

360° Surface Regression with a Hyper-Sphere Loss

Antonis Karakottas Nikolaos Zioulis Stamatis Samaras Dimitrios Ataloglou

Vasileios Gkitsas Dimitrios Zarpalas Petros Daras

Centre for Research and Technology Hellas (CERTH) - Information Technologies
Institute (ITI) - Visual Computing Lab (VCL), Thessaloniki, Greece

{ankarako, nzioulis, sstamatis, ataloglou, gkitsasv, zarpalas, daras}@iti.gr
vcl.iti.gr

Abstract

Omnidirectional vision is becoming increasingly relevant as more efficient 360° image acquisition is now possible. However, the lack of annotated 360° datasets has hindered the application of deep learning techniques on spherical content. This is further exaggerated on tasks where ground truth acquisition is difficult, such as monocular surface estimation. While recent research approaches on the 2D domain overcome this challenge by relying on generating normals from depth cues using RGB-D sensors, this is very difficult to apply on the spherical domain. In this work, we address the unavailability of sufficient 360° ground truth normal data, by leveraging existing 3D datasets and remodelling them via rendering. We present a dataset of 360° images of indoor spaces with their corresponding ground truth surface normal, and train a deep convolutional neural network (CNN) on the task of monocular 360° surface estimation. We achieve this by minimizing a novel angular loss function defined on the hyper-sphere using simple quaternion algebra. We put an effort to appropriately compare with other state of the art methods trained on planar datasets and finally, present the practical applicability of our trained model on a spherical image re-lighting task using completely unseen data by qualitatively showing the promising generalization ability of our dataset and model. The dataset is available at: vc13d.github.io/HyperSphereSurfaceRegression.

1. Introduction

Understanding 3D geometry from a single image is one of the most challenging and actively studied problems in computer vision. With the advent of efficient deep learning frameworks, many methods emerged that present state of the art results in tasks such as depth estimation [30, 19], surface normal prediction [16, 56] or a joint combination of

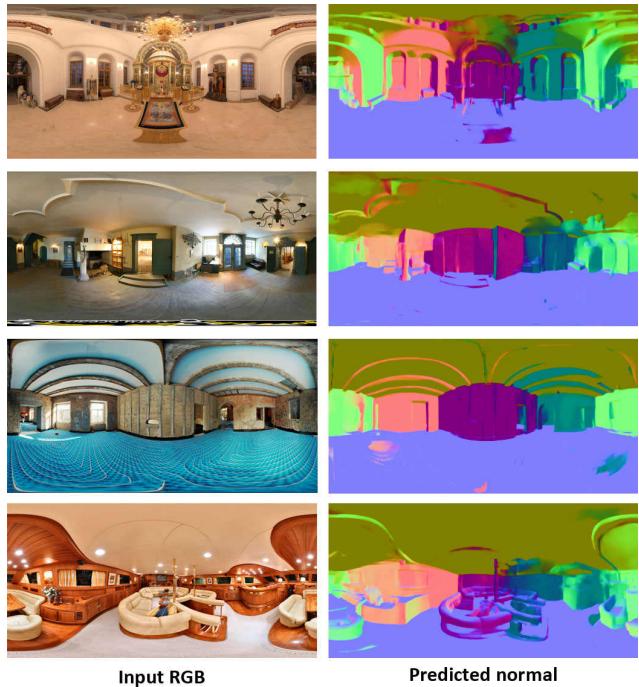


Figure 1: Qualitative results on samples of the realistic and unseen Sun360 [57] that contains indoors scene panoramas. Our model infers valid surface estimates, even on these challenging scenes, even though trained on a mix of synthetic and real - but different distribution (i.e. saturation, lighting, content) - scenes.

both [13]. 3D visual perception can trace the path for a number of applications, like autonomous driving [60, 58], robot navigation [63], 3D reconstruction [4] or even the fusion of two heterogeneous media, such as traditional 2D images with 3D objects for Augmented Reality (AR) applications [38].

