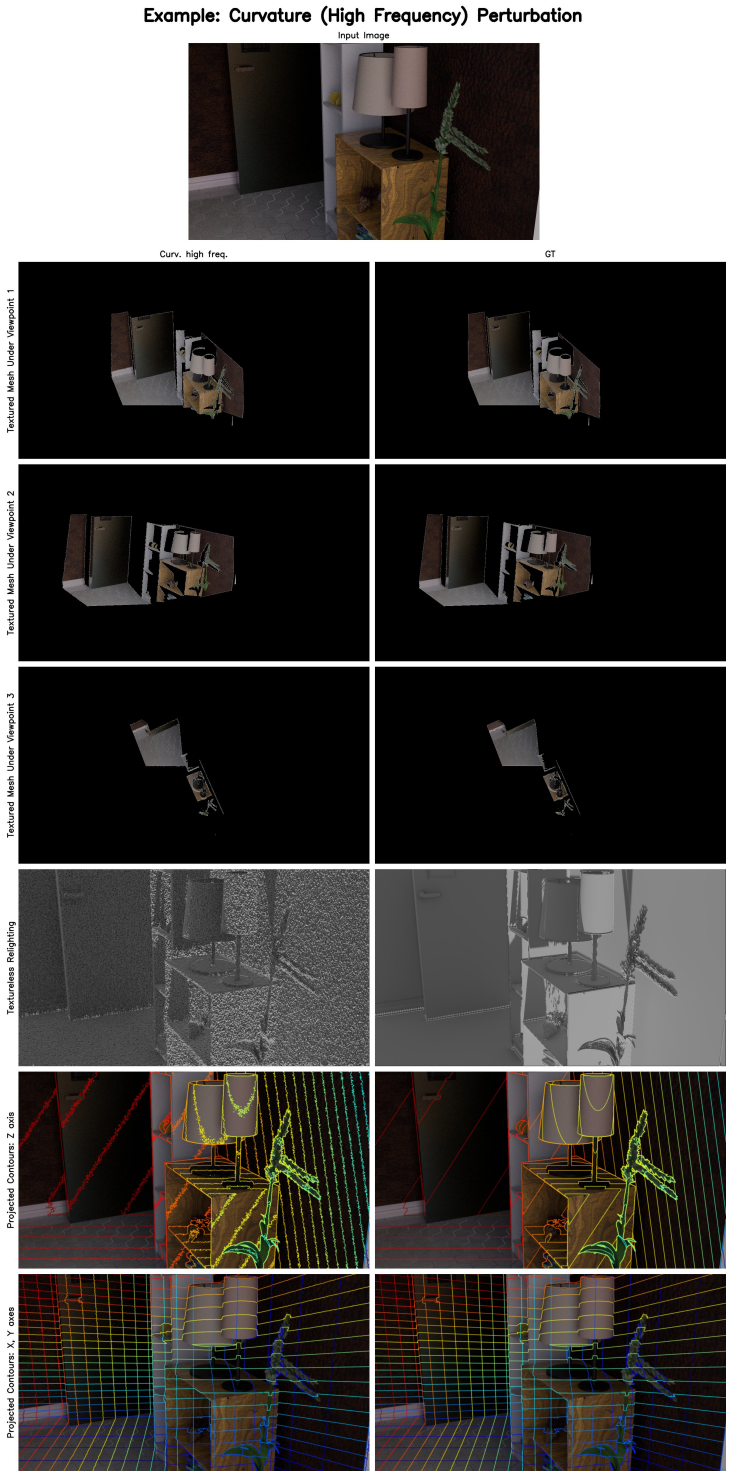

Figure 15: Example of Curvature (High Frequency) Perturbation. In visualization of the textured mesh, images on the same row are captured under the same camera pose using ground truth camera intrinsics. There are many high frequency spikes in the perturbed geometry, which can be easily observed in Textureless Relighting and Projected Contours.

