Graphical Models '0x
Special Issue on
STAG 2024 Best papers

Volume 0 (202x), Number 0

DDD++: Exploiting Density map consistency for Deep Depth estimation in indoor environments

Giovanni Pintore[1,2], Marco Agus[3], Alberto Signoroni[4], Enrico Gobbetti[1,2]

[1]CRS4, Italy; [2]National Research Center in HPC, Big Data, and QC, Italy; [3]HBKU, Qatar; [4]University of Brescia, Italy

Abstract

We introduce a novel deep neural network designed for fast and structurally consistent monocular 360° depth estimation in indoor settings. Our model generates a spherical depth map from a single gravity-aligned or gravity-rectified equirectangular image, ensuring the predicted depth aligns with the typical depth distribution and structural features of cluttered indoor spaces, which are generally enclosed by walls, floors, and ceilings. By leveraging the distinctive vertical and horizontal patterns found in man-made indoor environments, we propose a streamlined network architecture that incorporates gravity-aligned feature flattening and specialized vision transformers. Through flattening, these transformers fully exploit the omnidirectional nature of the input without requiring patch segmentation or positional encoding. To further enhance structural consistency, we introduce a novel loss function that assesses density map consistency by projecting points from the predicted depth map onto a horizontal plane and a cylindrical proxy. This lightweight architecture requires fewer tunable parameters and computational resources than competing methods. Our comparative evaluation shows that our approach improves depth estimation accuracy while ensuring greater structural consistency compared to existing methods. For these reasons, it promises to be suitable for incorporation in real-time solutions, as well as a building block in more complex structural analysis and segmentation methods.

Keywords: Depth estimation, Spherical Images, Indoor Environments, Structural Consistency

CCS Concepts
• Computing methodologies → Computer vision; Shape inference; Neural networks;

1. Introduction

The automatic 3D modeling of indoor scenes has gained significant research attention in recent years, emerging as a well-defined 3D subfield [PMG*20]. Specialized techniques for common, highly structured environments such as residential, office, and public buildings are one of the main targets. Such buildings constitute the majority of the built environment, and 3D reality-based models are required for many purposes [IYF15, PGGS16]. Moreover, the standardized construction methods offer degrees of similarity that can be exploited for improving reconstruction from noisy and partial input [PMG*20, PAG24].

Fast depth estimation from images is a fundamental sub-problem in this context since associating metric information with visual data is necessary for 3D reconstruction, and rapid solutions open the door to many applications, including mobile extended reality, indoor mapping, and autonomous navigation. Although traditional methods have utilized the correlation among multiple views captured simultaneously (e.g., stereo) or sequentially over time (e.g., video), the interest in monocular 360° depth estimation is growing [PAG24].

A 360° image, quickly and easily acquired with affordable cameras, captures the full scene from a single viewpoint, providing rich context for depth inference and scene understanding [YJL*18]. Monocular depth estimation remains, however, very challenging in furnished indoor environments. Even though structure priors characterize the architectural shape that bounds the scene, it is often hard to recognize them, since walls are often composed of large untextured regions, and objects can be cluttered and arranged arbitrarily in the near field, masking large portions of a room's walls and floors.

These challenges have led to the introduction of indoor-specific 360° solutions that have reached impressive results, especially in conjunction with supervised deep-learning approaches that learn hidden relations from large sets of examples [PJVH*24]. Although such state-of-the-art approaches can predict high-detail depth maps with good accuracy at the pixel level, the salient features of an indoor environment, such as wall planarity and edge sharpness [PAAG21], as well as the regularity and consistency of the architectural man-made structures [SRFL21, PAA*21] are less well preserved (see Sec. 2). Such consistency becomes of critical importance when depth is used for room layout reconstruction, exploited for immersive exploration [WGSJ20, PBAG23, PJVH*24], or to merge and analyze multiple single-view reconstructions for structural segmentation of complex multi-room environments [CQF22, YKSE23]. Moreover, it is not uncommon, see Sec. 5, to require in the order of hundreds of millions of tunable parameters and over hundreds of GFLOPs to infer depth from a 512×1024 image. Such high memory and computational costs make it difficult to use them for large images or low-latency depth generation [PBAG23].

This paper proposes a lightweight end-to-end deep learning approach, dubbed *DDD++*, for depth estimation from a single 360° image in an equirectangular format. This work is a significantly extended version of extension of the paper "DDD: Deep indoor panoramic Depth estimation with Density maps consistency" [PASG24], presented at the EG STAG (Eurographics Smart Tools and Applications for Graphics) 2024 conference.

We exploit the characteristics of indoor environments to reduce computational costs and both structural consistency and depth accuracy. To design our network, see Sec. 3 and Fig. 1, we start from the assumption that, due to gravity, world-space vertical and horizontal features have different characteristics in most, if not all, man-made environments. These characteristics are preserved in gravity-aligned or gravity-rectified images [PAA*21, SSC21]. To this end, we perform a contractive encoding to reduce the input equirectangular tensor only along the vertical direction to obtain a compact and flattened sequence of slices made of a set of Gravity-Aligned Features (GAFs). To preserve global information, we perform slicing over different resolution levels, concatenating the result at the end. In addition to optimizing the flow of information contained in features, as done in previous works [PAA*21,SSC21], this representation allows in our design subsequent processing directly through a vision transformer, which takes into account the spherical nature of the input and recovers long- and short-term spatial relationships among features.

To optimize the depth map in terms of its consistency when interpreted as a sampling of a 3D architectural environment, the network is trained through a novel indoor-specific metric and loss function (see Sec. 4). We do that by transforming depths into density maps computed using planar and cylindrical projections and comparing predicted with ground truth ones. These maps, which accumulate the occurrence of 3D points derived from depths and projected onto the floorplan and on two planes orthogonal to it, are known to provide good summaries of the characteristics of indoor environments characterized by vertical walls [CQF22, YKSE23].

The overall approach, as originally presented in our conference paper, combines a network design that enables, through feature flattening, the direct use of a vision transformer, without the need to sequence the input map by arbitrary patches and positional encoding [SLL*22], with the idea to supervise training by minimizing the error on density maps, leading to a better identification and preservation of permanent indoor structures. As a result, the method's lightweight architecture has a low computational impact and provides greater structural consistency than other current approaches (see Sec. 5). Such a lean network can be integrated as a component in multi-stage pipelines, for instance, for multi-room reconstruction (e.g., [YKSE23]) or view translation and synthesis for immersive applications [PBAG23,PJVH*24]. Its fast inference time also makes it ideal for real-time usage.

In this extended paper, in addition to improving the presentation of our original idea, we introduce significant new material with respect to our original conference paper [PASG24]. In particular: we devised an improved loss function, which now includes a novel cylindrical, full-view projection to better account for the geometric information available from the spherical view respective to the original proposal of Manhattan-aligned cube maps (see Sec. 3); we discuss results of new experiments with commonly used synthetic large-scale datasets and benchmarks where ground truth is fully available [ZZL*20] and compare them with results obtained by state-of-the-art for indoor panoramic depth estimation [SZL*23, ACC*23, AW24]; we include tests on large scale real-world data, using commonly available annotated public datasets [Mat17]; we include an improved ablation and discussion, highlighting the differences between using arbitrary views as in the original paper [PASG24] and the new approach proposed here and illustrating the effects of the various design choices.

2. Related work

Depth estimation from monocular input and 3D reconstruction of indoor environments are fundamental computer vision problems, which have recently attracted renewed interest with the emergence of deep learning techniques. A full review is beyond the scope of this paper, and we refer the reader to established surveys for wider coverage [PMG*20,dSPMLJ22,PAG24]. Here, we focus on the solutions most closely related to our work.

Depth from perspective images. Data-driven monocular depth estimation was introduced over a decade ago (e.g., Make3D [SSN09]). The emergence of deep learning and the availability of large-scale 3D datasets have led to significant performance improvements. After the introduction of CNNs for regressing dense depth maps from a single image [EPF14, EF15], Laina et al. [LRB*16] introduced the now standard *FCRN* encoder-decoder architecture, combining *ResNet* [HZRS16] for the encoding and an up-projection module for decoding and the reverse Huber loss [LLZ16] to improve depth estimation. Following these trends, many solutions have been further introduced, including predicting depth from several cropped images combined in the Fourier domain [LHKK18], using an ordinal regression loss to preserve the spatial relation among neighboring classes [FGW*18], exploiting Conditional random fields (CRF) to refine predictions [LCG15, PXZ*15, CWS18, XWT*18], and many more follow-ups. However, directly applying perspective methods to 360° images, does not permit the full exploitation of their characteristics, and in particular, their global context, leading to sub-optimal results [ZSTX14, ZKZD18]. As a result, much of the research on reconstruction of indoors from sparse imagery is now focused on 360°-specific solutions.