Typical end-to-end deep learning pipelines usually require a large amount of ground truth annotated data. While

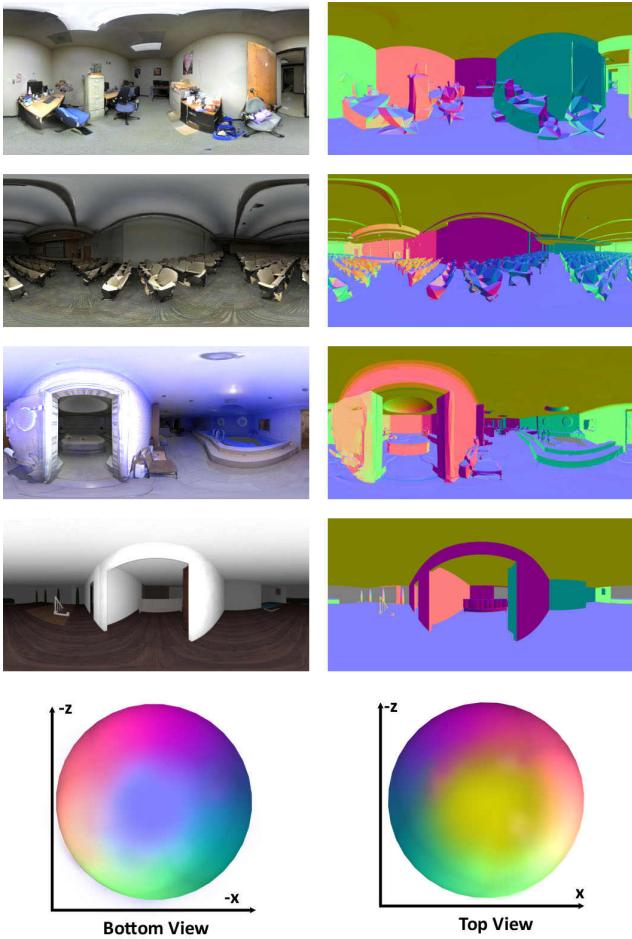


Figure 2: Samples of our generated dataset, with the color images next to their corresponding ground truth surface normal map. Below, the two color-spheres map pixel color to normal vector orientation. We consider a left-hand coordinate system, with the z and y axes representing the camera’s look-up and up vectors respectively. Left: a bottom-view of the color-sphere, where the center-color represents the positive y -axis (y^+). Right: a top-view of the color-sphere, where the center-color represents the negative y -axis (y^-). The main coordinate axes x^+ , x^- , z^+ , z^- , have the same colors in both the color-spheres.

this is partially addressed for datasets captured by traditional techniques following the typical pinhole camera projection model using depth sensors [48] or laser scanners [45], the same cannot be said for 360° content¹, which is still considered a novel research domain with limited work done regarding 3D perception.

Nowadays, with good quality and efficient commercial based 360° spherical cameras and rigs, omnidirectional

¹The terms 360°, omnidirectional, spherical, equirectangular are equivalently used in this document.

content is becoming increasingly popular and more easily produced. This expanded the usage of spherical sensors in a number of fields, such as Virtual Reality (VR) [33, 14], indoor navigation [35], or even real-estate.

In this work we train a deep CNN on the task of single image 360° surface normal estimation. We address the lack of sufficient training data by generating a novel dataset of 360° indoors scenes with their corresponding ground truth surface annotations by rendering existing 3D datasets. The dataset is publicly available to enable further research in 360° visual perception².

Inspired by the simplicity and numerical stability of quaternions when representing rotations, we train a deep CNN to predict 360° surface normal maps, by utilizing a novel loss function defined on the hyper-sphere using quaternions to express angular differences. Our experimental results (Table 1) show additional performance boost compared to models trained with losses commonly used in similar regression tasks.

Additionally, we compare with other state of the art normal estimation methods trained on planar images by inferring their predictions on equirectangular as well as cubemap projections of our dataset’s images. Finally, we present promising qualitative results of our network applied on completely unseen challenging samples of the Sun360 dataset [57], and further present the feasibility of our model for a 360° image-relighting application.

2. Related Work

Since the goal of this work is to learn surface normal from a single 360° image, and to the best of our knowledge, similar work does not exist, we first present learning-based methods for the 360° domain, followed by similar work done on planar 2D datasets.