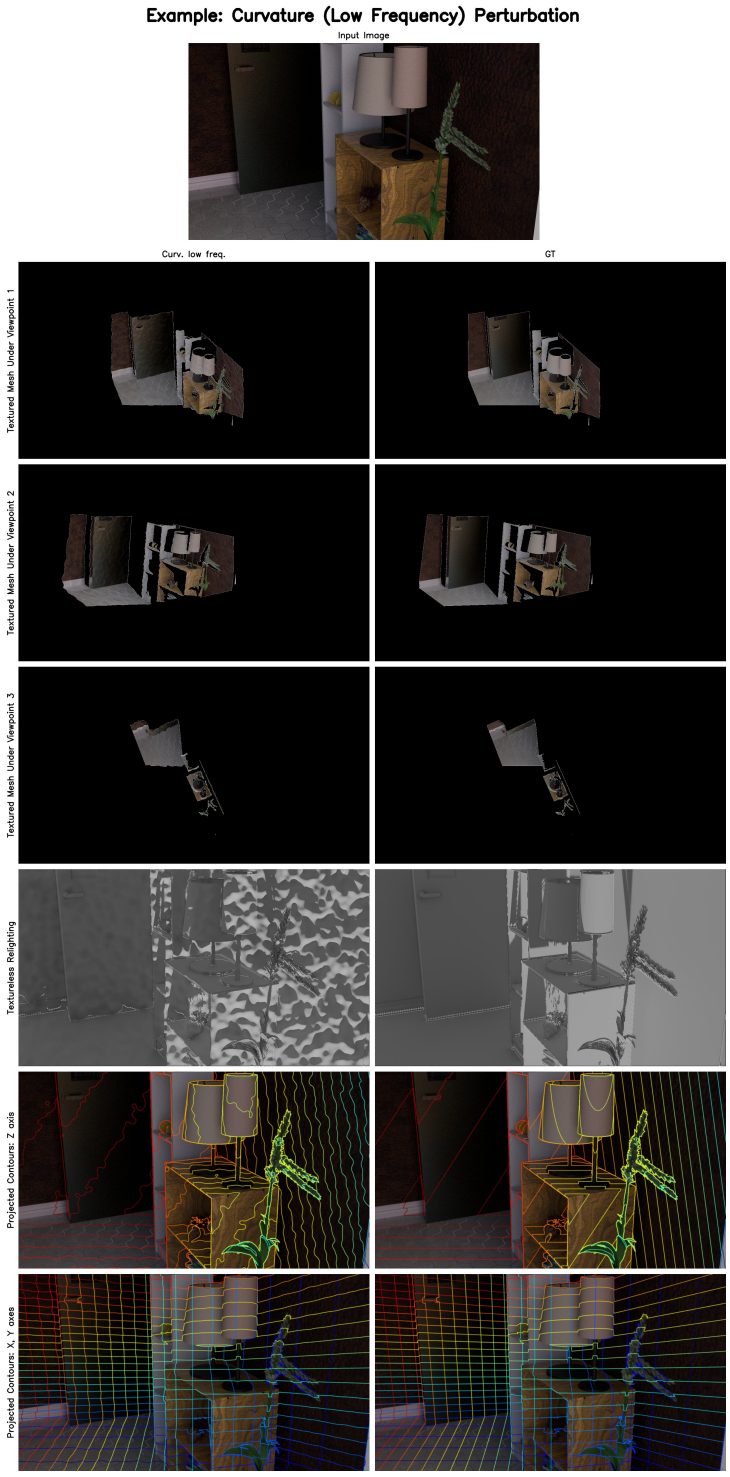

Figure 16: Example of Curvature (Low Frequency) Perturbation. In visualization of textured mesh, images on the same row are captured under the same camera pose and using ground truth camera intrinsics. There are many bumps in the perturbed geometry, which can be easily observed in Textureless Relighting and Projected Contours.

