STAG: Smart Tools and Applications in Graphics (2024)
A. Caputo, V. Garro, U. Castellani, and A. Giachetti (Editors)

DDD: Deep indoor panoramic Depth estimation with Density maps consistency

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 for rapid and structurally consistent monocular 360° depth estimation in indoor environments. The network infers a depth map from a single gravity-aligned or gravity-rectified equirectangular image of the environment, ensuring that the predicted depth aligns with the typical depth distribution and features of cluttered interior spaces, which are usually enclosed by walls, ceilings, and floors. By leveraging the distinct characteristics of vertical and horizontal features in man-made indoor environments, we introduce a lean network architecture that employs gravity-aligned feature flattening and specialized vision transformers that utilize the input's omnidirectional nature, without segmentation into patches and positional encoding. To enhance the structural consistency of the predicted depth, we introduce a new loss function that evaluates the consistency of density maps by projecting points derived from the inferred depth map onto horizontal and vertical planes. This lightweight architecture has very small computational demands, provides greater structural consistency than competing methods, and does not require the explicit imposition of strong structural priors.

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 sub-field within 3D reconstruction [PMG*20]. One of the main focuses concerns specialized techniques for common, highly structured environments such as residential, office, and public buildings, that constitute the majority of the built environment and require the availability of 3D reality-based models for many purposes [IYF15, PGGS16].

Fast depth estimation from images is a fundamental sub-problem in this context since associating metric information to visual data is necessary for 3D reconstruction and scene understanding, and rapid (real-time) and accurate 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. A 360° image, that can be quickly captured with widely available and affordable consumer-level and professional cameras, encompasses the complete scene visible from a specific viewpoint within a 360° field of view at a given instant, offering ample context for monocular depth inference and scene understanding [YJL*18]. Moreover, as a presentation medium, a single panoramic image is not consumed at once, and its exploration is inherently more dynamic than traditional 2D imagery. Especially when presented through a Head-Mounted Display (HMD), 360° images have become a key component for creating immersive content directly from real scenes and for supporting a range of Virtual Reality (VR) applications [PAAG22], where an associated depth is used to enhance immersion (e.g, through stereo and motion parallax) [MCE*17].

Despite the full context provided by 360° images, monocular depth estimation remains very challenging in indoor environments, even more than in typical outdoor depth-estimation settings, e.g., as found in autonomous driving contexts [WWH*22]. Indoor scenes are typically characterized by narrow spaces filled with objects, including but not limited to furniture, and bound by architectural structures, such as walls, floors, and ceilings. As a result, indoor depth is unevenly distributed between near and far ranges, e.g., zoom-in views of close-by furniture vs. ceilings, making it very challenging to predict accurate metric depths. Even though structure priors characterize the architectural shape that bounds the scene, it is hard to recognize them, since objects can be cluttered and arranged arbitrarily in the near field, masking large portions of a room's walls and floors. Moreover, the fact that the bounding structures, such as walls, are often composed of large untextured regions makes the commonly used photometric losses ambiguous.

To this end, specific 360° solutions targeting indoor environments have been introduced, reaching impressive results, especially in conjunction with deep-learning approaches capable of learning