2.1. Learning on 360° images

The 360° field of view of spherical images benefits many applications, such as autonomous driving [46], robotics [42] or VR [22]. Typically, omnidirectional images are modeled as a sphere, and its pixel coordinates map to the longitudinal and latitudinal spherical coordinates. Despite their advantages, omnidirectional content suffers from distortion, especially near the sphere’s poles, making it very difficult to process them with typical CNN architectures. Nowadays, the most usual ways to apply neural network pipelines on spherical input are either employing standard CNN architectures and run their predictions directly on the projected (typically equirectangular [50]) image, or projecting the image to the faces of a cube (cubemap) and then back-projecting them to equirectangular. However, there are a number of efforts

²vc13d.github.io/HyperSphereSurfaceRegression

that model the distortion of spherical images in the neural network’s architectural processing pipeline.

To address spherical image distortion, many techniques utilized the gnomonic projection [50] to either model equirectangular distortion in the representation of the input data, or to guide convolution kernels’ sampling pattern in order to learn distortion invariant features. In [24] a graph-learning approach for omnidirectional image classification is presented. The graph representing the image is constructed using the gnomonic projection and a method for designing convolutional kernels to have similar responses for the same patterns in the image regardless of the position and lens distortion is proposed. In a similar manner, a distortion aware convolution kernel sampling pattern is presented in [53], which models the distortion in spherical images. The convolution kernels sample equirectangular images w.r.t. the gnomonic projection, and thus can be used with models trained on regular 2D images. SphereNet [10], a framework for learning spherical image representations uses the same kernel sampling pattern, further boosting its computational performance by additionally sampling uniformly on the sphere using the method described in [44].

Other efforts, try to model equirectangular image distortion with more typical neural network architectures like [52], where the authors focus on learning to transfer trained 2D models to the spherical domain, by adjusting their network’s kernel sizes w.r.t. to the latitudinal angle and enforcing consistency between the predictions of the 2D projected views and those in the 360° image.

Additionally, there is limited work addressing 3D perception problems on the omnidirectional domain, such as [54], in which the authors follow the steps of [62], to learn depth and camera motion from 360° videos, using two networks; one for inferring depth and the other for predicting the camera pose. They train their networks on cubemap projections of 360° video sequences rendered from the SunCG [51] dataset. Moreover, in [64], the authors use an end-to-end approach to learn 360° depth from equirectangular indoors scenes. They present a dataset generated via rendering existing 3D datasets and two neural network architectures, one more typical and the other constructed with rectangular filters and dilated convolutions [59] to account for the distortion in the spherical domain.

Finally, in [15] the authors focus on the task of learning a 3D room layout from a single 360° image, using the edges that are formed from wall-ceiling-floor intersections and their end-points, i.e. their corners, as their ground truth data. They achieve this by introducing equirectangular convolution kernels and a neural network trained on a subset of the Sun360 [57] annotated with ground truth edge and corner data.

2.2. Surface normal estimation from a single image

The use of standard feedforward CNNs to predict a surface normal from a single RGB image has been employed by many recent works. Eigen and Fergus [13] propose a deep learning model for per-pixel regression using a sequence of three scales to generate features and refine predictions in a coarse to fine approach. Their network can be adapted to predict depth, surface normal or semantic segmentation by making small modifications to the architecture.

In a more recent work [61], the authors introduce a synthetic dataset of indoors scenes, generated via physically-based rendering, with ground truth normal annotations, segmentation and object boundary masks. They pre-train a UNet [43] - VGG16 [49] hybrid neural network model on their synthetic dataset and fine-tune it on NYUv2 [48]. A similar network architecture is adopted by [5], that presents an effort to retrieve 3D objects from 2D images. Their neural network is trained to predict surface normals that serve as input to another two-stream network that estimates the pose and the style of the depicted object in order to retrieve the object’s 3D model from a large CAD library [3].

One of the first approaches to propose a non-standard feedforward CNN architecture [56], treats surface normal prediction as a classification problem instead of a regression one, based on [29]. A three-model neural network architecture is presented which comprises a top-down, a bottom-up and a fusion network. The first learns a coarse global normal map and a room layout hypothesis incorporating vanishing point labels under a Manhattan World assumption. The second learns normals for a local patch of the input image and classifies the edges of the depicted scene as convex, concave and occlusion edges. Finally, the latter network fuses the predictions of the two input networks and outputs a final surface normal estimation of the input image.