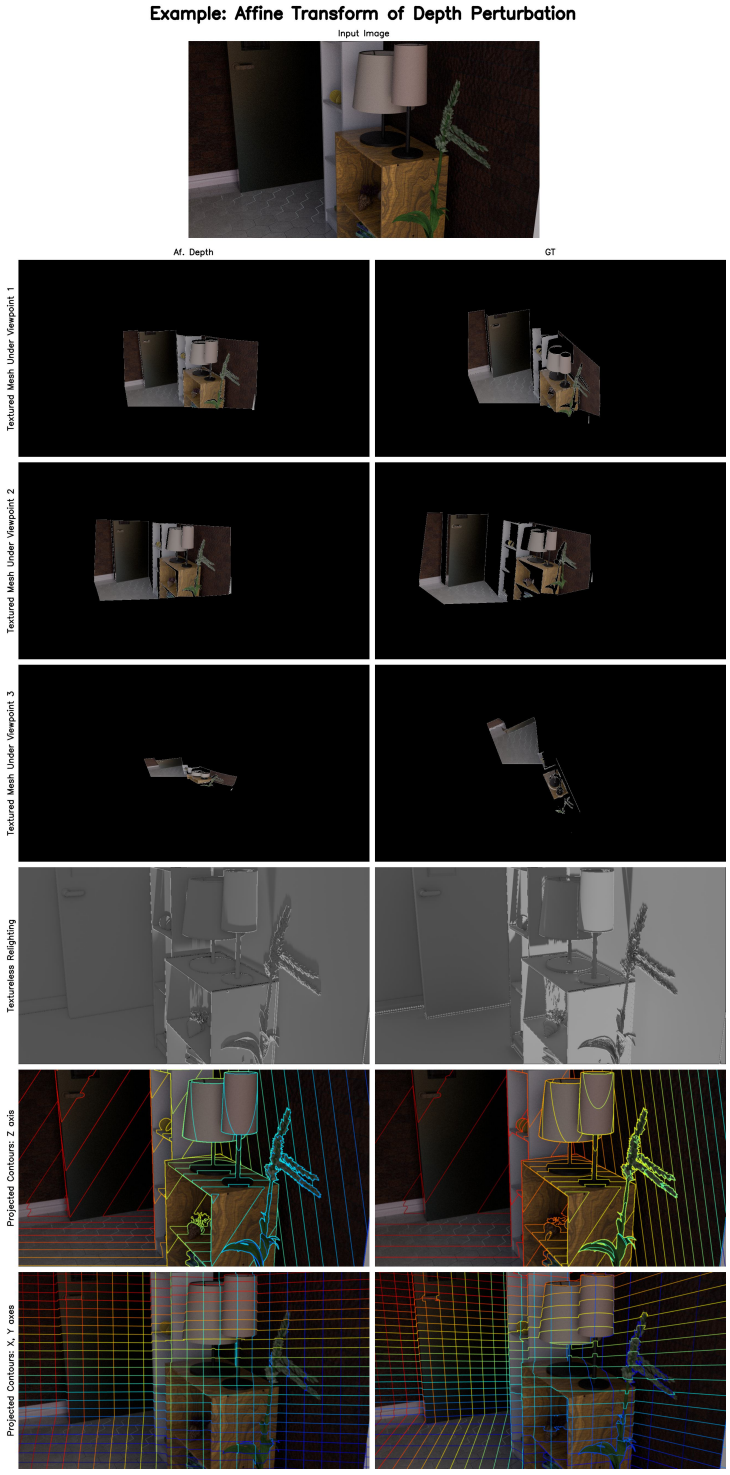

Figure 17: Example of Affine Transform of Depth Perturbation. In visualization of the textured mesh, images on the same row are captured under the same camera pose and using ground truth camera intrinsics. Depth is scaled by 0.2x before being translated to retain the same median depth, so the perturbed geometry is flattened.

