Pano3D: A Holistic Benchmark and a Solid Baseline for 360° Depth Estimation

Georgios Albanis1 Nikolaos Zioulis1,2 Petros Drakoulis1 Vasileios Gkitsas1 Vladimiros Sterzentsenko1 Federico Alvarez2 Dimitrios Zarpalas1 Petros Daras1 1 Centre for Research and Technology Hellas, Thessaloniki, Greece 2 Universidad Politécnica de Madrid, Madrid, Spain {galbanis,nzioulis,petros.drakoulis,gkitsasv,vladster}@iti.gr fag@gatv.ssr.upm.es {zarpalas,daras}@iti.gr vcl3d.github.io/Pano3D

Abstract

Pano3D is a new benchmark for depth estimation from spherical panoramas. It aims to assess performance across all depth estimation traits, the primary direct depth estimation performance targeting precision and accuracy, and also the secondary traits, boundary preservation and smoothness. Moreover, Pano3D moves beyond typical intra-dataset evaluation to inter-dataset performance assessment. By disentangling the capacity to generalize in unseen data into different test splits, Pano3D represents a holistic benchmark for 360° depth estimation. We use it as a basis for an extended analysis seeking to offer insights into classical choices for depth estimation. This results into a solid baseline for panoramic depth that followup works can built upon to steer future progress.

1. Introduction

Benchmarks are the drivers of progress as they facilitate measurable technical increments, and can also provide explainable insights for diverging technical approaches. They must be unbiased, especially given the emergence of data-driven methods that can easily exploit any hidden bias in the data. Besides, the expressiveness of deep models necessitates the enrichment of benchmarks with varying data distributions to allow for the assessment of their generalization and their capacity to exploit different sources of data.

The recent availability of 360° depth datasets out of stitched raw sensor data [1, 7], 3D reconstruction renderings [56, 55], and photorealistic synthetic scenes [34, 53] has stimulated research in monocular 360° depth estimation [44, 13, 45, 26, 52, 43]. Still, the progress in monocular depth estimation has been mainly driven by research for traditional cameras, and assessed on perspective datasets, starting with the pioneering work of [14]. Even though other approaches exist (e.g. ordinal regression [16, 3]), depth estimation is most typically addressed as a dense regression objective. Various estimator choices are available for the direct objective like L1, L2, or robust versions like the reverse Huber (berHu) loss [32]. Complementary errors have also been introduced like the virtual normal loss [51] which captures longer range depth relations. Additional smoothness ensuring losses can be used to enforce a reasonable and established prior of depth maps, which is their piece-wise smoothly spatially varying nature [23].

Depth maps also exhibit sharp edges at object bound-
aries [23], whose preservation is important for various downstream applications. Recent works which focus explicitly on improving the estimated boundaries introduced new metrics to measure boundary preservation performance [21, 39]. Since convolutional data-driven methods spatially downscale the encoded representations, predicting neighboring values relies on neighborhood information, leading to interpolation blurriness. Counteracting approaches, like encoder-decoder skip connections or guided filters [47], can lead to texture transfer artifacts, hurting the predictions’ smoothness. The latter (smoothness) is also an important trait for some tasks like scene-scale 3D reconstruction which usually relies on surface orientation information [29] to preserve structural planarities, while the former (boundaries) are necessary for applications like view synthesis [2] or object retrieval [28]. Still, smoothness related metrics are usually presented on surface [46, 27] or plane [12] estimation works. Further, the balance between them needs to be tuned as they are conflicting objectives.

In this work we set to deliver an unbiased and holistic benchmark for monocular 360° depth estimation that provides performance analysis across all traits, i) depth estimation, ii) boundary preservation, iii) smoothness. We also consider an orthogonal evaluation strategy that seeks to assess the models’ generalization as well, across its different facets, i) varying depth distributions, ii) adaptation to the scenes’ contexts, and iii) different camera domains. To support the benchmark, we design a set of solid baselines that respect best practises as reported in the literature and rely on standard architectures. Our results, data, code, configurations and trained models are publicly available at vcl3d.github.io/Pano3D/.

- We show that recently made available datasets contain significant biases or artifacts that prevent them from being suitable as solid benchmarks.
- We provide 360° depth estimation performance results for all different traits, across different domains, contexts, distributions and resolutions, while also taking depth refinement advances into account.
- We demonstrate the effectiveness of skip connections, a rare architectural choice for (360°) depth estimation.

2. Related Work

Monocular Omnidirectional Depth Estimation. The first works addressing the monocular data-driven omnidirectional depth estimation task were [44] and [56]. The former applied traditional CNNs trained on perspective images in a distortion-aware manner to spherical images, while the latter introduced a rendered spherical dataset of paired color and depth images, in addition to a simplistic rectangular filtering preprocessing block. Pano Popups [12] simultaneously predict depth and surface orientation to construct planar 3D models, showing the insufficiency of depth estimates along to approximate planar regions.

The generalized Mapped Convolutions [13] were applied to omnidirectional depth estimation, showing how accounting for the distortion when using equirectangular projection increases performance in the image regions closer to the equator. Although these spatially imbalanced predictions are an important issue to address for 360° depth estimation methods, the usual evaluation methodologies do not address this apart from [55]. The omnidirectional extension networks [9] employ a near field-of-view (NFoV) perspective depth camera to accompany the spherical one, offering a necessary, albeit not full FoV, constraint to enhance the preservation of details in the inferred depth map.

Recent omnidirectional depth estimation works diverged in two paths. One route is to exploit the nature of the spherical images within network architectures, with BiFuse [45] fusing features from a cubemap and an equirectangular representation, while UniFuse [26] shows that the fusion of cubemap features to the equirectangular ones is more effective. HoHoNet [43] adapts classical CNNs to operate on 360° images by flattening the meridians to DCT coefficients, allowing for efficient dense feature reconstruction, and applying it to monocular depth estimation from spherical panoramas. Other recent works [34, 52] explore the connection between the layout and depth estimation tasks, while [15] relies on the joint optimization between depth and surface orientation estimates using a UNet model [41].