2.3. Joint normal and depth estimation

As depth and surface normals follow a strong geometric correlation [47], there are a number of methods that concentrate on learning surface normals and depth in a joint manner. Specifically in [55], a four-branch neural network architecture that predicts dense depth and normals along with plane and edge probability maps is presented. The predictions are regularized by a dense conditional random field (DCRF) [26] that encourages the consistency of depth and normals within planar regions and enforces surface predictions to have unit length via the predicted edges and planes.

Li *et al.* [32], use a pre-trained part of AlexNet [27] for depth estimation and VGG16 for surface prediction with non-trainable weights, which they feed with superpixel patches of different sizes sampled from the input image. Their network makes as many predictions as the input patches, which are then concatenated and fed to two fully-

Table 1: Quantitative results of our model trained on our dataset’s train-split and evaluated on our test-split with four different loss configurations. We present the mean, median and root mean square angular error across our dataset’s test-set. We also provide an additional threshold of 5° along with the most commonly used thresholds (11.25° , 22.5° , 30°). \downarrow means lower is better, while \uparrow means higher is better.

Network	Loss	Mean \downarrow	Median \downarrow	RMSE \downarrow	$5^\circ \uparrow$	$11.25^\circ \uparrow$	$22.5^\circ \uparrow$	$30^\circ \uparrow$
VGG16-UNet	L2	7.72	7.23	8.39	73.55	79.88	87.72	90.43
VGG16-UNet	Cosine	7.63	7.14	8.31	73.89	80.04	87.29	90.48
VGG16-UNet	Quaternion	7.24	6.72	7.98	75.8	80.59	87.3	90.37
VGG16-UNet	Quaternion + Smoothness	7.14	6.66	7.88	76.16	80.82	87.45	90.47

connected layers with learnable parameters that produce the final depth (or surface normal) output. As a final refinement step they use a hierarchical CRF that incorporates the relationship between the patches and the pixels of the image.

In [40], Geonet is presented; a two-branch neural network trained to estimate depth and surface normal, using two new modules, the depth-to-normal and normal-to-depth networks, that both use pinhole camera geometry and the prediction of each branch to further refine the quality of their estimations.

An interesting method is presented in [8], where the authors build on top of their previous work [7], and create a dataset by crowd-sourcing the annotation of images collected randomly from Flickr³. They manage to train a neural network to estimate depth and surface normal using relative point-to-point depth and normal annotations evaluating their method on [48].

Finally, the authors in [31] consider fusing two different sources of information other than depth with surface normal, namely optical flow and semantic segmentation, introducing a novel synthetic dataset of outdoor nature scenes, for general scene understanding. They show that joint features efficiency and the complementary refinement of one prediction from the other two, improves object boundaries and region consistency in predictions.

3. Dataset Creation

The data-driven nature of deep CNN architectures is partially addressed with datasets such as [48] and [45], for learning depth or surface normals given scenes captured by the pinhole camera projection model. However, it is difficult to obtain similar datasets of spherical images.

We overcome this limitation by following the steps of [64], and create a mixed dataset of spherical images of indoors scenes. Similarly, we used a path-tracing renderer⁴ and Blender⁵ to render existing 3D datasets and annotate

our rendered images with their corresponding ground truth surface normal maps that are produced as a result of the rendering process.

Specifically, we utilized the same 3D datasets, namely Matterport3D [6], Stanford2D3D [2, 1] and SunCG [51] to generate a dataset composed of a mixture of computer generated (CG) and realistic scenes of indoors spaces. The dataset consists of 24933 unique viewpoints, from which we split 7868 scenes for training, 1098 for validation and 2176 for benchmarking our trained models. We consider the remaining ones as invalid due to inaccuracies during rendering. We provide the dataset publicly to enable further research in 360° visual perception. We showcase a sample of our dataset in Fig. 2.

4. 360° Surface Normals Estimation

Following most background work, we treat training a fully convolutional neural network (FCN) to learn surface normal from a single spherical image as a regression task. In most learning-based normal regression problems the approach is to minimize either the L_2 norm [32, 40, 5, 12] of the difference of the predicted normal map and the ground truth, or their normalized per-pixel dot-product [13, 61] that implies their angular differences.