Depth from a single omnidirectional image. Several solutions adapted perspective established methods to 360° depth prediction by using projections into a cube map [CCD*18] or by replacing regular convolutions with spherical convolutions to cope with distortions [SG17, TNT18, PdLGAAB18, ZKZD18, SG19]. Wang et

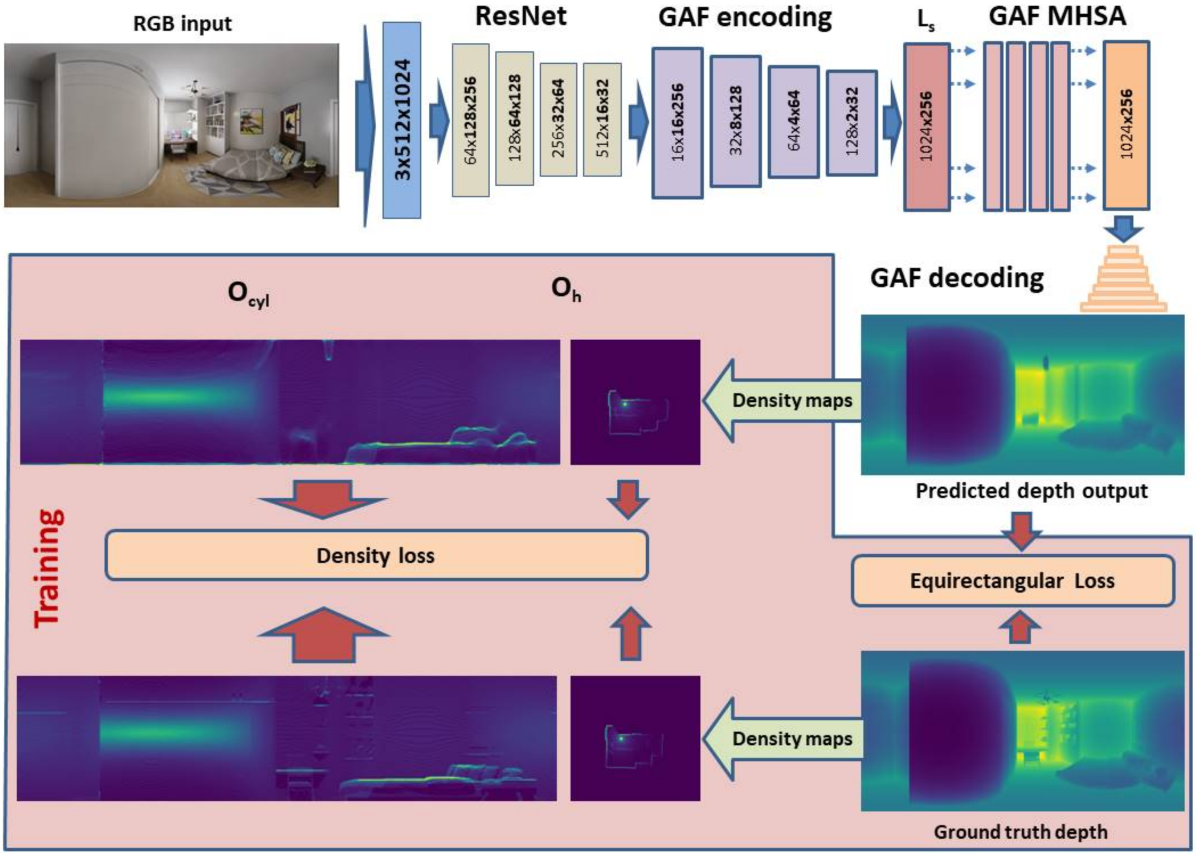


Figure 1: ***Overview***. *The network maps a gravity-aligned 360° image to its depth. The input image is first transformed by a ResNet block (light green) into feature maps with different depths and spatial sizes. Through a Gravity-Aligned Features (GAF) encoding block (light purple), we perform a gravity-aligned anisotropic contractive encoding to obtain latent features. Once assembled in a single sequence, latent features are processed by a single-layer multi-head self-attention scheme (light pink) to produce the final set of features whose decoding, through convolution and upsampling, produces the desired depth map. During training, we produce density maps respectively from predicted and ground truth depths and exploit them for structural loss computation. It should be noted that density maps are computed only at training time. Thus, the memory and computational costs do not affect network performance at inference time. Note that the depth image and density map colors have been visually enhanced for illustration purposes.*

al. [WYS*20] combined the approaches through a two-branch network, respectively for the equirectangular and the cube map projection, based on a distortion-aware encoder [ZKZD18] and the FCRN decoder [LRB*16]. Several recent methods leverage perspective views sampled on panoramic images [LGY*22, RAYR22] before combining depth maps using patch-based vision transformers [SZL*23, ACC*23, AW24]. Building on this trend, Wang et al. [WL24] propose Depth Anywhere, a framework that enhances 360° monocular depth estimation by distilling knowledge from a pretrained perspective depth estimator into a panoramic model. Complementary to this, Cao et al. [CAVW24] address the challenge of high-resolution 360° depth estimation without high-res depth ground truth by introducing a weakly-supervised framework that transfers structural knowledge through a Scene Structural Knowledge Transfer (SSKT) module. Another breed of solutions for panoramic depth estimation in indoor spaces [SSC21, PAA*21] proposes, instead, to work directly on the equirectangular images produced by spherical cameras and to leverage the concept of gravity-aligned features to reduce network size while supporting the exploitation of short- and long-range relations. In this work, we show how to directly use gravity-aligned features [PAA*21] to feed a self-attention vision transformer, without the need to arbitrarily partition the image into patches [SZL*23, ACC*23, AW24]. Moreover, we incorporate the concept of density maps, as used in reconstruction and segmentation tasks [CQF22, YKSE23], into 360° depth prediction to define a structural loss that enhances the accuracy and consistency of depth predictions with architectural structures in indoor models. Improving over our original proposal [PASG24], we exploit here both a cylindrical and a planar projection to recover structural information. Moreover, As a result, we achieve state-of-the-art performance at a lower inference cost than previous solutions.

3. Network architecture

Our network takes as input a 360° gravity-aligned image in equirectangular format and produces as output its per-pixel depth. Assuming gravity-alignment allows us to design a particularly efficient solution, while not limiting the domain of application of the method. Gravity-aligned capture is very common, and nearly all public 3D indoor datasets commonly used for training and testing reconstruction solutions exhibit minimal orientation deviations. [PAA*21, SSC21]. This is because maintaining the upright position for capturing, besides being natural for free-form single-shot images, is usually enforced by exploiting data from the IMUs present in most modern capture devices or by mechanical setups such as tripods. Moreover, even in the few cases where these assumptions are not verified at capture time, many orthogonal and fast solutions can be applied to gravity-rectify images in a preprocessing step to connect the direct output from the capture device to our depth estimation network (e.g., [XLF*19, JLAB19, DAH20]).

Our lightweight network architecture for depth estimation in indoor environments combines gravity-aligned features obtained by asymmetric convolution of the input with multi-head self-attention. The structure of our network is depicted in Fig. 1.

From the input image, a cascade of five residual layers [HZRS16] returns four feature maps having different depths and spatial sizes. Given the spherical nature of the image, we also adopt circular padding along the horizon for convolutions, to overcome the longitudinal boundary discontinuity, and reflection padding to alleviate the singularities at the poles [GSZ*21].

To support an efficient gathering of information from the extracted features, we perform a specifically indoor-designed feature compression exploiting our knowledge of preferential directions, based on the fact that gravity-aligned images preserve the fact that world-space vertical and horizontal features have different characteristics in most, if not all, man-made environments [SSC21, SHSC19, PAA*21, PAAG21]. For instance, it is fairly natural, if only for physical reasons, to have horizontal planes both in architectural (e.g., floors) and impermanent (e.g., tabletops) structures, as well as vertical ones (e.g., walls and supporting parts of furniture). Exploiting this assumption, we perform an *anisotropic contractive encoding* that reduces the vertical direction while keeping the horizontal direction unchanged, so that separated vertical features can be better preserved. Specifically, we reduce the vertical dimension by a factor of 8 through an asymmetric convolution module with stride $(2,1)$, applied three times, that contains a 2D convolution and an ELU module. We apply such compression for each encoded feature map (i.e., four maps), obtaining a set of latent features $L_s = (l_1 \dots l_4)$. Compressed features L_s are reshaped to the same size and joined in a flattened latent feature, as a single sequence of s feature vectors of dimension l (i.e., s horizontal size of the less deep feature map - $s = 1024$ and $l = 256$ for a 512×1024 input). Such a compressed representation contains a wealth of information about the scene's local and global geometry, which can be exploited to recover depth and layout and provide a latent representation of the scene.

Note that our flattening of gravity-aligned features constructs a structured linear sequence that can be directly used as input to a self-attention-based vision transformer. This design bypasses the need for arbitrary image patching or complex positional encodings, as commonly required in transformer-based architectures [SZL*23, ACC*23]. By leveraging the inherent alignment with the gravity direction, our representation preserves the spatial coherence of the scene, particularly the vertical semantic structure typical of indoor environments. To exploit long-range dependencies in this structured representation, we integrate a single-layer multi-head self-attention (MHSA) module [VSP*17]. This module effectively captures contextual relationships across distant regions of the image—an especially valuable property for omnidirectional imagery, where objects and structural cues may span wide angular fields.

Crucially, our approach avoids the need for recurrent modules, which are commonly used in prior GAF-based methods such as SliceNet [PAA*21] to model sequential dependencies. While effective, recurrent neural networks (RNNs) introduce significant computational overhead and latency due to their inherently sequential nature.

In contrast, our MHSA implementation processes all tokens in parallel and maintains a much lower computational footprint, enabling faster training and inference without sacrificing the ability to reason over global context. This design choice not only improves scalability but also enhances the model's capacity to integrate complementary information from spatially distant but semantically related regions, making it particularly well-suited for the rich geometric structure captured in panoramic indoor scenes.

Our self-attention module takes the latent features $L \in \mathbb{R}^{s \times l}$ as input, and outputs a self-attention weight matrix $A \in \mathbb{R}^{s \times s}$:

$$A = softmax\left(\frac{(LW_q)(LW_k)^T}{\sqrt{l}}\right) \quad (1)$$

where $W_q, W_k \in \mathbb{R}^{l \times l}$ are learnable weights. The MHSA module has a particularly lightweight design with four heads and only one inner layer. We have verified experimentally that increasing the number of layers and heads heavily increases the number of parameters and computational load without significantly improving reconstruction accuracy. Once passed to the MHSA module, the decoding of the latent feature $(1 \times 1 \times s)$ is very fast, through convolutions, upsampling modules, and ELU activations, until we reach the target output resolution $(1 \times h \times w)$.