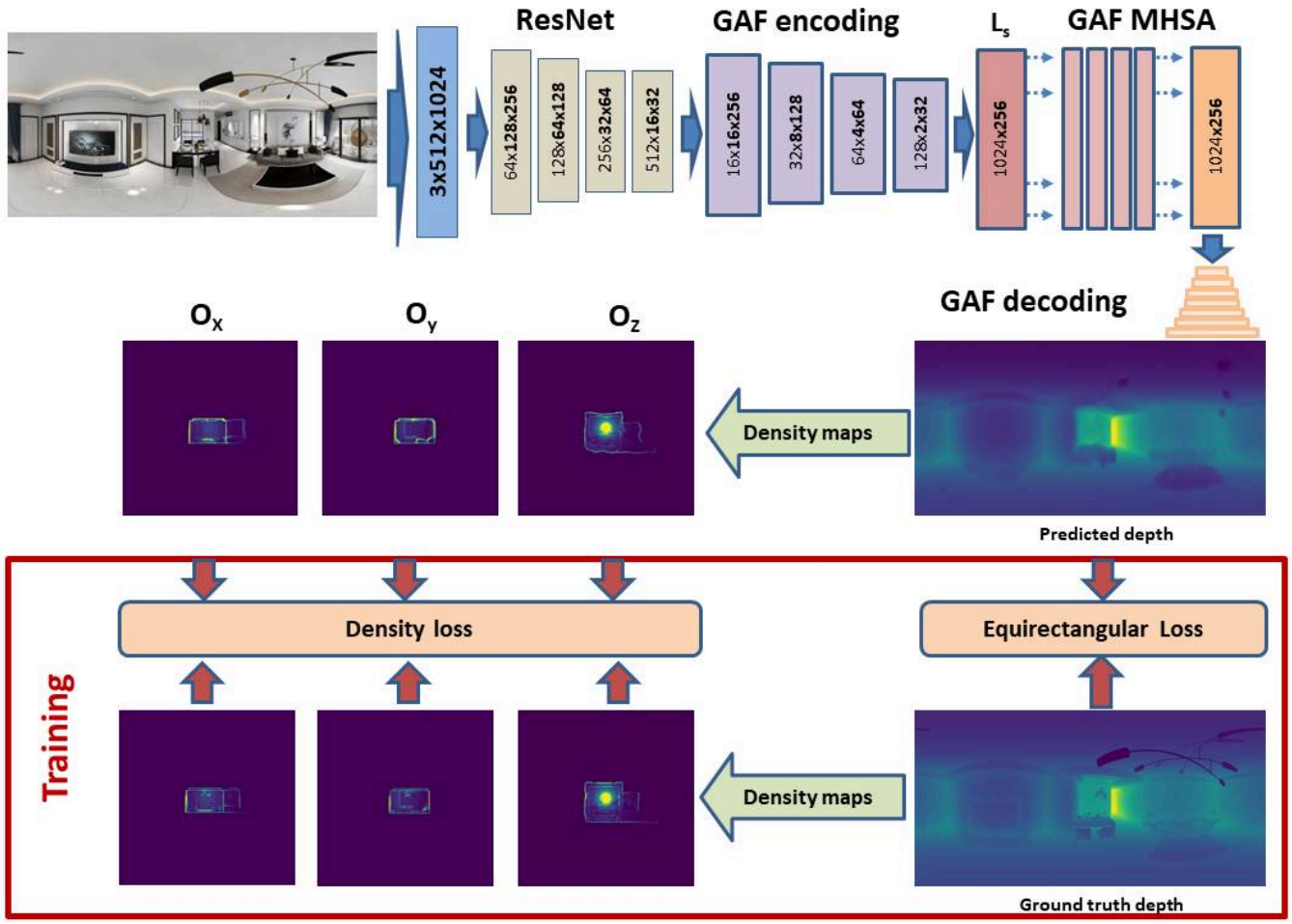


Figure 1: **Overview**. The network maps a gravity-aligned 360° image to its depth. The input image is first transformed by a ResNet block (light green) into feature maps with different depths and spatial sizes. Through a Gravity-Aligned Features (GAF) encoding block (light purple), we perform a gravity-aligned anisotropic contractive encoding to obtain latent features, that, once assembled in a single sequence, 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. At training time, we produce density maps that are used for our structural loss computation (red rectangle) through differentiable rendering. This network's section is only used for training, and its memory and computational costs do not count at inference time. Note that the image depth and density maps have been visually enhanced for illustration purposes.