Monocular Perspective Depth Estimation. The pioneering work for data-driven monocular dense depth estimation [14] employed a scale-invariant loss and established the set of metrics used to evaluate follow-up works. Naturally, the progress in monocular depth estimation for perspective images is larger, as traditional images find more widespread use. While impressive gains have been presented using ordinal regression [16] or adaptive binning [3], they have not been applied to 360° depth estimation, which exhibits more complex depth distributions than perspective depth maps due to its holistic FoV.

Results like the berHu loss presented in [32] have found traction in omnidirectional models as they are more easily transferable. On the contrary the more recently presented virtual normal loss [51] has not been applied to 360° depth, albeit its longer range depth relation modelling is highly aligned with the global reasoning required for the spherical task. Recently, the balance between the multitude of losses required to balance smoothness, boundary preservation and depth accuracy were investigated in [33] to help models initially focus on easier to optimize losses (i.e. depth accuracy), and then on harder ones (i.e. smoothness, boundary).

Regarding depth discontinuity preservation performance, [21] showed that a combination of three different loss terms, a depth, a surface and a spatial derivative one,
Figure 2. Existing 360° depth benchmarks are either synthetic (Structured3D – tomato red) or biased (3D60, with an extra light source – orange). In addition to a holistic evaluation scheme, our benchmark Pano3D fixes the lighting issues (Matterport3D data – violet), and additionally offers a variety of zero-shot cross dataset transfer splits (generated from GibsonV2), permitting the validation of close-to-real-world performance. The tiny (dark green) and medium (dark blue) splits contain residential building scenes but with a different depth distribution as presented on the left of the figure. The fullplus (light green) split, albeit of similar distribution, contains different context scenes, like super-markers, garages and construction sites. Finally, we additionally generate camera domain adapted splits like the tiny filmic (purple) on the right, or the fullplus filmic (not shown), effectively capturing all different generalization performance traits.

help increase performance at object boundaries. Similarly, a boundary consistency was introduced in [25] to overcome blurriness and bleeding artifacts. Another approach, based on learnable guided filtering [47], exploits the color image as guidance. Recently, displacement fields [39] showed that predicting resampling offsets instead of residuals is more suitable to increase performance at sharp depth discontinuities, while preserving depth estimation accuracy.

3. Methodology

Our goal is two-fold, first to deliver a new benchmark for 360° depth estimation, and second, to methodically analyze the task in light of recent developments, to identify a set of solid baselines, which future works will use as the starting points for assessing performance gains. Section 3.1 introduces the benchmark data, which set the ground for the subsequent analysis. Section 3.2 describes the benchmark’s holistic approach in terms of evaluation, while Section 3.3 presents the experiment design rationale.

3.1. Dataset

Up to now 360° depth datasets either rendered purely synthetic scenes like Structured3D [53] or 360D’s SunCG and SceneNet parts [56], or relied on 3D scanned datasets like Matterport3D [7] and Stanford2D3D [1]. The latter offer both panoramas and the 3D textured meshes, with some works using the original Matterport camera derived data and others the rendered panoramas from the 3D scanned meshes. Both approaches come with certain drawbacks, the original data contain invalid (i.e. true black) regions towards the sphere’s poles, while the 3D rendered data contain invalid regions where the 3D scans failed to reconstruct the surface. At the same time, the original data present with stitching artifacts (mostly blurring), while the rendered data sometimes suffer from 3D reconstruction errors which manifest in color discontinuities. We opt for the generation via rendering approach [56] as it produces true spherical panoramas and higher quality depth maps at lower resolution compared to nearest neighbor sampling. However, we fix a critical issue of the 360D [56] and 3D60 [55] datasets, namely, the introduction of a light source that alters the scene’s photorealism. Instead, we only sample the raw diffuse texture, preserving the original scene lighting, a crucial factor for unbiased learning and performance evaluation.

Zero-shot Cross-Dataset Transfer. However, there is a need to move beyond traditional train/test split performance analysis to support model deployment in real-world conditions. Thus, assessing generalization performance is very important. Towards that end, apart from re-rendering the Matterport3D scans for training, we introduce a new 360° color-depth pair generated dataset from the 572 GibsonV2 (GV2) [48] 3D scans. Compared to Matterport3D’s (M3D) 90 buildings, it is a vastly larger dataset, whose scenes offer higher variety as well. These renders can be used for assessing generalization performance across its different splits: tiny, medium, full [3] and fullplus. After removing outlier scans and filtering samples (keeping those with ≤ 10% invalid pixels), we are left with 7170/1527 train/test M3D samples, and 2740, 6999, 3284, 21203 GV2 tiny, medium, fullplus, and full split samples respectively.

3.2. Metrics

Since the introduction of the first set of metrics for data-driven depth estimation [14], namely root mean squared error (RMSE), room mean squared logarithmic error (RM
Table 1. Direct depth metrics performance across models and supervision schemes. Best three performers are denoted with bold faced light green (1st), light blue (2nd) and light purple (3rd) respectively following the ranking order. Same scheme applies to all tables.