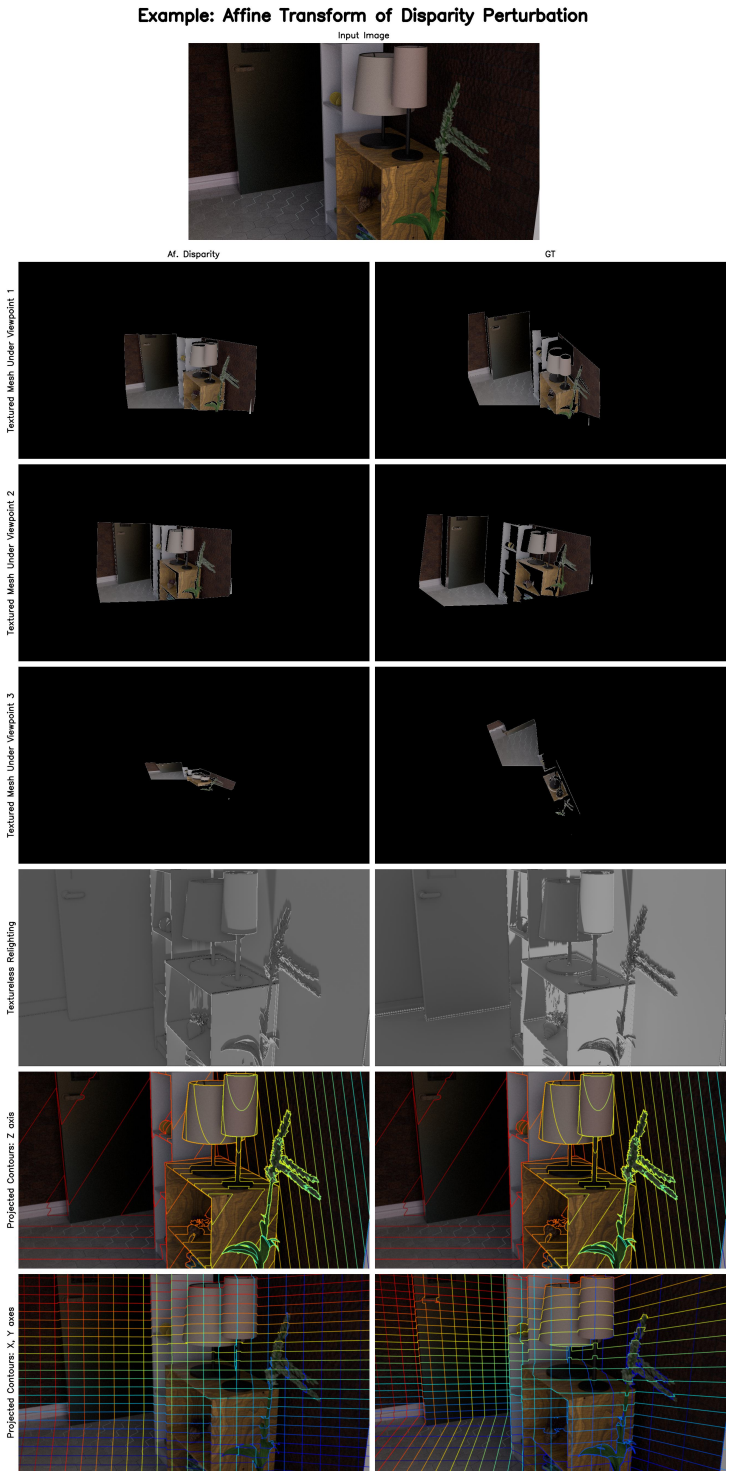

Figure 18: Example of Affine Transform of Disparity Perturbation. In visualization of the textured mesh, images on the same row are captured under the same camera pose and using ground truth camera intrinsics. Disparity is scaled by 0.2x before being translated to retain the same median disparity, so the perturbed geometry is flattened.