hidden relations from large sets of curated examples. However, 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 with other single-view reconstruction for structural segmentation of complex multi-room environments [CQF22, YKSE23]. Moreover, it is not uncommon, see Sec. 5 to require in the order 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. Its design considers the observed scene's indoor features to improve depth estimation and facilitate the recognition of the indoor architectural structure in reconstruction tasks while ensuring a low computational cost at inference time, i.e., about half of current state-of-the-art solutions. To design our network, see Sec. 3 and Fig. 1, we start from the assumption that gravity is a key factor in shaping interior environments. Thus, 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 on the point cloud resulting from the depth (see Sec. 4). We do that by computing density maps along preferential horizontal and vertical directions and comparing them 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, provide good summaries of the characteristics of indoor environments characterized by vertical walls, and are currently used for layout reconstruction and segmentation tasks, and not for depth prediction [CQF22, YKSE23].

Our contributions are summarized as follows:

- We introduce a novel metric and loss function which, starting from a spherical depth and its point cloud, supervises the training by minimizing the error on selected density maps of the 3D points, improving the overall quality of the depth prediction and simplifying the integration into pipelines for geometric reconstruction, especially for permanent indoor structures;
- We propose a very lightweight network architecture that moves most of the computational load related to geometric priors to the training phase, leading to a particularly fast inference and a low number of learnable weights. The network design combines gravity-aligned features obtained by asymmetric convolution of the input with multi-head self-attention. Our peculiar feature flattening enables the direct use of a vision transformer, without the need to sequence the input map by arbitrary patches and positional encoding [SLL*22].

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 makes it also ideal for real-time usage.

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]. 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 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]. 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 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. Moreover, 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]. 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 linear sequence that can be used to directly feed a self-attention vision transformer, without the need to arbitrarily partition the image into patches and use positional encoding [SZL*23, ACC*23]. In particular, we adopt a single-layer multi-head self-attention (MHSA) scheme [VSP*17] to leverage complementary features in distant portions of the image rather than only local regions, to maximize the wide contextual information provided by omnidirectional images while keeping the computational cost low. Our self-attention module takes the latent features $L \in \mathbb{R}^{s \times l}$ as input, and outputs a self-attention weight matrix $A \in \mathbb{R}^{s \times s}$:

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

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 $(1 \times h \times w)$.

The described network path completes the mapping from input colors to output depths. While the depths are the desired outcome of the network at inference time, in our architecture we also produce density maps along preferential horizontal and vertical directions by differentiable rendering of the depth map (see Fig. 1). Since these maps are used only at training time, they are described in Sec. 4. Here, it is important to note that this section of the network is removed at inference time and, thus, the memory and computational costs of density map computation are only relevant for training.

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]:

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

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 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