| Model | Depth Error ↓ | Depth Accuracy ↑ |
|-------|--------------|------------------|
|       | RMSE | RMSEL | AbsRel | SqRel | δ1,05 | δ1,1 | δ1,25 | δ1,255 | δ1,25 | δ1,255 |
| L1    | 0.4817 | 0.0780 | 0.1213 | 0.0933 | 34.59% | 59.98% | 87.25% | 96.30% | 98.50% |
| Lcosine | 0.4825 | 0.0782 | 0.1216 | 0.1014 | 37.04% | 60.96% | 87.48% | 96.36% | 98.46% |
| Lgrad  | 0.4616 | 0.0749 | 0.1163 | 0.0889 | 37.40% | 62.46% | 88.39% | 96.63% | 98.57% |
| Lcomb  | 0.4613 | 0.0740 | 0.1143 | 0.0892 | 38.56% | 63.31% | 88.70% | 96.68% | 98.62% |
| LcI   | 0.4640 | 0.0743 | 0.1165 | 0.0920 | 37.67% | 62.60% | 88.47% | 96.64% | 98.65% |
| UNet  | 0.4215 | 0.0233 | 0.1138 | 0.0744 | 37.54% | 60.47% | 88.05% | 97.01% | 98.81% |
| Lcosine | 0.4152 | 0.0841 | 0.1170 | 0.0736 | 34.06% | 59.75% | 88.13% | 97.13% | 98.99% |
| Lgrad  | 0.4061 | 0.0264 | 0.1135 | 0.0682 | 37.49% | 60.93% | 88.50% | 97.17% | 98.91% |
| Lcomb  | 0.4041 | 0.1459 | 0.1146 | 0.0692 | 37.24% | 60.44% | 88.31% | 97.15% | 99.04% |
| LcI   | 0.3967 | 0.1182 | 0.1095 | 0.0672 | 38.62% | 62.16% | 89.08% | 97.35% | 99.03% |
| DenseNet | 0.4672 | 0.5380 | 0.1223 | 0.0896 | 37.35% | 60.52% | 86.72% | 96.27% | 98.37% |
| Lcosine | 0.4603 | 0.0752 | 0.1145 | 0.0817 | 37.57% | 62.61% | 88.03% | 97.65% | 98.64% |
| Lgrad  | 0.4488 | 0.3847 | 0.1210 | 0.0827 | 33.25% | 59.71% | 87.39% | 96.73% | 98.63% |
| Lcomb  | 0.4490 | 0.2565 | 0.1129 | 0.0806 | 38.30% | 63.02% | 88.56% | 96.66% | 98.54% |
| LcI   | 0.4481 | 0.6177 | 0.1142 | 0.0805 | 39.28% | 63.34% | 88.49% | 96.66% | 98.43% |
| ResNet | 0.4755 | 0.1639 | 0.1310 | 0.0942 | 31.22% | 55.89% | 85.56% | 96.27% | 98.57% |
| Lcosine | 0.4700 | 0.0804 | 0.1279 | 0.0949 | 37.37% | 57.92% | 85.32% | 96.35% | 98.62% |
| Lgrad  | 0.4734 | 0.2495 | 0.1278 | 0.0916 | 35.23% | 57.34% | 85.54% | 96.20% | 98.50% |
| Lcomb  | 0.4573 | 0.1200 | 0.1272 | 0.0894 | 34.53% | 57.97% | 86.26% | 96.56% | 98.71% |
| LcI   | 0.4607 | 0.2938 | 0.1236 | 0.0862 | 34.75% | 59.16% | 86.11% | 96.60% | 98.66% |
| ResNeXt | 0.4373 | 0.2430 | 0.1161 | 0.0783 | 37.07% | 60.60% | 87.68% | 96.66% | 98.73% |
| Lcosine | 0.4347 | 0.0170 | 0.1139 | 0.0772 | 39.80% | 61.31% | 88.27% | 97.02% | 98.81% |
| Lgrad  | 0.4107 | 0.2710 | 0.1089 | 0.0717 | 38.93% | 63.31% | 89.51% | 97.32% | 98.92% |
| Lcomb  | 0.4165 | 0.0843 | 0.1102 | 0.0722 | 36.71% | 61.92% | 89.17% | 97.24% | 98.90% |
| LcI   | 0.4260 | 0.0967 | 0.1125 | 0.0756 | 39.92% | 62.53% | 88.22% | 97.04% | 98.88% |

SLE), absolute relative error (AbsRel), squared relative error (SqRel), and the relative threshold (t) based accuracies (δt), these metrics are the standard approach for evaluation depth estimation performance. More recent works have identified some shortcomings of these metrics. Specifically in [31] an expanded analysis of depth estimation quality measures was conducted, and focused on two important traits, planarity and discontinuities. The latter is very important for some downstream applications like view synthesis, and apart from the completeness (comp) and accuracy (acc) of the depth boundary errors (dbe) proposed in [31], another set of accuracy metrics were proposed in [21]. The precision, recall and their harmonic mean (F1-score) were used after extracting different boundary layers via Sobel edge thresholding. Planarity is also very important for various downstream applications, especially for indoor 3D reconstruction. Finally, to overcome resolution [4] and focal length variations [8], recent perspective depth estimation works resort to nearest-neighbor 3D metrics.

**Direct Depth Metrics.** We build upon these developments and design our benchmark to provide a holistic evaluation of depth estimation models. Given the progress of recent data-driven models we expand the δ accuracies with two lower thresholds, a strict at δ1,1, and a precise, δ1,05, similar to [25]. However, these metrics, when applied directly on equirectangular images, are biased by its distortion towards the poles. To remove this bias we take the spherically weighted mean (denoting with a w prefix), which is standard practise for 360° image/video quality assessment [50] and was also used in [55]. For the δt accuracies though, we turn to uniform sampling on the sphere using the projected vertices of a subdivided icosahedron, denoted as δt,cosK, K being the icosahedron’s order.

**Depth Discontinuity Metrics.** We complement the direct depth performance metrics with a set of secondary metrics measuring performance at preserving the depth discontinuities, usually manifesting at object boundaries. While [31] used manual annotation and structured edge detection [11], we follow the approach of [39] that relies on automatic Canny edge detection [5]. In addition, we complement the depth boundary errors (dbeacc and dbeimp) with the accuracy metrics of [21] (prec, rec) using the same thresholds, t ∈ {0.25, 0.5, 1.0}, for both set of metrics.

**Depth Smoothness Metrics.** While the planarity metric of [31] required the manual annotation of samples, its goal is to measure the smoothness of the inferred depth with respect to dominant structures. A straightforward adaptation that alleviates annotations is the use of surface orientation metrics, which is a property directly derived from the depth measurements. Using spherical-to-Cartesian coordinates conversion the depth/radius measurements are lifted to 3D points, with the surface orientation extracted by exploiting the structured nature of images. Similar to how surface estimation methods measure performance, we use the angular RMSE (RMSEo), and a set of accuracies αdbe, with pre-defined angle thresholds d, using those from [46] (d ∈ {11.25°, 22.5°, 30°}).