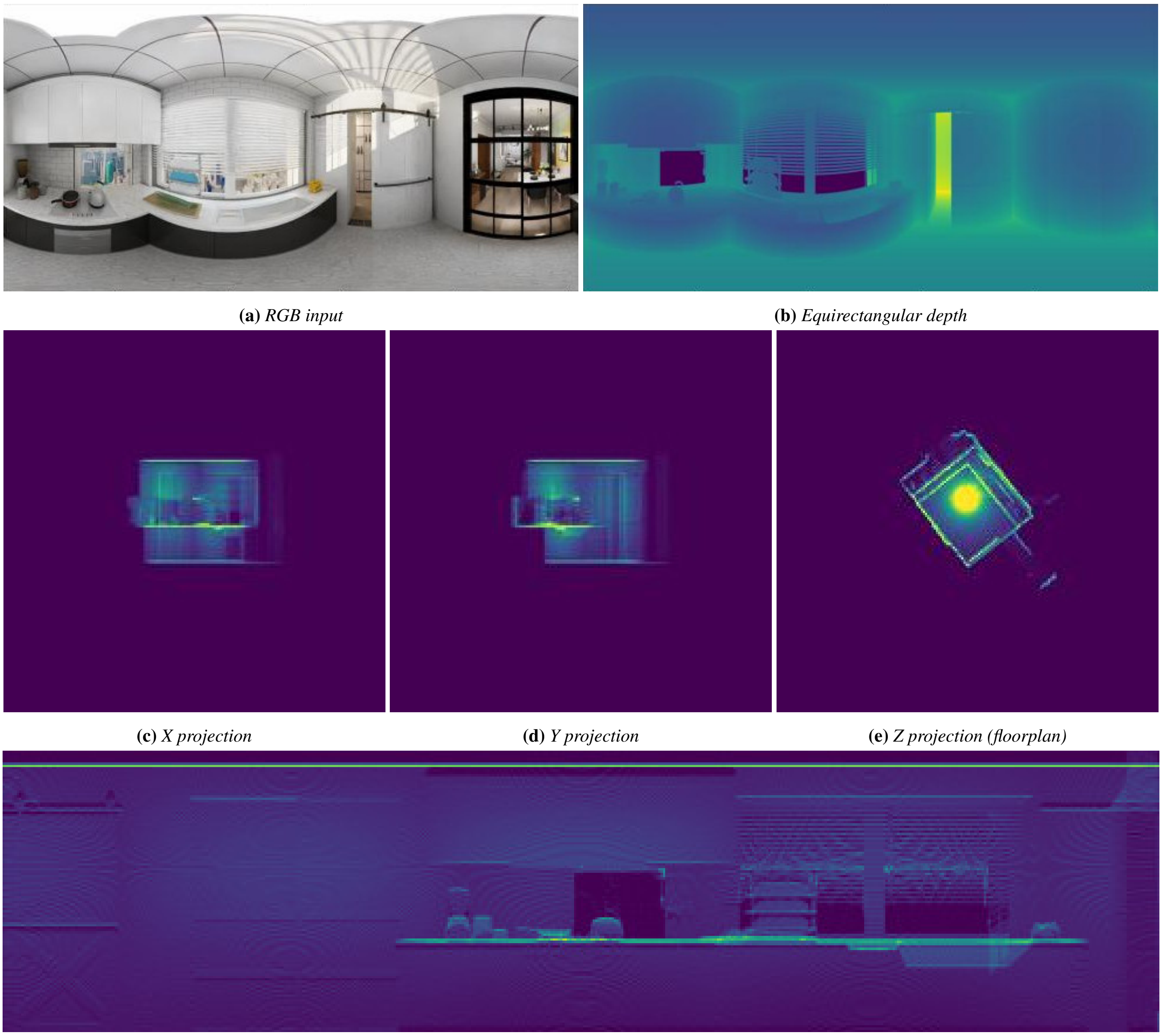
4. Indoor-specific loss function and training strategy

To train our network, we designed a loss function that is a combination of a conventional equirectangular loss term ($\mathcal{L}_{eq}$) with a novel, structure-driven component ($\mathcal{L}_{ds}$), i.e., $\mathcal{L} = \mathcal{L}_{eq} + \mathcal{L}_{ds}$.

The equirectangular loss term $\mathcal{L}_{eq}$ penalizes per-pixel deviations of the inferred depth from the ground truth value. As common for depth estimation frameworks, we build it on top of the robust *Adaptive Reverse Huber Loss (BerHu)* [LLZ16]:

$$H(e) = \begin{cases} |e| & |e| \leq c \\ \frac{e^2+c^2}{2c} & |e| > c \end{cases} \quad (2)$$

where e is the error term and the parameter c determines where to switch from L1 to L2. To set the c value adaptively, we follow



(a) *RGB input*

(b) *Equirectangular depth*

(c) *X projection*

(d) *Y projection*

(e) *Z projection (floorplan)*

(f) *Cylindrical projection*

Figure 2: ***Different density maps examples***. *In the first row we show the input image Fig. 2a and the equirectangular depth Fig. 2b. In the second row we show examples of density maps recovered from the equirectangular depth. In this case we illustrate the density maps adopted by the DDD baseline [PASG24]: Fig. 2c and Fig. 2d, respectively projections along the X and Y axes; Fig. 2e along Z axis, that is the density map projected on the floorplan (i.e., vertical projection). In this paper, we keep Fig. 2e, which is highly representative of the room shape, and we replace the arbitrary projections Fig. 2c and Fig. 2d with a more general and comprehensive cylindrical projection Fig. 2f, which cover the whole horizon (i.e., horizontal projection). As also evident visually, this projection best captures seamless structural details.*

the approach originally introduced by Laina et al. [LRB*16], so that c is set, in every gradient step, to 20% of the maximal error of the current batch. When applied to the depth maps, we have $e = D_{ij} - D^*_{ij}$ at each pixel (i, j), where D and D^* are, respectively, the predicted and the ground-truth depth maps, and, thus:

$$\mathcal{L}_{eq}(D, D^*) = \sum_{ij} H(D_{ij} - D^*_{ij}) \tag{3}$$

Using only this term, however, that measures, per-pixel, distances from training data, would not take into account the peculiar features of indoor environments, and especially of the architectural structures, that we expect made of large fairly regular surfaces with preferential orientations. For instance, we expect to find mostly horizontal floors and mostly vertical walls, rather than curved/wobbly surfaces, that can, instead, more commonly be found on objects.

To drive the solutions toward plausible depth reconstructions, we introduce in this work a structural term $\mathcal{L}_{ds}$, rather than using a regularization term. Using such an approach allows us to learn these regularities from data, rather than imposing them upfront through specific penalty functions.

In our original conference contribution [PASG24], we have shown how density maps computed from point clouds can provide an adequate structural summary to be exploited for loss computation. The basic idea was to extract important features through projections on horizontal and vertical directions. This was done, however, by imposing an arbitrarily oriented Manhattan world prior (i.e., Fig. 2c and Fig. 2d), introducing unwanted discontinuities and wide variations in the summary characteristics, for the same environment, depending on the relative alignments of the arbitrary Manhattan axes with the main equirectangular view direction. Here, we significantly improve $\mathcal{L}_{ds}$ by introducing a loss based on a single density map computed by vertically projecting the horizontal planes and a single density map computed by horizontally projecting on a cylinder (i.e., Fig. 2f), thereby encompassing the entire visible horizon and seamlessly capturing better and more stable structural features.

To compute the $\mathcal{L}_{ds}$ structural term, we transform D and D^* into the equivalent point clouds P_D and P_D^* in Cartesian coordinates using the spherical transformation associated to the equirectangular projection. We then scale 3D points to the same absolute scale, by setting a maximum distance from the observer (20 meters in the examples presented in this work). Assuming the gravity-vertical direction as the Z axis of our reference system, we produce predicted and ground truth density maps from P_D^* and P_D with respect to vertical and horizontal direction. Specifically, we render two density maps, Oz (i.e., vertical projection looking to the floor) and Oc (i.e., horizontal projection, looking to the walls), respectively from the from depth prediction and from the ground truth depth-point cloud. Since O represents a map of the occurrences of 3D points falling on the same pixel, the structural parts of the scenes become more evident. For instance, the vertical projection along Oz highlights the floor plan (e.g., Fig. 2e), since the many vertically aligned points on walls in ground truth data identify room boundary locations. For this reason, such a projection is often used to automatically derive the floor plan of one or more rooms from a point cloud [CLWF19], but, to the best of our knowledge, has not been used to define indoor-specific cost functions for depth recovery. At the same time, the horizontal projection Oc (e.g., Fig. 2f) emphasizes the shapes of horizontal planes in the scene, as well as geometric patterns distributed along the horizon.

Given a 3D point cloud $P_D = \{\mathbf{p}_i\}_{i=1}^N$, where each point $\mathbf{p}_i = (x_i, y_i, z_i)$ represents the Cartesian coordinates of the i-th point, the vertically projected density map Oz, along Z axis, is easily obtained just removing z_i and counting the occurrences of two points (x_i, y_i) on the floorplan, after normalization and rescaling to the density map size (512×512 in our experiments). For computing the horizontally projected density map, we transform 3D points into cylindrical coordinates.

Specifically, assuming the cylindrical occupancy map $O_c \in \mathbb{R}^{h \times w}$ (256×1024 in our experiments), for each point $\mathbf{p}_i = (x_i, y_i, z_i)$, we compute the azimuthal angle θ_i in the xy-plane as:

$$\theta_i = \operatorname{atan2}(y_i, x_i).$$

The angle θ_i is normalized to the range $[0, 2\pi)$ using:

$$\theta_i = \begin{cases} \theta_i + 2\pi & \text{if } \theta_i < 0, \\ \theta_i & \text{otherwise.} \end{cases}$$

Given the minimum and maximum z-values in the point cloud as:

$$z_{\min} = \min_i z_i, \quad z_{\max} = \max_i z_i.$$

The z-coordinate is normalized to the range $[0, 1]$ using:

$$z_i' = \frac{z_i - z_{\min}}{z_{\max} - z_{\min}}.$$

The normalized cylindrical coordinates (θ_i, z_i') are then discretized into indices (j, k) for the occupancy map:

$$j = \left\lfloor \frac{\theta_i}{2\pi} \cdot w \right\rfloor, \quad k = \lfloor z_i' \cdot h \rfloor.$$

The indices are clamped to ensure they lie within the valid range:

$$j = \operatorname{clamp}(j, 0, w-1), \quad k = \operatorname{clamp}(k, 0, h-1).$$

The occupancy map O_c is initialized as a zero matrix of size $h \times w$. For each point $\mathbf{p}_i$, the corresponding bin in O_c is incremented:

$$O_c(k, j) \leftarrow O_c(k, j) + 1.$$

The resulting cylindrical occupancy map O_c is a 2D grid where each cell (j, k) represents the number of points projected onto the corresponding cylindrical bin. The z-axis is oriented with the origin at the bottom.