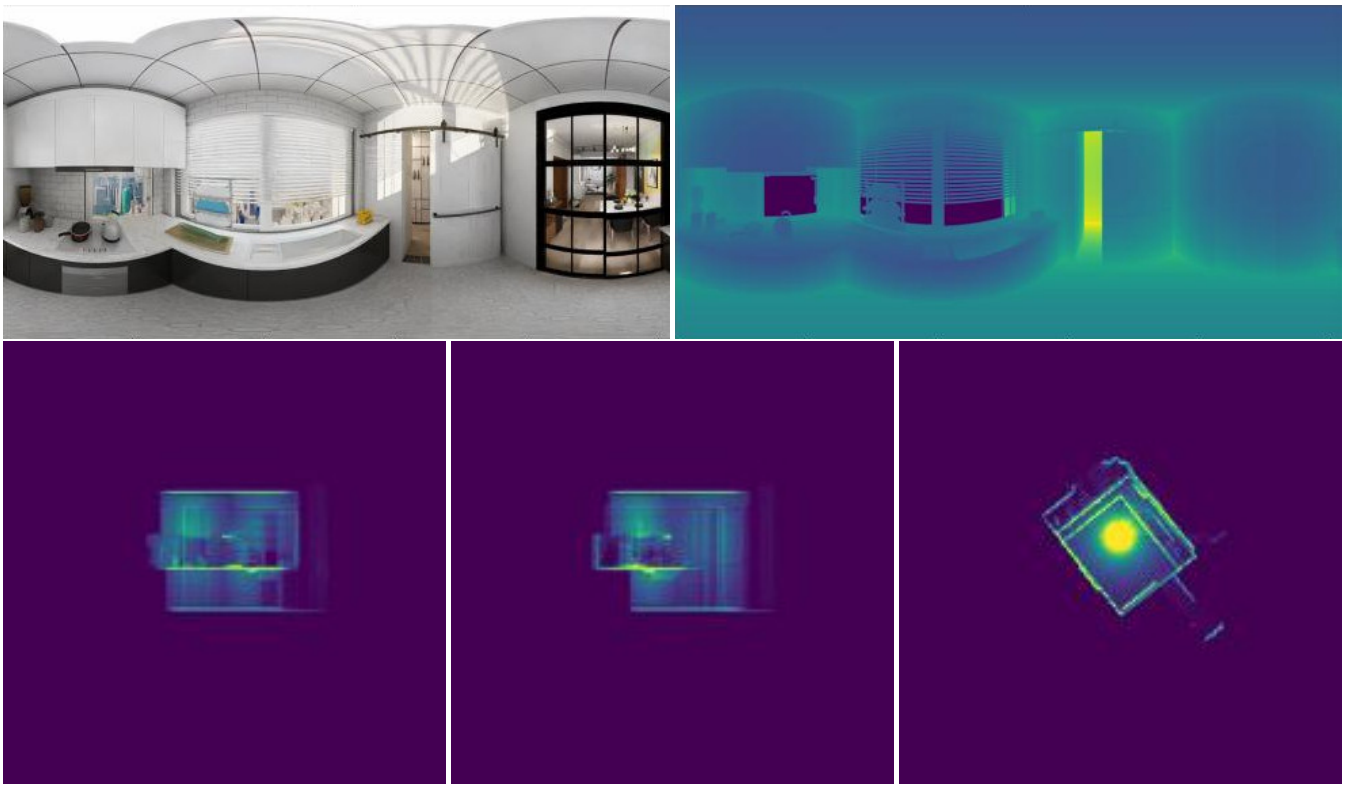


Figure 2: **Density maps example**. We show an example of a view of a scene not aligned with canonical Manhattan axes. In this case, the projection on the floorplan (bottom right image) represents the shape of the walls, while the horizontal projections (bottom left and center images) highlight the surfaces perpendicular to the walls, and in particular floor and ceiling. Note, also, how standard heights of several kinds of furniture are also highlighted.

introduce in this work a structural term, 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.

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 along preferential horizontal and vertical directions by differentiable rendering into fixed-size buffers (512×512 in our experiments). To do that, we render three density maps, Ox, Oy (i.e., horizontal projections) and Oz (i.e., vertical projection) from depth prediction D and three density maps Ox^*, Oy^*, Oz^* 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 Oz highlights the floor plan, 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 projections Ox and Oy, independently from the horizontal orientation of the projection, emphasize the shapes of ceilings and floors aligned with the projection axes are evident. Fig. 2 shows an example of a scene aligned only relative to the gravity axis, but with an arbitrary orientation around the axis. In this case, the projection on the floorplan represents the shape of the walls, while the horizontal projections highlight the surfaces perpendicular to the walls.

To exploit the information available in our projected density maps, therefore, we calculate the structural loss term 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:

$$\begin{aligned}\mathcal{L}_{ds}(Ox,Oy,Oz,Ox^*,Oy^*,Oz^*) = &\sum_{kl}\mathcal{H}(Ox_{kl}-Ox^*_{kl}) \quad + \\ &\sum_{kl}\mathcal{H}(Oy_{kl}-Oy^*_{kl}) \quad + \\ &\sum_{kl}\mathcal{H}(Oz_{kl}-Oz^*_{kl}) \qquad (4)\end{aligned}$$

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.

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.