Quaternions can represent arbitrary rotations and surface orientation in a very simple and compact form. To train our network, we consider normal vectors as pure quaternions and try to minimize their difference in terms of rotation, showing to further boost the performance of our model (Table 1).

We first formulate our novel quaternion loss function, followed by the description of the neural network architecture used for our experiments.

4.1. Angular loss on the hypersphere

According to Euler’s rotation theorem, a transformation of a fixed point $\mathbf{p}(p_x, p_y, p_z) \in \mathbb{R}^3$ can be expressed as a rotation given by an angle θ around a fixed axis $\mathbf{u}(x, y, z) = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}} \in \mathbb{R}^3$, that runs through \mathbf{p} . This kind of rota-

³<https://www.flickr.com/>

⁴<https://www.cycles-renderer.org/>

⁵<https://www.blender.org/>

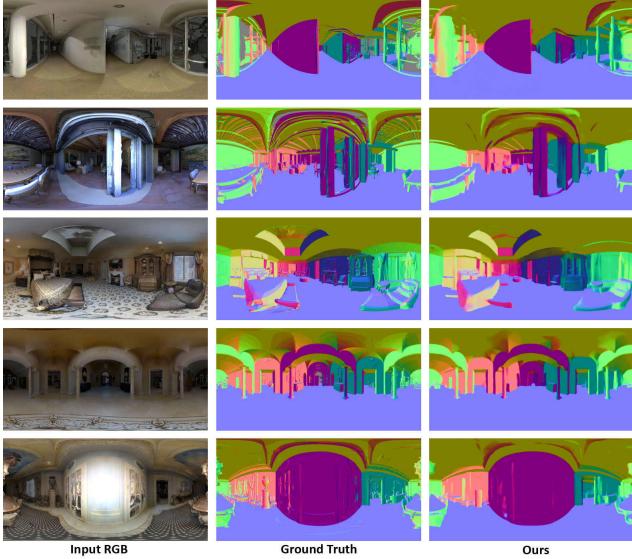


Figure 3: Qualitative results on samples of our test-split. From left to right: input equirectangular image, ground truth surface normal, our model’s prediction.

tion is easily represented by a unit quaternion $\mathbf{q}(w, x, y, z)$, where w is the quaternion’s real part, described by the following formula:

$$\mathbf{q} = e^{\frac{\theta}{2}(x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}})} \implies \quad (1)$$

$$\mathbf{q} = \cos(\theta) + \mathbf{u}\sin(\theta) \quad (2)$$

where $\|\mathbf{q}\| = 1$, and $\cos(\theta)$, $\sin(\theta)$ are the quaternion’s real and imaginary parts respectively.

Thereafter, we can represent two normal vectors $\hat{\mathbf{n}}_1(n_{1_x}, n_{1_y}, n_{1_z})$ and $\hat{\mathbf{n}}_2(n_{2_x}, n_{2_y}, n_{2_z})$ as the pure quaternions $\mathbf{q}_1(0, n_{1_x}, n_{1_y}, n_{1_z})$ and $\mathbf{q}_2(0, n_{2_x}, n_{2_y}, n_{2_z})$. Then, the angular difference between the two normal vectors can be expressed by their transition quaternion [28], which represents a rotation from $\hat{\mathbf{n}}_1$ to $\hat{\mathbf{n}}_2$:

$$\mathbf{t} = \mathbf{q}_1 \mathbf{q}_2^{-1} \quad (3)$$

Because \mathbf{q}_1 and \mathbf{q}_2 are unit quaternions $\mathbf{q}^{-1} = \mathbf{q}^*$, where \mathbf{q}^* is the conjugate quaternion of \mathbf{q} , and $\mathbf{q}^* = -\mathbf{q}$, due to being a pure quaternion, and:

$$\mathbf{q}_1 \mathbf{q}_2^* = \mathbf{q}_1 \cdot \mathbf{q}_2 - \mathbf{q}_1 \times \mathbf{q}_2 \quad (4)$$

Therefore, because \mathbf{q}_1 and \mathbf{q}_2 are pure unit quaternions, their multiplication is reduced to a simple dot (real part) and cross product (imaginary part). As a result, calculating the hypersphere angle represented by the transition quaternion can be straightforwardly implemented in most modern deep learning frameworks.