Starting from the predicted and ground-truth density maps, respectively (Oz, Oc) and (Oz^*, Oc^*), we calculate the structural loss term $\mathcal{L}_{ds}$ as the sum of the adaptive Reverse Huber loss of the individual predicted density map value relative to ground truth for each pixel (k, l) in the projections:

$$\mathcal{L}_{ds}(Oz, Oc, Oz^*, Oc^*) = \sum_{kl} \mathcal{H}(Oz_{kl} - Oz^*_{kl}) + \sum_{kl} \mathcal{H}(Oc_{kl} - Oc^*_{kl}) \tag{4}$$

The same parameters used for tuning Equation 2 for depth values are used for the density maps. In Sec. 5, we show how we achieve good performance using only these data terms even without adding other regularization terms.

It should be noted that our approach does not require strict alignment of the panorama and layout to the Manhattan World axes but only needs the more common gravity alignment (see Sec. 2). This allows us to limit geometric data augmentation to flips and random rotations around the Z axis during training. Furthermore, our geometric augmentation accounts for the fact that our density maps are mutually orthogonal and gravity-aligned, but arbitrarily rotated around the gravity vector. The augmentation through random rotations helps uncover hidden relationships that are independent of the view's alignment with world-space axes.

5. Results

Our approach was implemented using *PyTorch* and has been tested on several kinds of indoor scenes. In the following, we discuss the datasets used in this work (Sec. 5.1). We then briefly illustrate the training setup and the computational performance, also comparing inference times and costs to other state-of-the-art solutions (Sec. 5.2). Finally, we discuss the quantitative and qualitative results on depth reconstruction (Sec. 5.3), compared to the state-of-the-art.

5.1. Datasets

In this article, we significantly extended the experiments and comparisons from the original conference work [PASG24]. Following the latest state-of-the-art works [YSL*23, AW24], we adopt for training and testing Structured3D [ZZL*20]), a large-scale synthetic database of indoor scenes comprising 21,000 photorealistic scenes, which provides ground truth depth and layout information for each panoramic image. This benchmark provides full spherical coverage and provides very realistic, but artifact-free ground-truth color and depth information. This allows for a device-independent, reliable, and consistent evaluation of equirectangular depth estimation methods, free from the stitching or occlusion-related inconsistencies present in other datasets. For these reasons, Structured3D [ZZL*20]) is the main benchmark used in this work.

In addition, to illustrate performance on real-world capture data, we include results obtained on Matterport3D [Mat17]. Although Matterport3D [Mat17] is widely adopted as a benchmark for equirectangular depth estimation, it is important to note that the dataset was originally captured in perspective format using multiple RGB-D sensors. As such, transforming the data into a consistent equirectangular representation requires a non-trivial preprocessing pipeline involving stitching and blending of multiple perspective views. This preprocessing step introduces variability in the resulting images and depth maps, which in turn affects the comparability of results reported across different studies. Consequently, significant discrepancies can be observed among Matterport3D results in the literature, depending on the specific processing pipeline adopted. In this work, to ensure a fair and consistent comparison, we adhere to the preprocessing procedure introduced by Elite360D [AW24]. We also report their published results as a reference for all comparisons on this dataset.

Furthermore, to provide direct comparisons to the original paper baseline [PASG24] and to support the ablation study, we also discuss results obtained with the publicly available Shanghaitech-Kujiale Indoor 360° (SKI360) dataset [SK20]. The dataset contains 1,775 panoramic RGB images of scenes of furnished rooms accompanied by ground truth depth maps. The images are synthesized from 3D models with a photorealistic renderer based on path tracing to achieve realistic rendering [JXZ*20]. As in our previous work [PASG24], this benchmark is also used to compare our performance relative to other solutions exploiting geometric cues as priors and regularizers [JXZ*20].

5.2. Training setup and computational performance

We trained our DDD++ network using a single NVIDIA RTX 4090 GPU equipped with 24GB of VRAM. The training process employed the Adam optimizer with default momentum parameters, specifically $\beta_1 = 0.9$ and $\beta_2 = 0.999$, which are well-suited for stabilizing convergence in dense prediction tasks. We initialized the learning rate at 1×10^{-4} and employed an adaptive learning rate schedule that decayed based on validation loss plateaus, promoting efficient convergence without manual tuning. Training was performed with a batch size of 16, chosen to maximize GPU utilization while maintaining sufficient stability in gradient updates.

All experiments were conducted at a native resolution of 512×1024, preserving the full equirectangular spatial structure of the input panoramas. This resolution is the native one for Structured3D [ZZL*20]). Under these conditions, the average training time per image was approximately 32 milliseconds, enabling efficient iteration over large datasets. At inference time, our model achieves a latency of just 7 milliseconds per image on the same RTX 4090, making it well-suited for real-time or near-real-time applications.

During training, invalid depth values—such as points corresponding to views through windows or regions missing in the ground truth—are masked out and excluded from the loss computation. Additionally, these regions are set to zero in the output to prevent the network from hallucinating unreliable depth information.

Method	**Parameters↓**	**FLOPs↓**	**Inf. time↓**
Bifuse [WYS*20]	253 M	682 G	144 ms
SliceNet [PAA*21]	79 M	101 G	21 ms
Panoformer [SLL*22]	20 M	78 G	17 ms
EGFormer [YSL*23]	**15 M**	74 G	16 ms
Elite360D [AW24]	25 M	65 G	14 ms
DDD++ (our)	23 M	**38 G**	**7 ms**

Table 1: ***Computational performance of inference.*** *We show our computational performance compared to other state-of-the-art works for a* 512×1024 *image. We also show an example of inference time on a NVIDIA RTX 4090.*

Tab. 1 presents the computational performance of inference with our network. We compare it to major state-of-the-art depth estimation solutions for 360° indoor imagery, for which performance is reported in the original publication or the code is available for testing. As we can see, our approach has, by far, the lowest computational complexity (FLOPs) of the compared methods (see Sec. 5.3). We also show, for a more intuitive comparison, the average inference time for all methods on an NVIDIA RTX 4090 (24GB VRAM). Our computational cost is, in particular, less than half of the currently fastest method (Elite360D [AW24]).

The number of parameters is also in the ballpark of recent solutions based on vision transforms (Panoformer [SLL*22] and EGFormer [YSL*23]) and much less than prior solutions (SliceNet [PAA*21] and Bifuse [WYS*20]). Our method's reduced cost and footprint make it possible to scale our solution to larger image sizes than competitors when suitable higher-resolution training data will be available.

5.3. Evaluation and comparison with the state-of-the-art

DDD++ is a simple, lightweight architecture that quickly produces depth estimation and is trained by a simple loss function exploiting density maps. Since the density-map-based loss function is one of our main contributions, we strive to demonstrate that good results can be achieved at a low computational cost and without introducing other loss terms, based, e.g., on smoothness, planarity, verticality of walls, or other specific geometrical and architectural priors [RSL*24].

Tab. 2 presents a comparative evaluation of depth estimation performance across state-of-the-art methods on both real-world (Matterport3D [Mat17]) and synthetic (Structured3D [ZZL*20]) datasets. To reduce problems stemming from variations in architecture complexity, hyperparameters, and validation procedures used across different methods, we refer to the latest state-of-the-art work in this context, Elite360D [AW24], which performed retraining and evaluation under the same conditions as Jiang et al. [JSZ*21]. The error metrics used, where lower values indicate better performance (↓), include **Absolute relative error (Abs Rel)**, **Squared relative error (Sq Rel)**, and **Root mean squared error (RMSE)**. Additionally, the evaluation utilizes three threshold percentages, denoted as $\delta_1(\%)$, $\delta_2(\%)$, and $\delta_3(\%)$. These measure the percentage of pixels where the depth prediction error is less than a threshold defined as $\delta < \alpha^t$, where $\alpha = 1.25$ and $t = 1,2,$ or 3. For these threshold metrics, a higher value indicates better performance (↑).

Our method, DDD++, achieves the best performance on the synthetic Structured3D dataset across all metrics. This result is particularly significant as Structured3D provides complete and clean equirectangular views along with full ground-truth depth maps, ensuring the comprehensive geometric context that makes it possible to fully exploit our proposed occupancy-aware loss.

Conversely, on the real-world Matterport3D dataset, while DDD++ performs comparably with other top-performing methods, it does not achieve the best scores, especially when compared with solutions with a much increased computational cost. In addition to computational considerations (see Tab. 1), this can be attributed to two key factors. First, the equirectangular views in Matterport3D are incomplete and obtained through stitching, which leads to missing or distorted geometric information. Second, many of the competing approaches incorporate additional supervision signals – such as gradient consistency, surface normals, or structural priors – that, in addition to helping to compensate for the incomplete view, provide a better resistance to noise and increased detail preservation. These auxiliary loss terms are particularly effective in noisy real-world scenarios where geometric context is partially missing. We expect that, by incorporating some of these additional terms in our loss function, an even better solution could be achieved.

It is important to note that in Structured3D, such auxiliary losses play a more marginal role due to the dataset's completeness and reduced noise. This underscores the robustness and potential of our proposed loss, which directly exploits full geometric cues when available. Moreover, we emphasize that our novel loss formulation is agnostic to the network architecture and can be seamlessly integrated with other models and additional loss terms. We leave this broader integration and ablation analysis for future work.

Another important aspect highlighted in Tab. 2 is the impact of network backbone design on performance. While methods such as SliceNet employ structural priors to achieve strong results (particularly on the Structured3D dataset), these approaches often rely on heavier architectures like ResNet-50 combined with recurrent modules (RNN), resulting in high computational costs (see Tab. 1). In contrast, our DDD++ model leverages a lighter ResNet-18 backbone augmented with multi-head self-attention (MHSA), striking a better balance between efficiency and performance. Notably, the structural consistency enforced by our proposed loss term on density maps offers a compelling alternative to architectural complexity. Unlike general depth estimation methods such as Elite360D [AW24], which primarily optimize per-pixel depth accuracy, our formulation encourages data-driven regularization that captures broader structural patterns. We believe this performance gain stems from the synergy between our network design and the proposed loss, which jointly enable the model to better exploit medium- and large-scale regularities characteristic of indoor environments. This suggests that enhancing depth estimation through density-based geometric reasoning can be an effective and scalable strategy, especially in domains where structural coherence plays a central role.