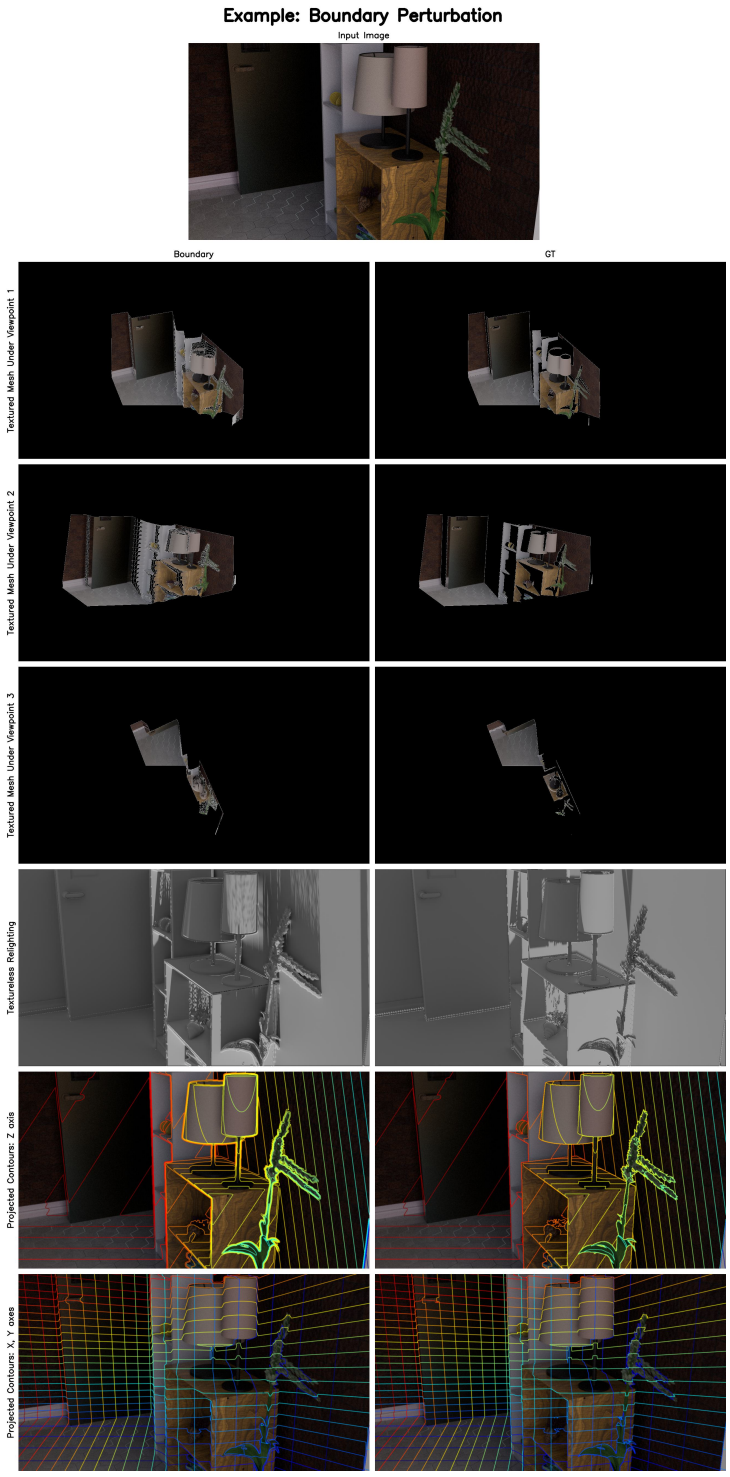

Figure 19: Example of Boundary Perturbation. In visualization of the textured mesh, images on the same row are captured under the same camera pose and using ground truth camera intrinsics. Depth values are blurred along boundaries, which can be easily observed in the textured mesh under different viewpoints and the thicker contour lines in Projected Contours.

DDAD

		Align w/ GT Depth, Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ Pred. Cam. Intr.		
		DAv1↑	DAv2↑	GeoWizard↑	Marigold↑	MiDaS↑	DepthPro↑	Metric3DV2↑	MoGe-2↑	UniDepthV2↑	ZoeDepth↑	DepthPro	MoGe-2	UniDepthV2
Components of SAWA-H	WKDR-No Align. ↑ (3.65x)	0.94	0.94	0.91	0.91	0.92	0.94	0.96	0.95	0.95	0.92	0.96	0.95	0.95
	$\delta^{0.125}$-Disparity Af. ↑ (0.18x)	0.26	0.29	0.14	0.09	0.24	0.29	0.43	0.37	0.39	0.28	0.29	0.37	0.39
	$\delta^{0.125}$-Depth Af. ↑ (0.01x)	0.07	0.08	0.12	0.10	0.08	0.15	0.29	0.07	0.21	0.10	0.15	0.08	0.21
	Boundary F1-No Align. ↑ (0.20x)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RelNormal ↓ (1.94x)	0.63	0.55	0.63	0.59	0.62	0.53	0.53	0.53	0.54	0.59	0.53	0.53	0.54
New Metric: SAWA-H ↓		1.80	1.62	1.90	1.83	1.84	1.59	1.49	1.52	1.55	1.78	1.59	1.52	1.55