It is important to note that 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, as done for cubemap representations. In future work, we plan to explore the use of cylindrical density maps to determine whether their increased continuity relative to angular orientation can enhance robustness and reduce the need for extensive augmentation.

5. Results

Our approach was implemented using *PyTorch* and has been tested on several kinds of indoor scenes. In the following, we first discuss the specific training dataset 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).

5.1. Training data

In this paper, we exploit the publicly available Shanghaitech-Kujiale Indoor 360° (SKI360) dataset [SK20] for both training and testing, as it has been generated to develop and evaluate solutions exploiting geometric cues as priors and regularizers to improve depth inference [JXZ*20]. 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]. This work does not use the additional data included in the dataset (e.g., unfurnished versions and precise 3D layout). As for most currently available datasets (e.g., Structured3D [ZZL*20]), RGB images and depth maps are provided at the resolution of 512×1024.

5.2. Training setup and computational performance

We trained our network with the Adam optimizer, with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and an adaptive learning rate from 0.0001, on an NVIDIA RTX A5000 (24GB VRAM) with a batch size of 8. At the native resolution of 512×1024, the average training time is 76 ms/image, while the inference time on the same NVIDIA RTX A5000 is 18 ms/image.

Tab. 1 presents the computational performance of inference with our network compared 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.

Method	Parameters↓	FLOPs↓
Bifuse [WYS*20]	253 M	682 G
SliceNet [PAA*21]	79 M	101 G
Panoformer [SLL*22]	20 M	78 G
EGFormer [YSL*23]	**15 M**	74 G
DDD (our)	23 M	**38 G**

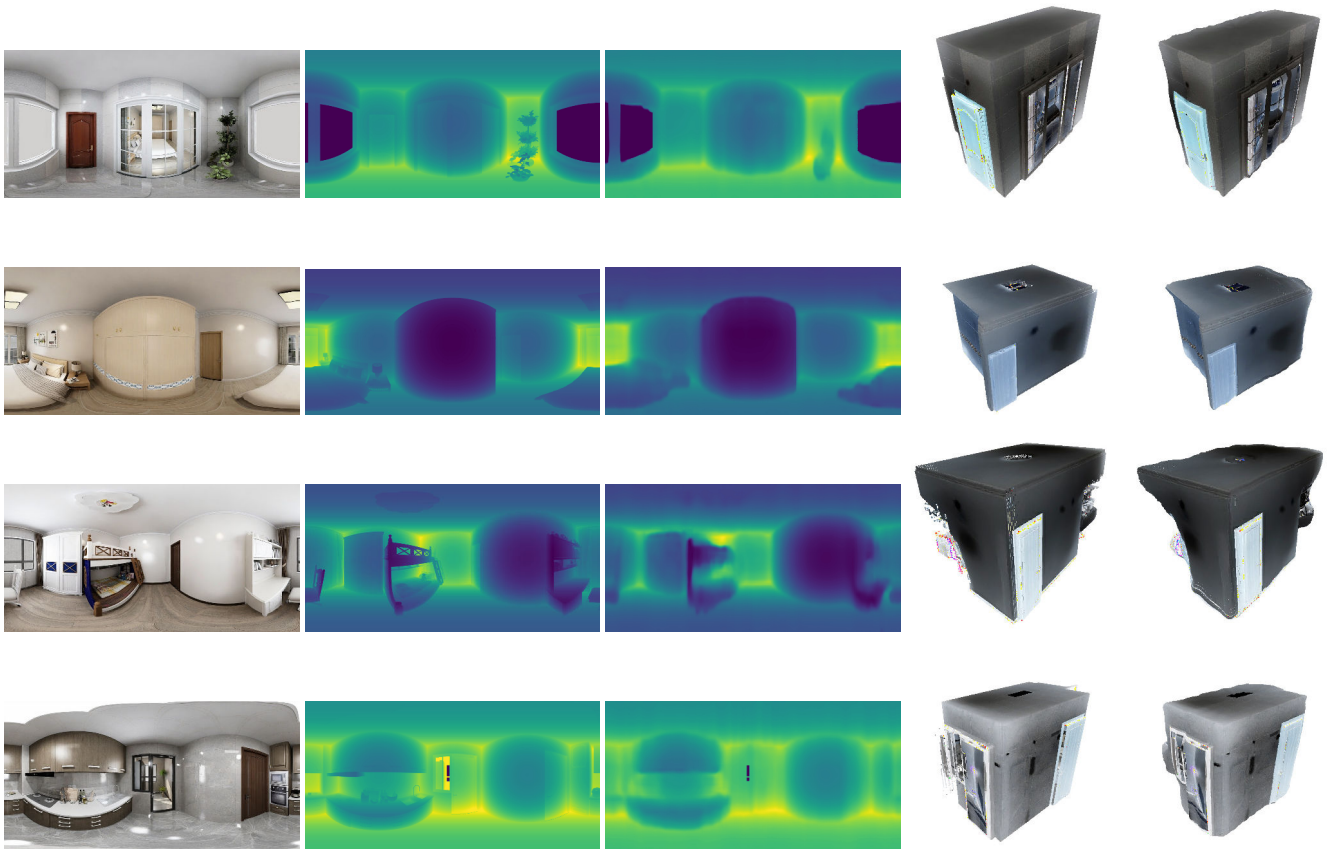
Table 1: **Computational performance of inference.** We show our computational performance compared to other state-of-the-art works for a 512×1024 image.