Fig. 3 presents qualitative examples of our method's predictions, showcasing both the estimated depth maps and the corresponding point clouds. These visualizations highlight the effectiveness of our approach in preserving fine architectural details. Notably, our method is capable of maintaining sharp geometric features, such as edges and corners, as well as ensuring the smoothness of continuous surfaces. This is achieved without relying on any explicit regularization terms or post-processing heuristics, underscoring the intrinsic strength of our formulation. Furthermore, Fig. 4 and Fig. 5 provide a comparative analysis on the Matterport3D [Mat17] and Structured3D [ZZL*20] datasets, respectively. In these figures, we compare our predictions with those of SliceNet [PAA*21], an established method that, like ours, exploits gravity alignment to streamline the network design, but uses a loss function that measures errors on data and gradient components.

5.4. Comparison with structured-guided baselines and ablation

To demonstrate the improvements of *DDD++* compared with other structure-guided methods, including the *DDD* baseline [PASG24], we present results on SKI360 dataset [SK20], which is the benchmark adopted by DDD work. With this benchmark, specific results with another structure-guided method (i.e., Jin et al. [JXZ*20]) are available. The framework introduced by Jin et al. [JXZ*20] represents a notable state-of-the-art pipeline for indoor scene understanding, jointly predicting per-pixel depth and room layout. The

Table 2: ***Depth estimation performance compared to SoA works.*** *We present results and comparisons on both real-world and synthetic datasets. For comparison, we follow the same setup of the recent publication Elite360D [AW24], which provides exhaustive results about the latest state-of-the-art approaches. To support further evaluations, we train, under the same conditions [AW24],* SliceNet, DDD, *and* DDD++.

Datasets	Backbone	Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1(%) ↑	δ_2(%) ↑	δ_3(%) ↑
Matterport3D [Mat17]	Transformer	EGFormer [YSL*23]	0.1473	0.1517	0.6025	81.58	93.90	97.35
		PanoFormer [SLL*22]	0.1051	0.0966	0.4929	89.08	96.23	98.31
	ResNet-18	UniFuse [JSZ*21]	0.1191	0.1030	0.5158	86.04	95.84	98.30
		Elite360D [AW24]	0.1272	0.1070	0.5270	85.28	95.28	98.49
	ResNet-34	BiFuse [WYS*20]	0.1126	0.0992	0.5027	88.00	96.13	98.47
		UniFuse [JSZ*21]	0.1144	0.0936	0.4835	87.85	96.59	98.73
		Elite360D [AW24]	0.1115	0.0914	0.4875	88.15	96.46	98.74
	ResNet-50	UniFuse [JSZ*21]	0.1185	0.0984	0.5024	86.66	96.18	98.50
		Elite360D [AW24]	0.1112	0.0980	0.4870	86.70	96.01	98.61
	ResNet-18+MHSA	DDD [PASG24]	0.1457	0.1522	0.6550	81.48	91.89	96.29
		DDD++(Ours)	0.1130	0.1026	0.5054	84.28	93.57	97.51
Structured3D [ZZL*20]	Transformer	EGFormer [YSL*23]	0.2205	0.4509	0.6841	79.79	90.71	94.55
		PanoFormer [SLL*22]	0.2549	0.4949	0.7937	74.70	89.15	93.97
	ResNet-34	BiFuse [WYS*20]	0.1573	0.2455	0.5213	85.91	94.00	96.72
		UniFuse [JSZ*21]	0.1506	0.2319	0.5016	85.42	93.99	96.76
		Elite360D [AW24]	0.1480	0.2215	0.4961	87.41	94.34	96.66
	ResNet-50+RNN	SliceNet [PAA*21]	0.1225	0.2214	0.5024	90.82	94.54	95.16
	ResNet-18+MHSA	DDD [PASG24]	0.0607	0.1128	0.1594	96.14	98.53	99.22
		DDD++(Ours)	0.0504	0.1092	0.1483	97.18	98.92	99.40

layout is expressed as a structural representation consisting of corners, boundaries, and planes, and serves as a strong geometric prior. The correlation between depth and layout introduces a robust form of structural consistency, which the method exploits through geometric structure-aware and regularized depth estimation.

To ensure a fair comparison, we used the same settings adopted by Jin et al. [JXZ*20], to train, test, and validate our proposed models, the original *DDD* [PASG24] model and *DDD++*, as well as SliceNet [PAA*21]. For Jin et al. [JXZ*20], we report their official depth estimation results on the same dataset. It is important to note that, due to the high computational cost of their model, the results by Jin et al. are reported at a downsampled resolution of 256 × 512, while our method runs at the resolution of 512 × 1024. This reduced resolution may lead to slightly overestimated performance metrics for Jin et al. [JXZ*20], as finer structural details are lost and prediction becomes less sensitive to high-frequency errors.

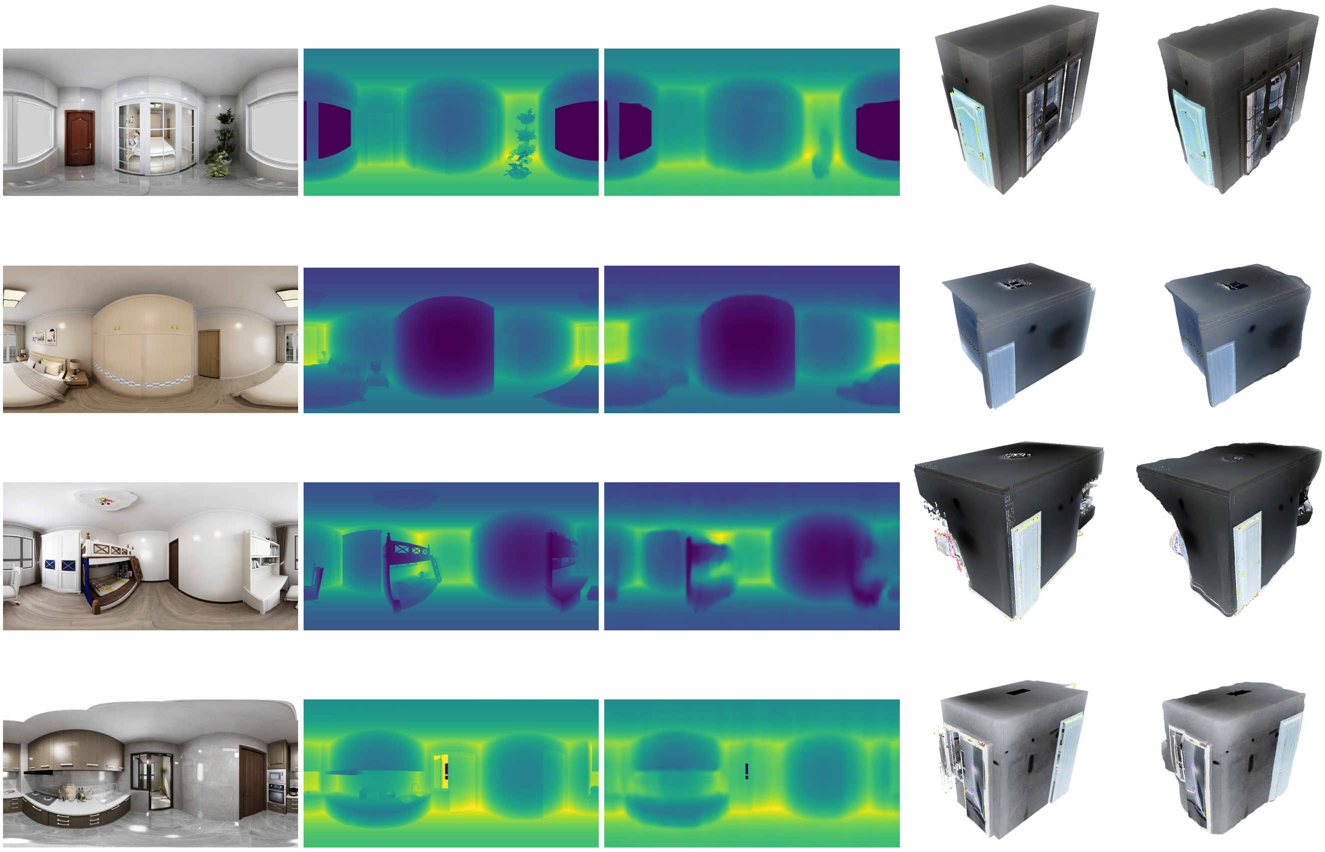
Tab. 3 summarizes this comparison. The first two rows show the results from Jin et al.'s network, where ("with SC") and ("no SC") indicate the presence or absence of structural consistency enforced by layout priors. The last three rows present the results obtained with SliceNet [PAA*21], DDD [PASG24], and our proposed DDD++ model.

All three of these latter methods share the same underlying principle of using gravity-aligned features (GAF) to better represent the vertical structure of indoor environments. However, they differ significantly in architectural choices and computational efficiency. SliceNet employs a ResNet-50 backbone combined with a recurrent neural network (RNN), which, while effective, is computationally expensive and less suitable for real-time applications. In contrast, both DDD and DDD++ adopt a more lightweight and efficient architecture, based on ResNet-18 augmented with a transformer-based self-attention mechanism. The primary difference between DDD and DDD++ lies in the introduction of a novel loss function in DDD++, which leverages structural priors in the form of density maps, further enhancing accuracy without additional computational cost. The improved performance in DDD++ demonstrates the advantage of the new cylindrical and planar projections.

Method	**MRE**↓	**RMSE**↓	δ_1(%) ↑
Jin [JXZ*20] no SC	0.114	0.721	89.4
Jin [JXZ*20] with SC	0.103	0.666	91.0
SliceNet [PAA*21]	0.102	0.273	90.4
DDD [PASG24]	0.063	0.254	91.9
DDD++	**0.050**	**0.242**	**92.8**

Table 3: ***Comparison with structure-guided baselines.*** *The first two rows report the results obtained with the network of Jin et al. [JXZ*20] without (no SC) and with (with SC) the inclusion of geometric priors and regularizers. The third row reports the results obtained by SliceNet [PAA*21], which uses GAFs but no structural consistency terms. The last two rows report the results obtained with the DDD baseline [PASG24] and with DDD++.*

Tab. 4 presents an ablation study aimed at evaluating the impact of each component in our proposed architecture. The bottom row corresponds to the full *DDD++* configuration, which integrates all key contributions: gravity-aligned features (GAF) for encoding structural priors, the lightweight multi-head self-attention (MHSA) module for capturing long-range contextual dependencies, and our novel cylindrical density loss term. This full-featured setting achieves the best overall performance across all metrics, confirming the effectiveness of our full design.