The rotation angle of the transition quaternion \mathbf{t} and therefore the angular difference between the two normal vectors $\hat{\mathbf{n}}_1$ and $\hat{\mathbf{n}}_2$ is calculated by the inverse tangent between the real and the imaginary parts of the transition quaternion, which are reduced to their dot and cross product, due to being unit quaternions.

$$\tan(\theta) = \frac{\sin(\theta)}{\cos(\theta)} = \frac{\|\mathbf{q}_1 \times \mathbf{q}_2\|}{\mathbf{q}_1 \cdot \mathbf{q}_2} \implies \quad (5)$$

$$\theta = \text{atan}\left(\frac{\|\mathbf{q}_1 \times \mathbf{q}_2\|}{\mathbf{q}_1 \cdot \mathbf{q}_2}\right) \quad (6)$$

In the above computation, the only different operation against other typically used error functions, like the cosine similarity error, is the calculation of a cross product and the $\text{atan}()$ operator (we should note that in our implementation we use the $\text{atan2}()$ operator). However, these kind of operations are simple to implement and are supported by most deep learning frameworks. Additionally, this simplicity makes this loss function practical and with relatively low performance overhead.

Due to imperfect scanning process, we do not consider invalid normals, during back-propagation by generating a mask $M(\mathbf{p})$ at training time with its values being equal to zero for invalid pixels and one for the remaining ones.

Additionally, to further enhance our model’s predictions on fine details and textureless regions, we add a weighted smoothness term (see 4.3) $E_{sm} = \|\nabla \tilde{N}(\mathbf{p})\|_2$ in the final error objective (for more information please refer to the supplementary material).

Finally, we minimize the following error:

$$E(\mathbf{p}) = (1-\alpha)M(\mathbf{p}) \cdot \text{atan}\left(\frac{\|\tilde{N}(\mathbf{p}) \times N(\mathbf{p})\|}{\tilde{N}(\mathbf{p}) \cdot N(\mathbf{p})}\right) + \alpha M(\mathbf{p})\|\nabla \tilde{N}(\mathbf{p})\| \quad (7)$$

4.2. Loss function comparison

In order to understand the efficiency of our novel quaternion loss function we provide 2D and 3D plots of its error landscape comparing them with the ones of the commonly used L_2 and *cosine* loss functions. The 2D plots are presented in Fig 4 while the 3D ones in Fig. 5.

In order to calculate the different error functions’ landscapes we consider a reference vector \vec{n} and calculate the error between all the vectors generated in a 512×256 grid and the reference one, using each compared error function.

The 2D plots in Fig 4 do not provide much information about the nature of the compared error functions. However, it is obvious that the error landscape of the quaternion loss function is more convex than the cosine one. Thus, we can justify the better performance of our network using the

Table 2: Quantitative results against other monocular surface normal estimation models. Rows with *equi* represent feeding the compared models with equirectangular images, and rows with *cube* with cubemap projections following the method we describe in Sec. 5.3.

Network	Mean ↓	Median ↓	$5^\circ \uparrow$	$11.25^\circ \uparrow$	$22.5^\circ \uparrow$	$30^\circ \uparrow$
VGG16-UNet	7.14	6.66	76.16	80.82	87.45	90.47
	41.85	27.76	11.4	31.5	45.2	51.8
Zhang <i>et al.</i> [61]	51.37	38.29	2.7	11.8	31.0	40.8
	26.12	20.83	9.1	26.9	53.3	66.0
Chen <i>et al.</i> [8]	27.10	19.42	6.2	25.9	56.0	68.9

quaternion loss function. In addition, we should note that despite the error landscape of the L_2 loss function appears to have similar convexity with the quaternion one, being just a difference between two values, it does not incorporate the three dimensional nature that is needed to solve 3D learning-based problems.