As we can see, our approach has, by far, the lowest computational complexity (FLOPs) of the compared methods (see Sec. 5.3). Its computational cost, is, in particular, less than half of the currently fastest method (EGFormer [YSL*23]). 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 much larger image sizes than competitors when suitable higher-resolution training data will be available.

5.3. Quantitative and qualitative evaluation of results

This work provides a preliminary evaluation of the method, focusing on the benefits of taking into consideration geometric consistency. For this purpose, we compare our results with those obtained by two representative state-of-the-art solutions. SliceNet [PAA*21] introduced the concept of gravity-aligned features for indoor depth estimation but, similarly to most other depth estimation networks, does not use specific structural consistency terms. The framework introduced by Jin et al. [JXZ*20], instead, is a representative state-of-the-art pipeline that jointly predicts per-pixel depth and layout, i.e., the structural shape of the room represented by a collection of corners, boundaries, and planes. The correlation between depth and layout provides a strong form of structural consistency that is exploited for geometric structure-based and regularized depth extraction. For SliceNet, we have retrained the network and run the benchmarks using the same settings as our solution, using the publicly available source code [PAA*21]. Both SliceNet and our method use the original split, with 1500 scenes selected for training and 275 for testing. For Jin et al. [JXZ*20], we compare instead with the reported official results on depth estimation performance for the same dataset. Note that, due to the high cost of their network, their results are reported for downscaled 256×512 resolution, and, thus, their performance may be slightly overestimated due to the reduced amount of details present in the half-resolution images.

Tab. 2 summarizes the depth estimation performance. 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 structural consistency through geometric priors and regularizers. The third row reports the results obtained by SliceNet [PAA*21], which uses GAFs but no structural consistency terms. Finally, the last two rows report the results obtained with our method (DDD) without ("no DL") and with ("DL") the loss term exploiting density maps. For this quantitative evaluation, we adopt the most common