(a) *RGB input* **(b)** *Ground truth depth* **(c)** *Our predicted depth* **(d)** *Ground truth PC* **(e)** *Our predicted PC*

Figure 3: ***Qualitative performance examples.*** *We show our reconstructions and ground-truth models as depth maps and as the associated point clouds from the SKI360 dataset [SK20]. Thanks to structural consistency and without specific post-processing, the method effectively preserves architectural details, such as sharp edges and smooth planes.*

Moving upward, the row labeled *DDD [PASG24]* uses the earlier version of the loss term (basic DL), based on planar density projection, while still exploiting GAF and MHSA. Despite being competitive, it underperforms compared to *DDD++*, highlighting the contribution of the improved cylindrical density loss (cyl DL).

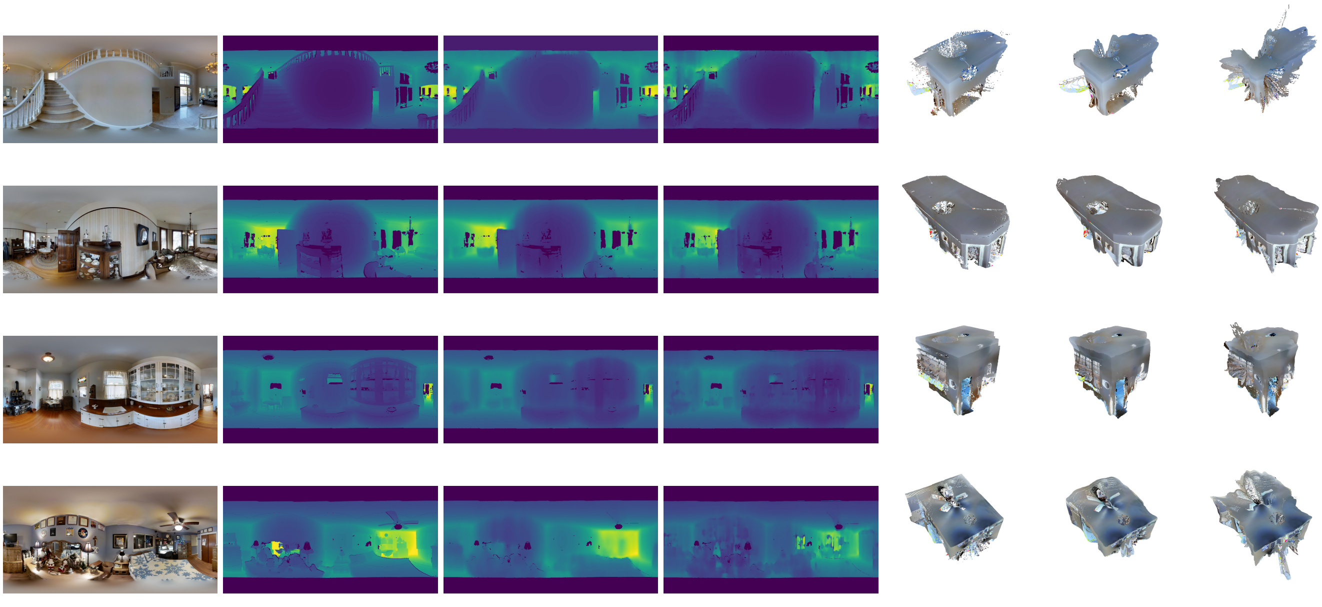
The upper three rows further dissect the contribution of the architectural components. Specifically, we progressively disable the density loss term (no DL), the MHSA (direct decoding), and GAF, observing significant degradation in accuracy, particularly in the configuration that lacks both modules (top row), which corresponds to a plain CNN without structural priors or attention mechanisms. This baseline suffers from poor geometric understanding and achieves the lowest accuracy, demonstrating the importance of each proposed design choice.

As additional experiments, Fig. 6 provides a qualitative comparison of the reconstructions obtained by our method to both ground truth and the SliceNet approach [PAA*21].

Baseline	**Loss**	**GAF**	**MHSA**	**MRE↓**	**RMSE↓**	**δ_1 ↑**
DDD cnn	✗	✗	✗	0.382	0.627	75.7
DDD direct	✗	✓	✗	0.162	0.312	88.9
DDD no DL	✗	✓	✓	0.075	0.278	90.7
DDD [PASG24]	basic DL	✓	✓	0.063	0.254	91.9
DDD++	**cyl DL**	✓	✓	**0.050**	**0.242**	**92.8**

Table 4: ***Ablation study.*** *The full model (DDD++, bottom row) combines gravity-aligned features (GAF), multi-head self-attention (MHSA), and cylindrical density loss (cyl DL), achieving top performance. Removing cyl DL (DDD row) causes a measurable drop, validating its improvement over planar density loss. Progressively disabling MHSA and GAF (upper rows) further degrades accuracy, with the plain CNN baseline (top row, no GAF/MHSA/DL) performing worst.*

In general, these experiments demonstrate how including structural consistency terms strongly benefits depth estimation, since both Jin et al. [JXZ*20] and DDD improve the most important metrics when incorporating their structural terms, independently from the different paths taken to design the networks. Moreover, our



(a) *RGB input* **(b)** *Ground truth depth* **(c)** *Our pred. depth* **(d)** *SliceNet pred. depth* **(e)** *Ground truth PC* **(f)** *Our PC* **(g)** *SliceNet PC*

Figure 4: ***Qualitative performance and comparisons on Matterport3D dataset [Mat17].*** *We illustrate our qualitative performance compared to the competitor [PAA*21] on real-world scenes. We show reconstructions and ground-truth models as depth maps and as canonical views of the associated point clouds.*

method, when including the structural consistency terms, achieves state-of-the-art performance despite the much lower computational burden compared to the other baselines. The significant improvement achieved when using the cylindrical loss introduced in this article over the previous one exploiting projections on Manhattan planes is due to its increased continuity along the horizontal direction and its reduced dependency on the relative alignment between 3D features and Manhattan planes.

The results in Tab. 3 show also that our method also provides increased performance when compared to a reference method that enforces a stronger architectural layout structure, such as Jin et al. [JXZ*20], which focuses on polyhedral rooms and was specifically designed using the dataset employed in this specific benchmark. Although further analysis is required to draw definitive conclusions, we hypothesize that for pure depth estimation, relying solely on density map similarities – rather than using corners, boundaries, and planes as priors and regularizers – makes our approach more robust to variations in the actual layout compared to the imposed prior model. Moreover, our depth inference solution is much leaner, since the complexity of generating and evaluating density maps is relevant only to the training phase.

6. Failure cases

Our network produces a pixel-wise depth map, whose accuracy depends on how closely the scene adheres to the indoor structural assumptions it was trained on. Learning this structure from data, as done with our density-map-based approach, makes the network less rigid than approaches that force architectural priors (e.g., Manhattan-based methods). Nonetheless, the proposed loss term, based on structural verticality and regularity, is less effective when these basic assumptions are violated. Fig. 7 illustrates what happens in an edge case, where the scene is part of a construction that opens to the outside and contains predominantly outdoor elements (Fig. 7a). In this scenario, it is evident that the resulting density map, particularly on the horizontal plane (Fig. 7c), lacks the typical features visible in typical indoor settings, where, e.g., walls have important prominence. While this does not prevent the network from estimating a plausible depth map, the result is not a very good match (Fig. 7b). This failure case highlights both the advantages and limitations of structure-guided approaches, which improve over general-purpose solutions in the restricted cases when their underlying assumptions are met.

7. Conclusions

Our work introduces a novel deep neural network designed for fast and structurally consistent monocular 360° depth estimation in indoor environments. This network infers a depth map from a single gravity-aligned or gravity-rectified equirectangular image, ensuring that the predicted depth matches the typical depth distribution and features of cluttered interior spaces. This is achieved by a network architecture that leverages the unique characteristics of vertical and horizontal features present in man-made interior environments through gravity-aligned feature flattening, feeding specialized vision transformers. To improve structural consistency, we introduced a novel purely data-driven loss function that measures the difference between the density maps constructed by projecting

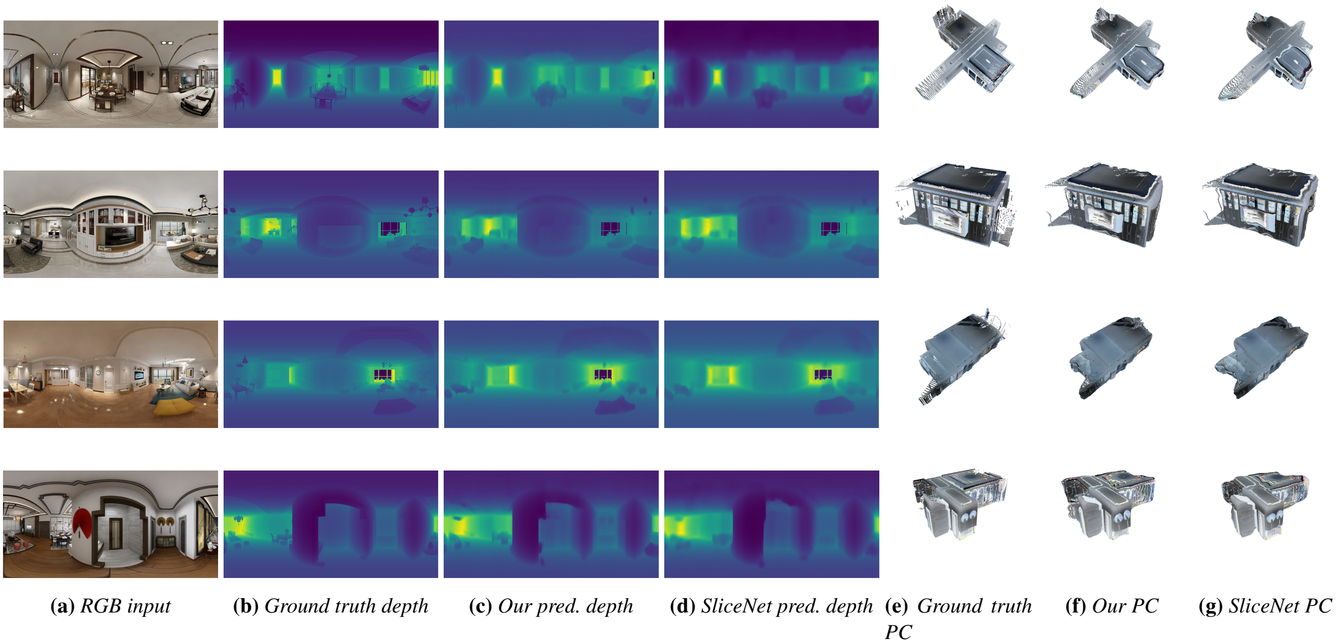


Figure 5: ***Qualitative performance and comparisons on Structured3D dataset [ZZL*20].*** *We illustrate our qualitative performance compared to the competitor [PAA*21] that reported the best performance in Tab. 2. We show reconstructions and ground-truth models as depth maps and as canonical views of the associated point clouds.*

predicted depth values onto horizontal (i.e., full-view cylindrical projection) and vertical planes and those built from training data.