**Geometric Metrics.** Depth estimations are typically
used in downstream applications for metric-scale 3D perception. Therefore, 3D performance metrics are reasonable to assess suitability for downstream tasks. We use two different metrics that aggregate the performance of the aforementioned different depth traits, i.e., accuracy and precision, boundary preservation and smoothness. The first geometric metric is computed on the point cloud level (c2c), using a point-to-plane distance between each point and its closest correspondence with the ground truth point cloud. The point-to-plane distance jointly encodes depth correctness and smoothness, while the closest point query will penalize boundary errors. The second geometric metric is computed on the mesh level, having each point cloud (predicted and ground truth) 3D reconstructed using the Screened Poisson Surface Reconstruction [29]. We then calculate the Hausdorff distance [10] between the two meshes. Similarly, Poisson reconstruction leverages both position and surface information when generating the scene’s mesh. Through this metric, we assess the capacity to represent the entire scene’s geometry with the estimated depth, an important trait for some downstream applications.

### 3.3. Experimental Setup

We design our experiments and search for a solid baseline taking recent developments into account.

**Supervision.** As shown in [6] the L1 ($L_1$) loss exhibits the best convergence for monocular depth estimation irrespective of the model size and architecture complexity, indicating that models behaving like median estimators are more appropriate. Most recent works for 360° depth estimation [45, 26, 13] use the berHu loss [32], with the exception to this rule being [43] that uses the L1 loss.

We additionally observe that these works rely solely on a single direct depth loss, while recent works on perspective depth estimation also include additional losses. MiDaS [40], MegaDepth [35] as well as [49] and [21] use a multi-scale ($K = 4$ scales) gradient matching term ($L_{grad}$) that enforces consistent depth discontinuities. While their terms are scale-invariant and operate in the log-space, 360° depth does not suffer from disparity/baseline or focal length variations, and since we do not use the L1 loss in log-space (as its performance is inferior to pure L1 [6]), we use a non-scale invariant version of this loss. Apart from boundary preservation, the piece-wise smooth nature of depth, necessitates the use of a suitable prior for the predictions. This was acknowledged in [21], where a surface orientation consistency loss was used ($L_{cosine}$). Prior works employed smoothness priors on the predictions, and, to overcome cross-boundary smoothing, relied on image gradient weighting [18]. Yet, image gradients do not necessarily align with depth discontinuities, making the normal loss a better candidate.

Finally, the newly introduced virtual normal loss [51] ($L_{vnl}$) is a long-range relationship oriented objective, which given the global context of spherical panoramas is well aligned with the task. In our experiments, we follow a progressive loss ablation starting with a $L_1$ objective, examining the effect of $L_{grad}$ and $L_{cosine}$ on the $L_1$ baseline, as well as their combined effect $L_{comb}$, and further extend the combined objective with $L_{vnl}$, with the latter experiment including all losses.

**Model Architecture.** The importance of high-capacity encoders, pre-training and multi-scale predictions is acknowledged in the literature [40]. Building on the first, we preserve a consistent convolution decoder and use a DenseNet-161 [22] (55M parameters) and ResNet-152 [20] (110M parameters) encoder as baselines. Inspired by recent work [33] we also include their Pnas model (99M parameters) whose encoder is a product of neural architecture search [36]. In addition, taking into account the boundary preservation performance of skip connections, we also use the – largely unpopular for depth estimation – UNet model [41] (27M parameters). Since it is a purely convolutional model, we additionally modify the ResNet-152 model with skip connections starting from the first residual block (112M parameters), in contrast to UNet’s very early layer encoder-to-decoder skip. Since pre-training weights are not available for UNet, we experiment with cold-started models, and also simplify training using a single-scale predictions as the multi-scale effect should be horizontal across all models with the same convolution decoder structure.

**Periodic Displacement Fields Refinement.** We additionally consider the refinement of the predicted depth, using a shallow hourglass module [38]. It is adapted for the task at hand, with two branches, one for the input color image and the other for the predicted depth map. Across each stage, we account for the varying nature of each branch’s feature statistics using Adaptive Instance Normalization [24]. We follow the recent approach of [39] that shows how predicting displacement fields instead of residuals produces higher quality depth refinement. However, the spherical domain is continuous, and thus, we need to account for the horizontal discontinuity of the equirectangular projection. To achieve this in a locally differentiable manner, we resort to a periodic reconstruction of the sampling coordinates. Considering the final sampling coordinates ($\phi, \theta$) after adding the displacement field, we wrap them around to ($\tau, \theta$), with $\tau = \arctan2(-\sin \phi, -\cos \phi) + \pi$.

**Training and Evaluation.** We train all models solely on the official train split of M3D, and evaluate them on its official test split as well. Evaluation is conducted across all the aforementioned axes of depth performance. Apart from this holistic performance analysis, we additionally take an orthogonal direction and assess the models’ generalization performance on zero-shot cross-dataset transfer using the GV2 tiny, medium and fullplus splits. Given that both GV2 and M3D scenes were scanned with same type
of camera (i.e. Matterport), we render another version of tiny which is tone mapped to a film-like dynamic range, dubbed tiny-filmic, changing the camera-related data domain. Our experiments are conducted on two different resolutions $512 \times 256$ and $1024 \times 512$ (we render all datasets to both) to assess cross-resolution performance.

4. Analysis

Implementation Details. We implement all experiments with moai [37], using the same seed across all experiments. For data generation we use Blender and the Cycles path tracer using 256 samples. Our ResNets are built with pre-activated bottleneck blocks [20] and all our models’ weights are initialized with [19]. We optimize all models for 60 epochs on a NVidia 2080 Ti, using Adam [30] with a learning rate of 0.0002 and default momentum parameters, and a consistent batch size of 4. All losses are unity (i.e. equally) weighted across all experiments. We use CloudCompare to calculate the $e2e$ distance [17], and MeshLab to calculate the $m2m$ distance [10]. During evaluation, we consider the raw values predicted by the models and clip the valid depth range to 10m.