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

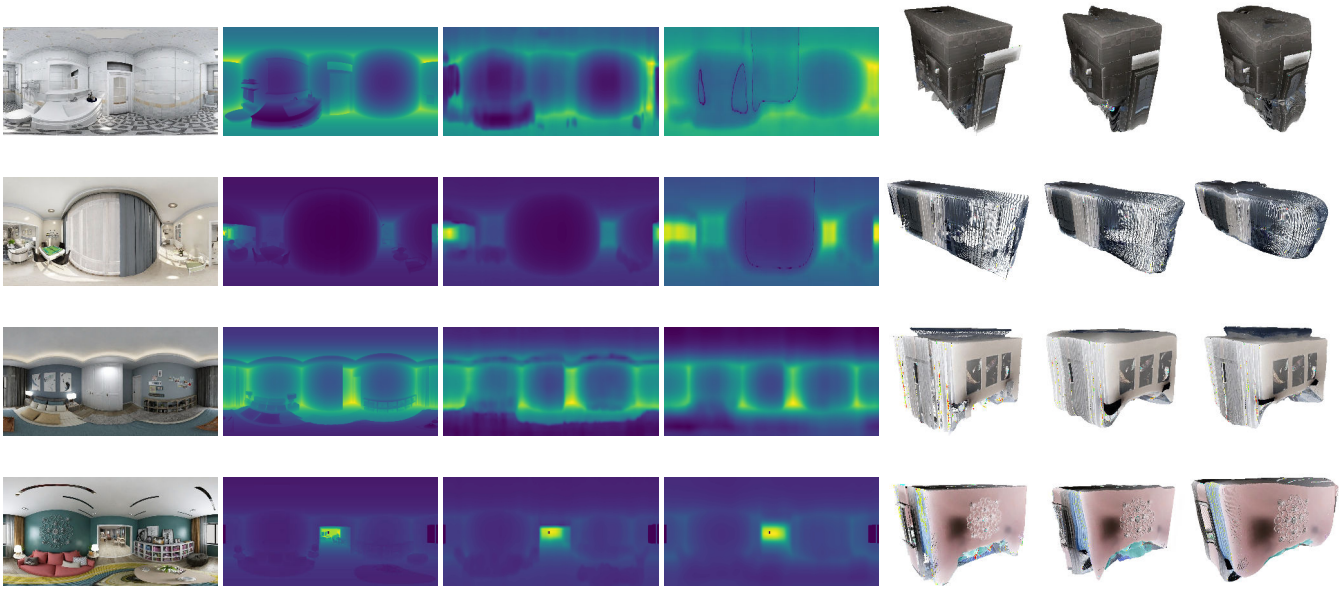
Figure 3: **Qualitative results.** Selected examples of our predictions compared to ground truth. We show reconstructions and ground-truth models as depth maps and as canonical views of the associated point clouds. Performance measures for the entire test dataset are summarized in Tab. 2.

metrics used for indoor depth estimation [EPF14], i.e., mean relative error (MRE) and root mean square error of linear measures (RMSE) and three relative accuracy measures δ_1, δ_2 and δ_3, defined, for an accuracy δ_n, as the fraction of pixels where the relative error is within a threshold of 1.25^n.

The results 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 including structural consistency, independently from the different paths taken. Moreover, our method, when including the structural consistency terms, achieves state-of-the-art performance despite the much lower computational burden compared to the other baselines. Fig. 3 shows our reconstruction results compared to the ground truth depth map and point cloud.

Notably, the data-driven structural consistency provided by our loss term on density maps can enhance depth accuracy, even in comparison with solutions like SliceNet [PAA*21], which focuses on optimizing per-pixel accuracy in depth maps. We believe this improvement arises because optimizing errors on our gravity-aligned density maps allows our method to leverage the medium- and large-scale regularities found in typical indoor structures more effectively. It is interesting to note that SliceNet is very competitive in terms of average error measures (MRE and MSE), surpassing the performance of the other solutions without the structural consistency term included. Moreover, it even surpasses the performance of Jin et al. [JXZ*20]'s solution with structural consistency enabled. However, δ_1 remains lower; thus, the number of pixels with a relative error larger than 25% is higher. This shows that SliceNet mostly improves because of the better preservation of details. Our method with the density-map-based loss, instead, achieves maximum performance on all measures.