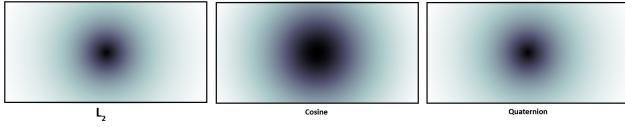


Figure 4: 2D loss error landscape of the commonly used L_2 and cosine loss functions compared with our presented quaternion error function. From left to right: L_2 , cosine and quaternion. (Darker color means higher error).

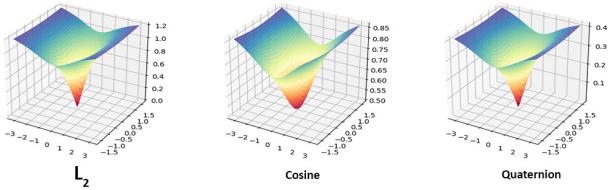


Figure 5: 3D loss error landscape of the commonly used L_2 and cosine loss functions compared with our presented quaternion error function. From left to right: L_2 , cosine and quaternion.

4.3. Smoothness term

Following the works of [21, 19, 64], we use a gradient smoothing term, i.e. an L_1 error on the prediction’s gradients, in the final loss function. As surface normal discontinuities often occur at image gradients, this encourages our network’s predictions to be locally smooth, penalizing gradients that may wrongly occur from the texture of the input image. We consider the weighting term α as a hyper-parameter, and experimentally choose $\alpha = 0.025$ in our

final implementation. In Table 3 we provide our network’s results on our test-set for range of different α values. Additionally, in Fig. 6 we present qualitative results on a sample of our dataset’s test-split.

From the results in Fig. 6 we can conclude that high α values result in over-smoothed predictions with very little level of detail. Additionally, when α is set to a very low value it has very little impact on the training of the model.

4.4. CNN architecture

Adopting the work of [61, 5], we utilize a fully convolutional (FCN) [34] encoder-decoder network with skip-connections that regresses towards the ground truth surface normals. The network architecture is based on UNet [43] combined with a VGG16 [49] encoder. Despite, training other models used in the literature, their performance was inferior to the selected architecture.

Typically a UNet architecture consists of an encoder that captures the input image’s context, and a symmetrical decoder that enables precise localization. In our implementation, the front-end encoder remains the same as *conv1-conv5* in VGG16, and the decoder is composed of symmetrical blocks of convolutions and bi-linear up-sampling layers. In order to localize the decoder’s upsampled features, we concatenate them with their symmetrical high-resolution features from the encoder via skip-connections. This technique is shown to prevent gradient degradation [20], and proved to be an important element in the network’s design. Our model outputs high resolution results and keeps fine object details that might otherwise disappear between pooling and up-sampling layers.

Further, we use ReLU [36] as the activation function and batch normalization [23] after each convolutional layer. Finally, the output of the network is fed to a convolution with a 3×3 kernel size to produce the final 3-channel prediction, which we explicitly normalize along the channel dimension.