Which loss combination offers better performance? Contrary to their focused nature both $L_{cosine}$ and $L_{grad}$ increase depth estimation performance across all models when complementing the direct $L_1$ objective, as evident in Table 1. In addition, they provide the expected boost in smoothness/discontinuity preservation across all models as presented in our supplementary material appended after the references. When viewed purely from a depth estimation perspective, it is observed that their combination, $L_{comb}$, benefits performance. But, when examining the specific depth traits that they seek to enforce, their conflicting nature is also apparent. Overall, we observe almost all models achieve highest overall performance when both losses are present, with, our without the virtuean normal loss (VNL) which is added in the $L_{vnl}$ case. The latter greatly boosts the UNet model, which is reasonable as the localised nature of skip connections is aided by the global depth constraints that VNL introduces.

Which architecture is better performing? We compare architectures after selecting the best performing models for each, which for UNet is $L_{vnl}$, and for the rest the $L_{comb}$. The rationale for choosing $L_{comb}$ for ResNet skip is that while $L_{grad}$ behaves better on direct depth metrics (except closer distances, as indicated by the RMSLE), there is a large performance gap in the discontinuity and smoothness metrics, compared to the performance discrepancy on depth estimation. Table 2 presents the results using the spherical metrics that account for the distortion. These are unbiased metrics, which is evident given the deteriorated performance across all metrics compared to those estimated on the image level on each equirectangular panorama. A more straightforward comparison is available in our supplementary material appended after the references. Interestingly, we observe that models employing encoder-decoder skip connections exhibit better performance both in direct depth metrics (Table 2). Curiously, contrary to the expectation as set by the literature [40] that high-capacity encoders are required, the UNet architecture showcases the best performance. Regarding domain oriented techniques, we train the better performing model with circular padding [42, 54] that connects features across the horizontal equirectangular boundary, denoted as circ. Evidently, this simple scheme increases the performance across all metrics, allowing the model to exploit its spherical nature.

Is this performance consistent when considering secondary traits? Regarding the discontinuity and smoothness traits as presented in Table 3 it is evident that skip connections result in higher performance, but especially for the dominating UNet, at the expense of the smoothness trait. This is reasonable as early layer skip connections result into texture transfer, and further evidenced by the improved performance of ResNet skip, which lacks early layer skips, on both discontinuity and smoothness metrics. Overall, UNet achieves the best performance on depth and discontinuity metrics at the expense of the smoothness trait and closer range performance as indicated by its inferior RMSLE. On the other side, the different metrics indicate that the PNAS model produces oversmoothed results that are more metrically accurate and precise at closer distances. Nonetheless, ResNet skip achieves a better balance without significant sacrifices across the secondary traits.

How helpful is depth refinement? We also examine the effect of a shallow depth refinement module on these models, with the results after training for 10 epochs presented in Table 3. All models, apart from UNet, improve their performance at boundary preservation while also preserving depth estimation performance, but at the expense of smoothness, with the exception in this case being ResNet skip. For UNet specifically, texture transfer leads to noise, which prevents an interpolation-based warping technique to improve results, as it was designed to improve smooth depth predictions. However, ResNet skip closes the performance gap and even improves smoothness performance, further solidifying its well-balanced nature.

Why this benchmark? Table 2 shows the performance of the two higher performing models when trained and tested on other recently introduced 360° depth datasets, namely 3D60 [55] which is an extension of [56] and Structured3D [53] with $512 \times 256$ and $1024 \times 512$ resolutions respectively. All metrics are significantly higher which evidences their insuitability to be used for benchmarking progress. This is largely because of their inherent bias which is the result of lighting for 3D60, which includes an extra light source at the center, as also explained in [26], an unfortunate bias
Table 2. Direct depth performance using spherical metrics. A UNet model with spherical padding is also presented (light pink), as well as the two better performing models trained and tested on the 3D60 (light orange) and Structured3D (light red) datasets.

| Model | wRMSE | wRMSL | wAbsRel | wSqRel | \(\delta^{\text{acc}}_{\text{1,05}}\) | \(\delta^{\text{prec}}_{\text{1,05}}\) | \(\delta^{\text{rec}}_{\text{1,05}}\) | \(\alpha_{\text{1,25}}\) | \(\alpha_{\text{1,25}}\) | \(\alpha_{\text{1,25}}\) | \(\alpha_{\text{1,25}}\) |
|-------|-------|--------|---------|--------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Pnas\textsuperscript{comb} | 0.5367 | 0.0811 | 0.1259 | 0.1193 | 36.44% | 60.52% | 86.80% | 95.83% | 98.11% |
| UNet\textsuperscript{comp} | 0.4520 | 0.1300 | 0.1147 | 0.0811 | 36.68% | 60.59% | 88.21% | 96.96% | 98.73% |
| DenseNet\textsuperscript{comb} | 0.5209 | 0.1982 | 0.1209 | 0.1013 | 35.97% | 60.41% | 87.02% | 95.96% | 98.09% |
| ResNet\textsuperscript{comb} | 0.5294 | 0.1365 | 0.1374 | 0.1127 | 32.03% | 55.31% | 84.74% | 95.81% | 98.21% |
| ResNet\textsuperscript{comp} | 0.4788 | 0.0927 | 0.1166 | 0.0893 | 36.20% | 60.64% | 87.99% | 96.62% | 98.49% |
| UNet\textsuperscript{comp} @ 3D60 | 0.4399 | 0.0685 | 0.1132 | 0.0769 | 36.85% | 61.38% | 88.84% | 97.25% | 98.89% |
| ResNet\textsuperscript{comb} @ 3D60 | 0.3140 | 0.0545 | 0.0741 | 0.0316 | 49.99% | 75.16% | 95.49% | 99.11% | 99.60% |
| ResNet\textsuperscript{comp} @ 3D60 | 0.3758 | 0.6100 | 0.0883 | 0.0481 | 46.03% | 70.29% | 93.12% | 98.41% | 99.34% |
| ResNet\textsuperscript{comp} @ 3D3 | 0.1815 | 0.0546 | 0.0919 | 0.0398 | 50.61% | 75.98% | 92.23% | 96.56% | 97.53% |

Table 3. Top half: Depth discontinuity/boundary preservation and depth smoothness performance metrics. Bottom half: Same metrics after refining all models (subscript \textit{ref}) with a periodic displacement field hourglass module.