DIODE

		Align w/ GT Depth, Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ Pred. Cam. Intr.		
		DAv1↑	DAv2↑	GeoWizard↑	Marigold↑	MiDaS↑	DepthPro↑	Metric3DV2↑	MoGe-2↑	UniDepthV2↑	ZoeDepth↑	DepthPro	MoGe-2	UniDepthV2
Components of SAWA-H	WKDR-No Align. ↑ (3.65x)	0.95	0.95	0.93	0.92	0.93	0.94	0.95	0.97	0.96	0.89	0.94	0.97	0.96
	$\delta^{0.125}$-Disparity Af. ↑ (0.18x)	0.56	0.54	0.44	0.42	0.49	0.59	0.67	0.67	0.67	0.39	0.59	0.67	0.67
	$\delta^{0.125}$-Depth Af. ↑ (0.01x)	0.47	0.43	0.43	0.41	0.41	0.50	0.65	0.50	0.57	0.32	0.50	0.50	0.57
	Boundary F1-No Align. ↑ (0.20x)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RelNormal ↓ (1.94x)	0.20	0.17	0.44	0.41	0.21	0.23	0.48	0.16	0.25	0.33	0.23	0.16	0.23
New Metric: SAWA-H ↓		0.86	0.81	1.41	1.39	0.96	0.93	1.38	0.71	0.88	1.34	0.93	0.70	0.84

ETH3D

		Align w/ GT Depth, Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ Pred. Cam. Intr.		
		DAv1↑	DAv2↑	GeoWizard↑	Marigold↑	MiDaS↑	DepthPro↑	Metric3DV2↑	MoGe-2↑	UniDepthV2↑	ZoeDepth↑	DepthPro	MoGe-2	UniDepthV2
Components of SAWA-H	WKDR-No Align. ↑ (3.65x)	0.95	0.95	0.93	0.93	0.93	0.94	0.96	0.97	0.97	0.91	0.98	0.97	0.97
	$\delta^{0.125}$-Disparity Af. ↑ (0.18x)	0.51	0.52	0.39	0.40	0.38	0.54	0.58	0.68	0.71	0.37	0.54	0.66	0.71
	$\delta^{0.125}$-Depth Af. ↑ (0.01x)	0.41	0.40	0.39	0.41	0.29	0.46	0.56	0.48	0.66	0.29	0.46	0.48	0.66
	Boundary F1-No Align. ↑ (0.20x)	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.00
	RelNormal ↓ (1.94x)	0.22	0.20	0.40	0.34	0.22	0.20	0.31	0.18	0.21	0.27	0.21	0.18	0.21
New Metric: SAWA-H ↓		0.90	0.85	1.33	1.22	1.02	0.89	1.04	0.74	0.78	1.19	0.89	0.74	0.78

HAMMER

		Align w/ GT Depth, Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ Pred. Cam. Intr.		
		DAv1↑	DAv2↑	GeoWizard↑	Marigold↑	MiDaS↑	DepthPro↑	Metric3DV2↑	MoGe-2↑	UniDepthV2↑	ZoeDepth↑	DepthPro	MoGe-2	UniDepthV2
Components of SAWA-H	WKDR-No Align. ↑ (3.65x)	0.91	0.92	0.93	0.93	0.88	0.94	0.95	0.96	0.96	0.89	0.94	0.96	0.96
	$\delta^{0.125}$-Disparity Af. ↑ (0.18x)	0.37	0.37	0.50	0.46	0.27	0.55	0.60	0.68	0.72	0.32	0.55	0.68	0.72
	$\delta^{0.125}$-Depth Af. ↑ (0.01x)	0.37	0.37	0.51	0.46	0.28	0.53	0.61	0.65	0.70	0.32	0.53	0.65	0.70
	Boundary F1-No Align. ↑ (0.20x)	0.01	0.05	0.02	0.03	0.02	0.04	0.01	0.04	0.03	0.00	0.04	0.04	0.03
	RelNormal ↓ (1.94x)	0.23	0.21	0.30	0.26	0.23	0.21	0.28	0.21	0.22	0.26	0.21	0.21	0.22
New Metric: SAWA-H ↓		1.10	1.02	1.12	1.06	1.23	0.89	1.00	0.82	0.81	1.25	0.89	0.82	0.81