Fig. 4 provides a qualitative comparison of the reconstructions obtained by our method to both ground truth and the SliceNet approach [PAA*21]. We specifically illustrate a few scenes where structural parts are more evident. The benefit of our method appears not only in the enhanced accuracy of depth prediction but also in its improved ability to delineate a room's bounding structure. We



(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 4: **Qualitative comparison with a pure indoor depth estimation method.** We illustrate our qualitative performance compared to a state-of-the-art solution for indoor depth estimation that only optimizes per-pixel depth measures (SliceNet [PAA*21]). We show reconstructions and ground-truth models as depth maps and as canonical views of the associated point clouds. Performance measures for the entire test dataset are summarized in Tab. 2.

Method	MRE↓	RMSE↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Jin [JXZ*20] no SC	0.114	0.721	0.894	0.973	0.989
Jin [JXZ*20] with SC	0.103	0.666	0.910	0.978	0.990
SliceNet [PAA*21]	0.102	0.273	0.904	0.977	0.989
DDD - no DL	0.075	0.278	0.907	0.977	0.989
DDD - with DL	**0.063**	**0.254**	**0.919**	**0.979**	**0.991**

Table 2: **Depth estimation performance and comparisons.** We show our quantitative performance compared to other representative state-of-the-art works. 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. Finally, the last two rows report the results obtained with our method (DDD) without (no DL) and with (DL) the loss term exploiting density maps.

expect that this feature will make it suitable for integration into pipelines that aim to extract the 3D layout.

The results in Tab. 2 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 paper. 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. Conclusions

Our work has introduced 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 predicted depth values onto horizontal and vertical planes and those built from training data.

Our initial 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.

For this study, we limited our loss function to the simplest possible form (i.e., using only one term for depth difference evaluation in addition to our novel loss terms). To further improve the depth estimation performance, particularly concerning fine-scale detail and non-structural objects, we plan to include other standard terms working at the small scale (e.g., ob gradients, normals, curvatures). We plan to evaluate the method against state-of-the-art methods on a larger variety of datasets.

We also plan to improve our approach further. First of all, we are experimenting with replacing the projection on two orthogonal planes for the horizontal density map computation with a cylindrical projection, with the expectation that this will reduce augmentation costs and increase robustness in the presence of arbitrarily aligned (i.e., non-Manhattan) layouts. We will, in addition, also 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.

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.

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. 3, 4

[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. 3

[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. 5

[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. 3

[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. 3

[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. 3

[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. 3, 7

[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). 3

[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. 3, 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

[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. 6, 7, 8

[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. 3

[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. 3

[LLZ16] LAMBERT-LACROIX S., ZWALD L.: The adaptive BerHu penalty in robust regression. *Journal of Nonparametric Statistics 28* (06 2016), 1–28. 3, 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. 3, 4

[MCE*17] MATZEN K., COHEN M. F., EVANS B., KOPF J., SZELISKI R.: Low-cost 360 stereo photography and video capture. *ACM TOG 36*, 4 (2017), 148:1–148:12. 1

[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. 2, 3, 4, 6, 7, 8

[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. 2, 4

[PAAG22] PINTORE G., AGUS M., ALMANSA E., GOBBETTI E.: Instant automatic emptying of panoramic indoor scenes. *IEEE Transactions on Visualization and Computer Graphics 28*, 11 (2022), 3629–3639. 1

[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, 3

[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. 3

[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. 2, 3

[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, 3

[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. 3

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

[SG17] SU Y.-C., GRAUMAN K.: Learning spherical convolution for fast features from 360 imagery. In *Proc. NeurIPS* (2017), pp. 529–539. 3

[SG19] SU Y., GRAUMAN K.: Kernel transformer networks for compact spherical convolution. In *Proc. CVPR* (2019), pp. 9434–9443. 3

[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. 6

[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. 3, 6

[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. 2

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

[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. 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. 3

[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

[WWH*22] WU C.-Y., WANG J., HALL M., NEUMANN U., SU S.: Toward practical monocular indoor depth estimation. In *Proc. CVPR* (2022), pp. 3804–3814. 1

[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, 6

[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. 3

[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. 6

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

[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. 6