| Model | \(\delta^{\text{acc}}_{\text{1,05}}\) | \(\delta^{\text{prec}}_{\text{1,05}}\) | \(\delta^{\text{rec}}_{\text{1,05}}\) | \(\alpha_{\text{1,25}}\) | \(\alpha_{\text{1,25}}\) | \(\alpha_{\text{1,25}}\) | \(\alpha_{\text{1,25}}\) |
|-------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Pnas\textsuperscript{comb} | 59.87% | 57.54% | 51.85% | 43.96% | 36.69% | 28.59% | 26.00% |
| UNet\textsuperscript{ref} | 58.97% | 57.54% | 51.85% | 43.96% | 36.69% | 28.59% | 26.00% |
| DenseNet\textsuperscript{comb} | 51.94% | 48.43% | 40.63% | 30.61% | 26.14% | 22.46% | 19.25% |
| ResNet\textsuperscript{comb} | 50.77% | 46.52% | 36.58% | 28.31% | 23.33% | 19.37% | 16.68% |
| ResNet\textsuperscript{comp} | 57.34% | 54.11% | 47.57% | 33.99% | 24.30% | 16.37% | 15.27% |
| UNet\textsuperscript{ref} @ 3D60 | 50.90% | 47.64% | 40.16% | 31.72% | 25.58% | 18.04% | 16.86% |
| ResNet\textsuperscript{comb} @ 3D60 | 51.94% | 48.43% | 40.63% | 30.61% | 26.14% | 22.46% | 19.25% |
| ResNet\textsuperscript{comp} @ 3D60 | 51.94% | 48.43% | 40.63% | 30.61% | 26.14% | 22.46% | 19.25% |
| ResNet\textsuperscript{comp} @ 3D3 | 60.78% | 58.09% | 51.49% | 37.79% | 32.55% | 27.23% | 15.05% |

Table 4. Consolidated performance on the GibsonV2 splits, across all depth traits, using a subset of the available metrics, for models trained on the Matterport3D data. The best performing model (UNet) trained with photometric augmentation is also presented (light pink).

| GV2 | Model | \(\delta^{\text{acc}}_{\text{1,05}}\) | \(\delta^{\text{prec}}_{\text{1,05}}\) | \(\delta^{\text{rec}}_{\text{1,05}}\) | \(\alpha_{\text{1,25}}\) | \(\alpha_{\text{1,25}}\) | \(\alpha_{\text{1,25}}\) | \(\alpha_{\text{1,25}}\) |
|-----|-------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Pnas\textsuperscript{comb} | 60.02% | 58.09% | 51.49% | 37.79% | 32.55% | 27.23% | 20.64% |
| UNet\textsuperscript{ref} | 60.02% | 58.09% | 51.49% | 37.79% | 32.55% | 27.23% | 20.64% |
| DenseNet\textsuperscript{comb} | 51.94% | 48.43% | 40.63% | 30.61% | 26.14% | 22.46% | 19.25% |
| ResNet\textsuperscript{comb} | 50.77% | 46.52% | 36.58% | 28.31% | 23.33% | 19.37% | 16.68% |
| ResNet\textsuperscript{comp} | 57.34% | 54.11% | 47.57% | 33.99% | 24.30% | 16.37% | 15.27% |


Table 5. Consolidated depth performance across all traits using a subset of the available metrics. Results are presented for the two best performing models on the GibsonV2 splits which have been trained on the Matterport3D train split.

| GV2 | Model | Error ↓ | Direct Depth Accuracy ↑ | Error ↓ | Depth Discontinuity Accuracy ↑ | Error ↓ | Depth Smoothness Accuracy ↑ |
|-----|-------|---------|--------------------------|---------|-------------------------------|---------|-----------------------------|
|     |       | m2m     | wRMSE | wRMSE | wAbsRel | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/w | δh/w | δd/
cilitate the assessment towards real-world applicability of data-driven models for 360° geometric inference.

Supplementary Material.

We provide extra quantitative and qualitative comparisons in the supplementary material following the references. Supplementing experiments also reproduce prior work used as a basis for designing our methodology. Finally, a live web demo of our baseline models can be found at share.streamlit.io/tzole1155/ThreeDit.

Acknowledgements

This work was supported by the EC funded H2020 project ATLANTIS [GA 951900].

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105, 2017. 1, 3
[2] Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. Matryoshka: Real-time 6dof video view synthesis using multi-sphere images. In European Conference on Computer Vision, pages 441–459. Springer, 2020. 2
[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. arXiv preprint arXiv:2011.14141, 2020. 1, 2
[4] Cesar Cadena, Yasir Latif, and Ian D Reid. Measuring the performance of single image depth estimation methods. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 4150–4157. IEEE, 2016. 4
[5] John Canny. A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence, (6):679–698, 1986. 4, 15
[6] Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Andrés Almansa, and Frédéric Champagnat. On regression losses for deep depth estimation. In 2018 25th IEEE International Conference on Image Processing (ICIP), pages 2915–2919. IEEE, 2018. 5, 11
[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In 7th IEEE International Conference on 3D Vision (3DV), pages 667–676. Institute of Electrical and Electronics Engineers Inc., 2018. 1, 3
[8] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 679–688, 2020. 4
[9] Xinjing Cheng, Peng Wang, Yanqi Zhou, Chenye Guan, and Ruigang Yang. Omnidirectional depth extension networks. In 2020 IEEE International Conference on Robotics and Automation (ICRA), pages 589–595. IEEE, 2020. 2
[10] Paolo Cignoni, Claudio Rocchini, and Roberto Scopigno. Metro: measuring error on simplified surfaces. In Computer graphics forum, volume 17, pages 167–174. Wiley Online Library, 1998. 5, 6
[11] Piotr Dollár and C Lawrence Zitnick. Fast edge detection using structured forests. IEEE transactions on pattern analysis and machine intelligence, 37(8):1558–1570, 2014. 4
[12] Marc Eder, Pierre Moulon, and Li Guan. Pano popups: Indoor 3d reconstruction with a plane-aware network. In 2019 International Conference on 3D Vision (3DV), pages 76–84. IEEE, 2019. 2
[13] Marc Eder, True Price, Thanh Vu, Akash Bapat, and Jan-Michael Frahm. Mapped convolutions. arXiv preprint arXiv:1906.11096, 2019. 1, 2, 5
[14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In Advances in neural information processing systems, pages 2366–2374, 2014. 1, 2, 3
[15] Brandon Yushan Feng, Wangjue Yao, Zheyuan Liu, and Amitabh Varshney. Deep depth estimation on 360° images with a double quaternion loss. In 2020 International Conference on 3D Vision (3DV), pages 524–533. IEEE, 2020. 2
[16] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2002–2011, 2018. 1, 2
[17] Daniel Girardeau-Montaut. Détection de changement sur des données géométriques tridimensionnelles. Theses, Télécom ParisTech, May 2006. 6
[18] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 270–279, 2017. 5
[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pages 1026–1034, 2015. 6
[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In European conference on computer vision, pages 630–645. Springer, 2016. 5, 6
[21] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1043–1051. IEEE, 2019. 2, 4, 5
[22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7470–7478, 2017. 5
[23] Jinggang Huang, Ann B Lee, and David Mumford. Statistics of range images. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662), volume 1, pages 324–331. IEEE, 2000. 1, 2
[24] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 5