iBims-1

		Align w/ GT Depth, Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ Pred. Cam. Intr.		
		DAv1↑	DAv2↑	GeoWizard↑	Marigold↑	MiDaS↑	DepthPro↑	Metric3DV2↑	MoGe-2↑	UniDepthV2↑	ZoeDepth↑	DepthPro	MoGe-2	UniDepthV2
Components of SAWA-H	WKDR-No Align. ↑ (3.65x)	0.95	0.96	0.94	0.95	0.94	0.96	0.95	0.97	0.96	0.91	0.96	0.97	0.96
	$\delta^{0.125}$-Disparity Af. ↑ (0.18x)	0.59	0.63	0.51	0.54	0.53	0.65	0.81	0.74	0.72	0.40	0.65	0.74	0.72
	$\delta^{0.125}$-Depth Af. ↑ (0.01x)	0.56	0.61	0.54	0.56	0.51	0.65	0.64	0.75	0.73	0.38	0.65	0.75	0.73
	Boundary F1-No Align. ↑ (0.20x)	0.09	0.17	0.12	0.15	0.12	0.18	0.13	0.20	0.10	0.03	0.18	0.20	0.10
	RelNormal ↓ (1.94x)	0.28	0.23	0.31	0.30	0.28	0.23	0.34	0.21	0.25	0.35	0.23	0.21	0.25
New Metric: SAWA-H ↓		0.98	0.83	1.07	1.05	1.02	0.83	1.07	0.74	0.85	1.33	0.83	0.74	0.85

KITTI

		Align w/ GT Depth, Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ Pred. Cam. Intr.		
		DAv1↑	DAv2↑	GeoWizard↑	Marigold↑	MiDaS↑	DepthPro↑	Metric3DV2↑	MoGe-2↑	UniDepthV2↑	ZoeDepth↑	DepthPro	MoGe-2	UniDepthV2
Components of SAWA-H	WKDR-No Align. ↑ (3.65x)	0.95	0.95	0.93	0.93	0.92	0.96	0.96	0.97	0.97	0.96	0.96	0.97	0.97
	$\delta^{0.125}$-Disparity Af. ↑ (0.18x)	0.30	0.31	0.23	0.20	0.21	0.41	0.38	0.53	0.55	0.44	0.41	0.53	0.55
	$\delta^{0.125}$-Depth Af. ↑ (0.01x)	0.22	0.22	0.19	0.18	0.14	0.29	0.34	0.38	0.43	0.31	0.29	0.38	0.43
	Boundary F1-No Align. ↑ (0.20x)	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.02	0.01	0.00	0.01	0.02	0.01
	RelNormal ↓ (1.94x)	0.50	0.47	0.55	0.51	0.52	0.46	0.47	0.46	0.47	0.48	0.46	0.46	0.47
New Metric: SAWA-H ↓		1.49	1.42	1.66	1.58	1.65	1.35	1.35	1.29	1.30	1.38	1.35	1.29	1.30

NYUv2

		Align w/ GT Depth, Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ Pred. Cam. Intr.		
		DAv1↑	DAv2↑	GeoWizard↑	Marigold↑	MiDaS↑	DepthPro↑	Metric3DV2↑	MoGe-2↑	UniDepthV2↑	ZoeDepth↑	DepthPro	MoGe-2	UniDepthV2
Components of SAWA-H	WKDR-No Align. ↑ (3.65x)	0.94	0.94	0.93	0.93	0.94	0.94	0.95	0.95	0.95	0.93	0.94	0.95	0.95
	$\delta^{0.125}$-Disparity Af. ↑ (0.18x)	0.50	0.51	0.42	0.42	0.45	0.52	0.50	0.61	0.63	0.45	0.52	0.63	0.63
	$\delta^{0.125}$-Depth Af. ↑ (0.01x)	0.51	0.50	0.45	0.44	0.46	0.54	0.42	0.63	0.63	0.44	0.54	0.63	0.63
	Boundary F1-No Align. ↑ (0.20x)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RelNormal ↓ (1.94x)	0.42	0.42	0.40	0.41	0.42	0.41	0.39	0.41	0.40	0.41	0.41	0.41	0.40
New Metric: SAWA-H ↓		1.31	1.31	1.34	1.35	1.35	1.29	1.25	1.23	1.21	1.36	1.29	1.24	1.21