Our experiments show that this approach achieves very good depth estimation results while maintaining a lightweight architecture with the low computational demands required by real-time usage in applications such as extended reality exploration and autonomous navigation. The solution offers greater structural consistency compared to existing methods that focus on optimizing pixel-wise depth estimation accuracy. Moreover, consistency is achieved by learning hidden relations from example sets, rather than implicitly or explicitly forcing the alignment with strict planar layouts. The results presented in this work highlight the benefit of the new density map computation and loss function design relative to the original conference contribution [PASG24], and include a much extended evaluation on real and synthetic data, a comparison with state-of-the-art solutions, a complete ablation study, and a critical discussion.

Since our work focused on evaluating the benefits of our novel density-map-based loss, we did not complement it with other loss components, demonstrating how we can achieve good structural preservation and depth estimation performance even in simplified settings. In the future, to further improve performance, we plan to extend this work by supplementing the loss with other gradient-based terms and normals, working especially at the fine-detail scale, as in other state-of-the-art methods.

Finally, while this work focused on the depth estimation task, we also plan to exploit our method as a building block inside a full processing pipeline. The two use cases that we are targeting are the extraction of multi-room 3D models from very sparse sampling (e.g., one image per room) and the generation of depth to support the synthesis and exploration of stereoscopic environments from a single surround-view panoramic image in extended reality settings. In both cases, the depth estimation task is very important, as is the preservation of structural consistency.

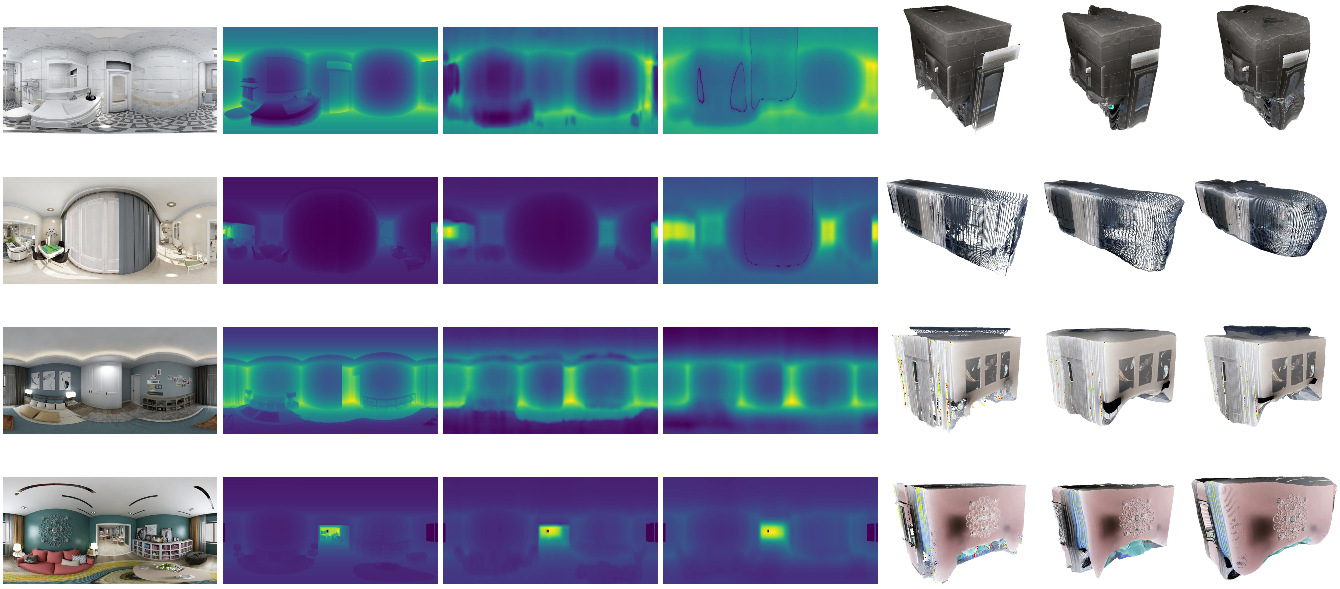
Acknowledgments This publication was supported by NPRP-S 14th Cycle grant 0403-210132 AIN2 from the Qatar National Research Fund (a member of Qatar Foundation). GP and EG also acknowledge the contribution of the Italian National Research Center in High-Performance Computing, Big Data, and Quantum Computing (Next Generation EU PNRR M4C2 Inv 1.4). The findings herein reflect the work and are solely the responsibility of the authors.

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

During the preparation of this work, the authors used ChatGPT solely to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

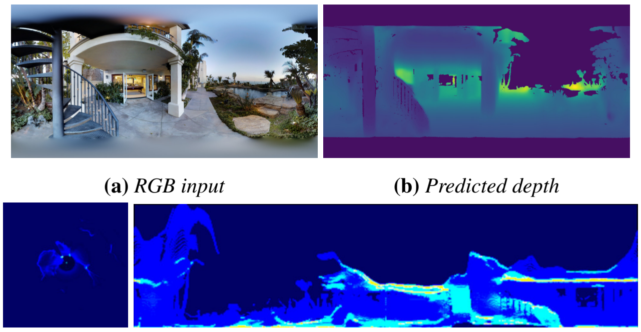
References

[ACC*23] AI H., CAO Z., CAO Y.-P., SHAN Y., WANG L.: HRDFuse: Monocular 360° depth estimation by collaboratively learning holistic-with-regional depth distributions. In *Proc. CVPR* (2023), pp. 13273–13282. 2, 3, 4



(a) *RGB input* **(b)** *Ground truth depth* **(c)** *Our predicted depth* **(d)** *SliceNet predicted depth* **(e)** *Ground truth point cloud* **(f)** *Our point cloud* **(g)** *SliceNet point cloud*

Figure 6: ***Qualitative comparison with structure-guided indoor depth estimation baseline.*** *We illustrate our qualitative performance on SKI360 dataset [SK20], compared to a state-of-the-art solution that, as ours, exploits gravity alignment (SliceNet [PAA*21]). We show reconstructions and ground-truth models as depth maps and as canonical views of the associated point clouds.*



(a) *RGB input* **(b)** *Predicted depth*

(c) *Floor OM* **(d)** *Cylindrical OM*

Figure 7: ***Failure case illustrating the limitations of the proposed loss term.****. The input RGB image (Fig. 7a) depicts a semi-outdoor scene that violates the indoor structural assumptions of the training data. As a result, the density map, particularly on the horizontal plane (Fig. 7c), lacks meaningful features. Despite this, the network is still able to produce a plausible pixel-wise depth prediction (Fig. 7b).*

[AW24] Ai H., Wang L.: Elite360d: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 9926–9935. 2, 3, 7, 8, 9

[CAVW24] Cao Z., Ai H., Vasilakos A. V., Wang L.: 360 high-resolution depth estimation via uncertainty-aware structural knowledge transfer. *IEEE Transactions on Artificial Intelligence* (2024). 3

[CCD*18] Cheng H., Chao C., Dong J., Wen H., Liu T., Sun M.: Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proc. CVPR* (2018), pp. 1420–1429. 2

[CLWF19] Chen J., Liu C., Wu J., Furukawa Y.: Floor-sp: Inverse cad for floorplans by sequential room-wise shortest path. In *Proc. CVPR* (2019), pp. 2661–2670. 6

[CQF22] Chen J., Qian Y., Furukawa Y.: HEAT: Holistic edge attention transformer for structured reconstruction. In *Proc. CVPR* (2022), pp. 3866–3875. 2, 3

[CWS18] Cao Y., Wu Z., Shen C.: Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE TCSVT 28*, 11 (2018), 3174–3182. 2

[DAH20] Davidson B., Alvi M. S., Henriques J. F. H.: 360 camera alignment via segmentation. In *Proc. ECCV* (2020), pp. 579–595. 4

[dSPMLJ22] da Silveira T. L., Pinto P. G., Murrugarra-Llerena J., Jung C. R.: 3D scene geometry estimation from 360° imagery: A survey. *ACM Computing Surveys 55*, 4 (2022), 1–39. 2

[EF15] Eigen D., Fergus R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proc. ICCV* (2015), pp. 2650–2658. 2

[EPF14] Eigen D., Puhrsch C., Fergus R.: Depth map prediction from a single image using a multi-scale deep network. In *Proc. NeurIPS* (2014), pp. 2366–2374. 2

[FGW*18] Fu H., Gong M., Wang C., Batmanghelich K., Tao D.: Deep ordinal regression network for monocular depth estimation. In *Proc. CVPR* (June 2018). 2

[GSZ*21] Gkitsas V., Sterzentsenko V., Zioulis N., Albanis G., Zarpalas D.: PanoDR: Spherical panorama diminished reality for indoor scenes. In *Proc. CVPR Workshops* (2021), pp. 3716–3726. 4