[25] Yu-Kai Huang, Tsung-Han Wu, Yueh-Cheng Liu, and Winston H Hsu. Indoor depth completion with boundary consistency and self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3, 4

[26] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *arXiv preprint arXiv:2102.03550*, 2021. 1, 2, 5, 6, 13

[27] Antonios Karakottas, Nikolaos Zioulis, Stamatis Samaras, Dimitrios Ataloglou, Vasileios Gkitsas, Dimitrios Zarpalas, and Petros Daras. 360° surface regression with a hypersphere loss. In *2019 International Conference on 3D Vision (3DV)*, pages 258–268. IEEE, 2019. 2

[28] Kevin Karsch, Zicheng Liao, Jason Rock, Jonathan T Barron, and Derek Hoiem. Boundary cues for 3d object shape recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2163–2170, 2013. 2

[29] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 2, 5, 16

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[31] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 4

[32] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 1, 2, 5, 11

[33] Jae-Han Lee and Chang-Su Kim. Multi-loss rebalancing algorithm for monocular depth estimation. In *Proceedings of the 2020 European Conference on Computer Vision (ECCV), Glasgow, UK*, pages 23–28, 2020. 2, 5

[34] Jin Lei, Xu Yanyu, Zheng Jia, Zhang Junfei, Tang Rui, Xu Shugong, Yu Jingyi, and Gao Shenghua. Geometric structure based and regularized depth estimation from 360° indoor imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[35] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 5

[36] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018. 5

[37] moai: Accelerating modern data-driven workflows. *https://github.com/ai-in-motion/moai*, 2021. 6

[38] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 5

[39] Michaël Ramamonjisoa, Yuming Du, and Vincent Lepetit. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14648–14657, 2020. 2, 3, 4, 5

[40] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 5, 6, 11

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 5

[42] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. HorizonNet: Learning room layout with 1d representation and pano stretch data augmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1047–1056, 2019. 6

[43] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hoohon: 360 indoor holistic understanding with latent horizontal features. *arXiv preprint arXiv:2011.11498*, 2020. 1, 2, 5, 11

[44] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 707–722, 2018. 1, 2

[45] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 462–471, 2020. 1, 2, 5, 12, 21

[46] Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, and Jan-Michael Frahm. Vpnet: Deep single view normal estimation with vanishing points and lines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 689–698, 2020. 2, 4

[47] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Fast end-to-end trainable guided filter. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1838–1847, 2018. 2, 3

[48] Fei Xia, Amir R Zamir, Zhixiang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9068–9079, 2018. 3

[49] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 611–620, 2020. 5
A. Supplementary Material

This supplementary material complements our original manuscript with additional quantitative results, offering extra ablation experiments, providing qualitative results on real data and comparisons between the different architectures.

A.1. Quantitative Results

Table 7 complements Table 1 of the main document, presenting the performance of all remaining metrics, namely the spherical direct depth metrics, the boundary preservation metrics, and the smoothness metrics. In addition, Figure 3 presents the different models’ performance in terms of three indicators, one for each trait. These indicators combine an error and an accuracy metric:

\[ i_d = \frac{1}{(1 - \delta_{1:25}) \times RMSE} \]  

\[ i_b = \frac{1}{(1 - (F_{0.25} + F_{0.5} + F_{1.0})/3) \times dbe_{acc}} \]  

\[ i_s = \frac{1}{(1 - (\alpha_{1.25} + \alpha_{2.50} + \alpha_{3.30})/3) \times RMSE} \]  

with \( i_d, i_b, \) and \( i_s \) the depth, boundary and smoothness performance indicators. Evidently, UNet performs significantly better than the other models, especially in the boundary consistency metrics, while all models benefit of the addition of extra losses. The addition, of skip connections in a common ResNet architecture offers better performance. While \( L_{grad} \) offers better depth performance for \( \text{ResNet}_{skip} \), the variant trained with \( L_{comb} \) offers higher performance across the two secondary traits.

In addition, we complement the main’s paper spherical metrics Table 8 by collating the traditional ones for a straightforward comparison.

Finally, Table 9 reproduces the grounds upon our methodology was designed, namely the efficacy of pre-trained models [40] and the L1 loss [6]. We use the DenseNet and Pnas models with the encoders initialized using weights pre-trained on ImageNet. Both claims stand, with all pre-trained models achieving better performance than the model trained from scratch. In addition, the L1 loss outperforms both berHu [32] and log loss. Interestingly, the performance drops significantly in DenseNet when trained with other losses, while for Pnas the performance gap is smaller. Therefore, when benchmarking different models, this needs to be taken into account as well. Only through consistent experimentation across different aspects measurable and explainable progress will be possible.

A.2. Qualitative Results

Finally we present additional qualitative results for different models. Apart from the collation of the predicted depth maps between the different models, we provide an advantage visualisation technique similar to that presented in HoHoNet [43]. The visualisation is the MAE difference across the two secondary traits.