Sintel

		Align w/ GT Depth, Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ Pred. Cam. Intr.		
		DAv1↑	DAv2↑	GeoWizard↑	Marigold↑	MiDaS↑	DepthPro↑	Metric3DV2↑	MoGe-2↑	UniDepthV2↑	ZoeDepth↑	DepthPro	MoGe-2	UniDepthV2
Components of SAWA-H	WKDR-No Align. ↑ (3.65x)	0.91	0.90	0.87	0.84	0.86	0.91	0.91	0.92	0.92	0.85	0.91	0.92	0.92
	$\delta^{0.125}$-Disparity Af. ↑ (0.18x)	0.18	0.16	0.14	0.12	0.15	0.24	0.21	0.23	0.23	0.13	0.24	0.23	0.23
	$\delta^{0.125}$-Depth Af. ↑ (0.01x)	0.13	0.12	0.14	0.12	0.10	0.20	0.17	0.17	0.20	0.12	0.20	0.17	0.20
	Boundary F1-No Align. ↑ (0.20x)	0.25	0.32	0.05	0.06	0.18	0.39	0.26	0.33	0.35	0.02	0.39	0.33	0.35
	RelNormal ↓ (1.94x)	0.42	0.38	0.64	0.63	0.46	0.37	0.56	0.33	0.39	0.53	0.39	0.33	0.37
New Metric: SAWA-H ↓		1.44	1.39	2.09	2.18	1.73	1.31	1.73	1.22	1.31	1.93	1.34	1.21	1.28

Spring

		Align w/ GT Depth, Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ Pred. Cam. Intr.		
		DAv1↑	DAv2↑	GeoWizard↑	Marigold↑	MiDaS↑	DepthPro↑	Metric3DV2↑	MoGe-2↑	UniDepthV2↑	ZoeDepth↑	DepthPro	MoGe-2	UniDepthV2
Components of SAWA-H	WKDR-No Align. ↑ (3.65x)	0.92	0.91	0.88	0.86	0.89	0.92	0.90	0.92	0.92	0.86	0.92	0.92	0.92
	$\delta^{0.125}$-Disparity Af. ↑ (0.18x)	0.19	0.16	0.10	0.10	0.15	0.16	0.16	0.18	0.19	0.13	0.16	0.18	0.19
	$\delta^{0.125}$-Depth Af. ↑ (0.01x)	0.14	0.13	0.10	0.10	0.11	0.12	0.15	0.13	0.14	0.09	0.12	0.13	0.14
	Boundary F1-No Align. ↑ (0.20x)	0.07	0.09	0.05	0.05	0.06	0.11	0.08	0.09	0.07	0.01	0.11	0.09	0.07
	RelNormal ↓ (1.94x)	0.62	0.58	0.76	0.73	0.66	0.56	0.72	0.55	0.59	0.65	0.57	0.54	0.57
New Metric: SAWA-H ↓		1.83	1.79	2.27	2.28	2.04	1.90	2.12	1.70	1.79	2.14	1.91	1.67	1.76

Figure 20: Performance of state-of-the-art monocular depth estimation methods under SAWA-H on each dataset.

Average Over HAMMER, iBims-1, Sintel, Spring

		Align w/ GT Depth, Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ GT Cam. Intr.					No Align., Un-proj. w/ Pred. Cam. Intr.		
		DAv1↑	DAv2↑	GeoWizard↑	Marigold↑	MiDaS↑	DepthPro↑	Metric3DV2↑	MoGe-2↑	UniDepthV2↑	ZoeDepth↑	DepthPro	MoGe-2	UniDepthV2
Components of SAWA-H	WKDR-No Align. ↑ (3.65x)	0.92	0.92	0.91	0.89	0.89	0.92	0.93	0.94	0.94	0.88	0.92	0.94	0.94
	$\delta^{0.125}$-Disparity Af. ↑ (0.18x)	0.33	0.33	0.31	0.30	0.28	0.40	0.40	0.46	0.46	0.25	0.40	0.46	0.46
	$\delta^{0.125}$-Depth Af. ↑ (0.01x)	0.30	0.31	0.32	0.31	0.25	0.37	0.39	0.42	0.44	0.23	0.37	0.42	0.44
	Boundary F1-No Align. ↑ (0.20x)	0.11	0.16	0.06	0.07	0.10	0.18	0.12	0.16	0.14	0.01	0.18	0.16	0.14
	RelNormal ↓ (1.94x)	0.39	0.35	0.50	0.48	0.41	0.34	0.47	0.33	0.36	0.45	0.35	0.32	0.35
New Metric: SAWA-H ↓		1.34	1.26	1.64	1.64	1.50	1.23	1.48	1.12	1.19	1.66	1.24	1.11	1.17

Figure 21: Performance of state-of-the-art monocular depth estimation methods under SAWA-H, averaged on datasets with accurate depth values near boundaries.