[HZRS16] He K., Zhang X., Ren S., Sun J.: Deep residual learning for image recognition. In *Proc. CVPR* (2016), pp. 770–778. 2, 4

[IYF15] IKEHATA S., YANG H., FURUKAWA Y.: Structured indoor modeling. In *Proc. ICCV* (2015), pp. 1323–1331. 1

[JLAB19] JUNG R., LEE A. S. J., ASHTARI A., BAZIN J.: Deep360Up: A deep learning-based approach for automatic VR image upright adjustment. In *Proc. IEEE VR* (2019), pp. 1–8. 4

[JSZ*21] JIANG H., SHENG Z., ZHU S., DONG Z., HUANG R.: Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters 6*, 2 (2021), 1519–1526. 8, 9

[JXZ*20] JIN L., XU Y., ZHENG J., ZHANG J., TANG R., XU S., YU J., GAO S.: Geometric structure based and regularized depth estimation from 360 indoor imagery. In *Proc. CVPR* (2020), pp. 889–898. 7, 8, 9, 10, 11

[LCG15] LIU F., CHUNHUA SHEN, GUOSHENG LIN: Deep convolutional neural fields for depth estimation from a single image. In *Proc. CVPR* (2015), pp. 5162–5170. 2

[LGY*22] LI Y., GUO Y., YAN Z., HUANG X., DUAN Y., REN L.: Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *Proc. CVPR* (2022), pp. 2801–2810. 3

[LHKK18] LEE J., HEO M., KIM K., KIM C.: Single-image depth estimation based on fourier domain analysis. In *Proc. CVPR* (2018), pp. 330–339. 2

[LLZ16] LAMBERT-LACROIX S., ZWALD L.: The adaptive BerHu penalty in robust regression. *Journal of Nonparametric Statistics 28* (06 2016), 1–28. 2, 4

[LRB*16] LAINA I., RUPPRECHT C., BELAGIANNIS V., TOMBARI F., NAVAB N.: Deeper depth prediction with fully convolutional residual networks. In *Proc. 3DV* (2016), pp. 239–248. 2, 3, 5

[Mat17] MATTERPORT: Matterport3D. https://github.com/niessner/Matterport, 2017. [Accessed: 2022-09-25]. 2, 7, 8, 9, 11

[PAA*21] PINTORE G., AGUS M., ALMANSA E., SCHNEIDER J., GOBBETTI E.: SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proc. CVPR* (2021), pp. 11536–11545. 1, 2, 3, 4, 7, 8, 9, 10, 11, 12, 13

[PAAG21] PINTORE G., ALMANSA E., AGUS M., GOBBETTI E.: Deep3DLayout: 3D reconstruction of an indoor layout from a spherical panoramic image. *ACM TOG 40*, 6 (2021), 250:1–250:12. 1, 4

[PAG24] PINTORE G., AGUS M., GOBBETTI E.: Automatic 3D modeling and exploration of indoor structures from panoramic imagery. In *SIGGRAPH Asia 2024 Courses (SA Courses '24)* (2024). 1, 2

[PASG24] PINTORE G., AGUS M., SIGNORONI A., GOBBETTI E.: Ddd: Deep indoor panoramic depth estimation with density maps consistency. In *2024 Eurographics Italian Chapter Conference on Smart Tools and Applications in Graphics, STAG 2024* (2024), Eurographics Association, pp. stag–20241336. 2, 3, 5, 6, 7, 8, 9, 10, 12

[PBAG23] PINTORE G., BETTIO F., AGUS M., GOBBETTI E.: Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image. *IEEE TVCG 29* (November 2023). 2

[PdLGAAB18] PAYEN DE LA GARANDERIE G., ATAPOUR ABARGHOUEI A., BRECKON T. P.: Eliminating the blind spot: Adapting 3D object detection and monocular depth estimation to 360 panoramic imagery. In *Proc. ECCV* (2018), pp. 812–830. 2

[PGGS16] PINTORE G., GANOVELLI F., GOBBETTI E., SCOPIGNO R.: Mobile mapping and visualization of indoor structures to simplify scene understanding and location awareness. In *Proc. ECCV Workshops* (October 2016), Springer, pp. 130–145. 1

[PJVH*24] PINTORE G., JASPE-VILLANUEVA A., HADWIGER M., SCHNEIDER J., AGUS M., MARTON F., BETTIO F., GOBBETTI E.: Deep synthesis and exploration of omnidirectional stereoscopic environments from a single surround-view panoramic image. *Computers & Graphics 119* (March 2024), 103907. 1, 2

[PMG*20] PINTORE G., MURA C., GANOVELLI F., FUENTES-PEREZ L., PAJAROLA R., GOBBETTI E.: State-of-the-art in automatic 3D reconstruction of structured indoor environments. *Comput. Graph. Forum 39*, 2 (2020), 667–699. 1, 2

[PXZ*15] PENG WANG, XIAOHUI SHEN, ZHE LIN, COHEN S., PRICE B., YUILLE A.: Towards unified depth and semantic prediction from a single image. In *Proc. CVPR* (2015), pp. 2800–2809. 2

[RAYR22] REY-AREA M., YUAN M., RICHARDT C.: 360MonoDepth: High-resolution 360° monocular depth estimation. In *Proc. CVPR* (2022), pp. 3762–3772. 3

[RSL*24] RAJAPAKSHA U., SOHEL F., LAGA H., DIEPEVEEN D., BENNAMOUN M.: Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey. *ACM Computing Surveys 56*, 12 (2024), 1–51. 8

[SG17] SU Y.-C., GRAUMAN K.: Learning spherical convolution for fast features from 360 imagery. In *Proc. NeurIPS* (2017), pp. 529–539. 2

[SG19] SU Y., GRAUMAN K.: Kernel transformer networks for compact spherical convolution. In *Proc. CVPR* (2019), pp. 9434–9443. 2

[SHSC19] SUN C., HSIAO C.-W., SUN M., CHEN H.-T.: HorizonNet: Learning room layout with 1D representation and pano stretch data augmentation. In *Proc. CVPR* (2019), pp. 1047–1056. 4

[SK20] SHANGHAITECH UNIVERSITY, KUJIALE.COM: Shanghaitech-Kujiale Indoor 360° dataset, 2020. [Online; accessed 2024-08-19]. URL: https://svip-lab.github.io/dataset/indoor_360.html. 7, 8, 10, 13

[SLL*22] SHEN Z., LIN C., LIAO K., NIE L., ZHENG Z., ZHAO Y.: PanoFormer: Panorama transformer for indoor 360 depth estimation. In *Proc. ECCV* (2022), Springer, pp. 195–211. 2, 7, 8, 9

[SRFL21] STEKOVIC S., RAD M., FRAUNDORFER F., LEPETIT V.: Montefloor: Extending MCTS for reconstructing accurate large-scale floor plans. In *Proc. CVPR* (2021), pp. 16034–16043. 1

[SSC21] SUN C., SUN M., CHEN H.-T.: HoHoNet: 360° indoor holistic understanding with latent horizontal features. In *Proc. CVPR* (2021), pp. 2573–2582. 2, 3, 4

[SSN09] SAXENA A., SUN M., NG A. Y.: Make3D: Learning 3D scene structure from a single still image. *IEEE TPAMI 31*, 5 (2009), 824–840. 2

[SZL*23] SHEN Z., ZHENG Z., LIN C., NIE L., LIAO K., ZHENG S., ZHAO Y.: Disentangling orthogonal planes for indoor panoramic room layout estimation with cross-scale distortion awareness. In *Proc. CVPR* (2023), pp. 17337–17345. 2, 3, 4

[TNT18] TATENO K., NAVAB N., TOMBARI F.: Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proc. ECCV* (2018), pp. 732–750. 2

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł., POLOSUKHIN I.: Attention is all you need. *Proc. NeurIPS 30* (2017). 4

[WGSJ20] WILES O., GKIOXARI G., SZELISKI R., JOHNSON J.: Synsin: End-to-end view synthesis from a single image. In *Proc. CVPR* (2020), pp. 7467–7477. 2

[WL24] WANG N.-H. A., LIU Y.-L.: Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. *Advances in Neural Information Processing Systems 37* (2024), 127739–127764. 3

[WYS*20] WANG F.-E., YEH Y.-H., SUN M., CHIU W.-C., TSAI Y.-H.: BiFuse: Monocular 360 depth estimation via bi-projection fusion. In *Proc. CVPR* (2020), pp. 462–471. 3, 7, 8, 9

[XLF*19] XIAN W., LI Z., FISHER M., EISENMANN J., SHECHTMAN E., SNAVELY N.: UprightNet: geometry-aware camera orientation estimation from single images. In *Proc. ICCV* (2019), pp. 9974–9983. 4

[XWT*18] XU D., WANG W., TANG H., LIU H., SEBE N., RICCI E.: Structured attention guided convolutional neural fields for monocular depth estimation. In *Proc. CVPR* (2018), pp. 3917–3925. 2

[YJL*18] YANG Y., JIN S., LIU R., , YU J.: Automatic 3D indoor scene modeling from single panorama. In *Proc. CVPR* (2018), pp. 3926–3934. 1

[YKSE23] YUE Y., KONTOGIANNI T., SCHINDLER K., ENGELMANN F.: Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries. In *Proc. CVPR* (2023). 2, 3

[YSL*23] YUN I., SHIN C., LEE H., LEE H.-J., RHEE C.-E.: EGformer: Equirectangular geometry-biased transformer for 360° depth estimation. In *Proc. ICCV* (2023), pp. 6078–6089. 7, 8, 9

[ZKZD18] ZIOULIS N., KARAKOTTAS A., ZARPALAS D., DARAS P.: OmniDepth: Dense depth estimation for indoors spherical panoramas. In *Proc. ECCV* (2018), pp. 453–471. 2, 3

[ZSTX14] ZHANG Y., SONG S., TAN P., XIAO J.: PanoContext: A whole-room 3D context model for panoramic scene understanding. In *Proc. ECCV* (2014), pp. 668–686. 2

[ZZL*20] ZHENG J., ZHANG J., LI J., TANG R., GAO S., ZHOU Z.: Structured3D: A large photo-realistic dataset for structured 3D modeling. In *Proc. ECCV* (2020), pp. 519–535. 2, 7, 8, 9, 12