To that end, Figure 4 demonstrates the comparison of ResNet and \( \text{ResNet}_{skip} \) architectures, Figure 5 that of the UNet and Pnas architectures, and, finally Figure 6 presents the differences between the UNet and \( \text{ResNet}_{skip} \) architectures.

Additionally, Figure 7 presents comparative results regarding the boundary preservation performance across models. Once again, UNet is able to capture finer-grained details while the Pnas model produces smoother results. Similarly, the differences between ResNet and \( \text{ResNet}_{skip} \), attributed to the addition of the skip connections are apparent across all samples.

Nonetheless, Pnas better captures the global context as seen in Figure 8, where the scene’s dominant planar surfaces are better preserved by it than UNet.

Figures 9, 10, 11 demonstrate qualitative results in GV2 tiny split for the UNet, Pnas, and \( \text{ResNet}_{skip} \) architectures respectively. Apart from the predicted point cloud we visualise the c2c error on the ground truth point cloud, with a
Table 7. Three axis depth metrics performance across models and supervision schemes. Best three performers are denoted with bold faced light green (1°), light blue (2°) and light purple (3°) respectively following the ranking order. Same scheme applies to all tables.

Table 8. Direct depth performance using spherical and conventional metrics. Bottom part results are the same as those presented in Table 2 of the original document. Top part are the corresponding results from Table 1 of the original manuscript.

blue-green-red colormap denoting the error’s magnitude. Finally, Figures 12 and 13 offer qualitative results of our best performing method in real world, in-the-wild, data captures. We also qualitatively compare our predictions with a state-of-the-art 360° depth estimation model (i.e. BiFuse [45]). It is worth highlighting that even the two of the three 360° images are captured by a panorama camera, and as
| model     | pretrained | $\mathcal{L}$ | $\text{RMSE}$ | $\text{RMSLE}$ | $\text{AbsRel}$ | $\text{SqRel}$ | $\delta_{1.25}$ |
|-----------|------------|----------------|----------------|----------------|-----------------|---------------|-----------------|
| **DenseNet** | ✓          | $\mathcal{L}_1$ | 0.4072         | 0.3194         | 0.1140          | 0.0694        | 88.91%          |
|           | ✓          | $\mathcal{L}_{\text{log}}$ | 0.5597         | 0.5720         | 0.1528          | 0.4475        | 80.48%          |
|           | ✓          | $\mathcal{L}_{\text{berHu}}$ | 0.4532         | 0.3754         | 0.1228          | 0.0852        | 86.68%          |
| **Pnas**   | ✓          | $\mathcal{L}_1$ | 0.3998         | 0.0634         | 0.0975          | 0.0661        | 91.91%          |
|           | ✓          | $\mathcal{L}_{\text{log}}$ | 0.4135         | 0.0656         | 0.0999          | 0.0697        | 91.09%          |
|           | ✓          | $\mathcal{L}_{\text{berHu}}$ | 0.4059         | 0.0643         | 0.0992          | 0.0666        | 91.56%          |

such there are artifacts. Yet, it seems that this does not greatly affect the performance of models. The UNet model produces higher quality depth estimates than BiFuse, albeit trained only on the train split of M3D, while the publicly available BiFuse model, as reported in UniFuse [26], has been trained on the entire M3D dataset.
Figure 4. Qualitative comparison between the ResNet and ResNet\textsubscript{skip} architectures. On the right the advantage visualization shows with blue color the areas where the former performs better, and with red color the areas where the latter performs better. The color magnitude corresponds to the MAE difference between the two models, illustrating the performance deviation between the two models. The addition of skip connections allows ResNet\textsubscript{skip} to capture finer-grained details.

Figure 5. Qualitative comparison between the UNet and Pnas architectures. On the right the advantage visualization shows with blue color the areas where the former performs better, and with red color the areas where the latter performs better. The color magnitude corresponds to the MAE difference between the two models, illustrating the performance deviation between the two models. Pnas provides smoother results while it is clear that UNet is able to capture finer-grained details.
Figure 6. Qualitative comparison between the UNet and ResNet skip architectures. On the right the advantage visualization shows with blue color the areas where the former performs better, and with red color the areas where the latter performs better. The color magnitude corresponds to the MAE difference between the two models, illustrating the performance deviation between the two models.

Figure 7. Boundary preservation qualitative comparison between the UNet, Pnas, ResNet and ResNet skip models. Boundaries are extracted by applying a Canny edge detector [5] with predefined thresholds on normalized predicted depth maps, and then are blended with the original color panorama. From left to right: i) GT depth (green), ii) UNet (orange), iii) Pnas (cyan), iv) ResNet (yellow), and v) ResNet skip (magenta).
Figure 8. Qualitative comparison of the Pnas and UNet models in surface reconstruction. From top to bottom: i) input color panorama, ii) Pnas normal map from the estimated depth map, iii) UNet normal map, iv) Pnas Screened Poisson Surface Reconstruction [29] 3D surface reconstruction, v) UNet 3D surface reconstruction, vi) overlaid Pnas (cyan) and UNet (pink) 3D surface reconstructions from birds eye view.
Figure 9. UNet qualitative results. From left to right: i) Input color panorama, ii) colored predicted point cloud, and iii) heatmap visualization of the $c2c$ error on the ground truth point cloud.
Figure 10. Pnas qualitative results. From left to right: i) Input color panorama, ii) colored predicted point cloud, and iii) heatmap visualization of the $c2c$ error on the ground truth point cloud.
Figure 11. ResNetquip qualitative results. From left to right: i) Input color panorama, ii) colored predicted point cloud, and iii) heatmap visualization of the $c2c$ error on the ground truth point cloud.
Figure 12. Qualitative results using the UNet model applied to in-the-wild real data captures. The top two rows are captures with a 360° camera, while the bottom two rows are stitched panoramas from a mobile phone. From left to right: i) Input color panorama, ii) predicted depth, and iii) normals derived from the predicted depth.
Figure 13. Qualitative results using in-the-wild data. On the left the input color panoramas are depicted. The two top rows are captured with a $360^\circ$ camera, while the bottom two rows are stitched panoramas from a mobile phone. The colored point clouds of the predicted depths from our UNet model (middle) and BiFuse [45] (right). Ceilings have been removed for visualization purposes.