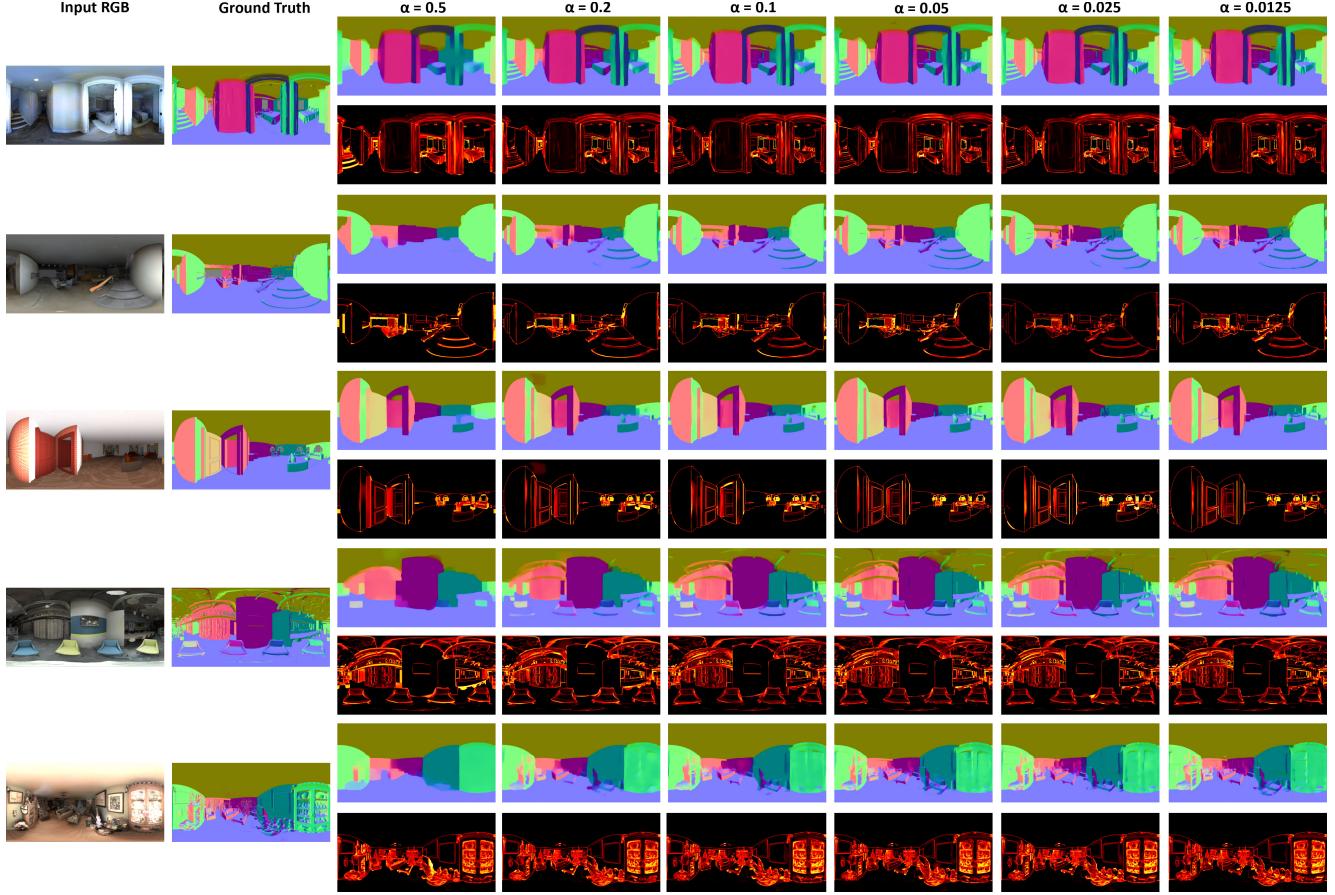


Figure 6: Qualitative results on a sample of our test-split for different loss weighting factors α . From left to right: The input color image, its corresponding ground truth surface normal, predictions for $\alpha = 0.5, 0.2, 0.1, 0.05, 0.025$ and 0.0125 , with their respective error maps underneath each prediction.

Table 3: Quantitative results of our model trained on our dataset’s train-split and evaluated on our test-split for different values of the loss weighting factor α . We also provide an additional threshold of 5° along with the most commonly used thresholds ($11.25^\circ, 22.5^\circ, 30^\circ$). ↓ means lower is better, while ↑ means higher is better.

α	Mean↓	Median↓	RMSE↓	5° ↑	11.25° ↑	22.5° ↑	30° ↑
$\alpha = 0.5$	8.61	8.12	9.41	75.4	79.44	85.4	88.19
$\alpha = 0.2$	7.38	6.84	8.13	76.21	80.62	87.02	89.99
$\alpha = 0.1$	7.2	6.7	7.93	76.18	80.84	87.35	90.36
$\alpha = 0.05$	7.18	6.66	7.91	76.12	80.78	87.38	90.41
$\alpha = 0.025$	7.14	6.66	7.88	76.16	80.82	87.45	90.47
$\alpha = 0.0125$	7.17	6.65	7.89	76.04	80.78	87.4	90.36

5. Experimental Results

This section provides an experimental evaluation of our method. To assess the efficiency of our quaternion loss function, we first train our model using the L_2 norm of the difference of the predicted and the ground truth surface nor-

mal, and additionally, with their normalized per-pixel dot product, i.e. their cosine similarity. We then compare their performance on our dataset’s test split.

We then evaluate its performance compared to other methods applied on cubemap projections of our dataset as well as the original equirectangular images.

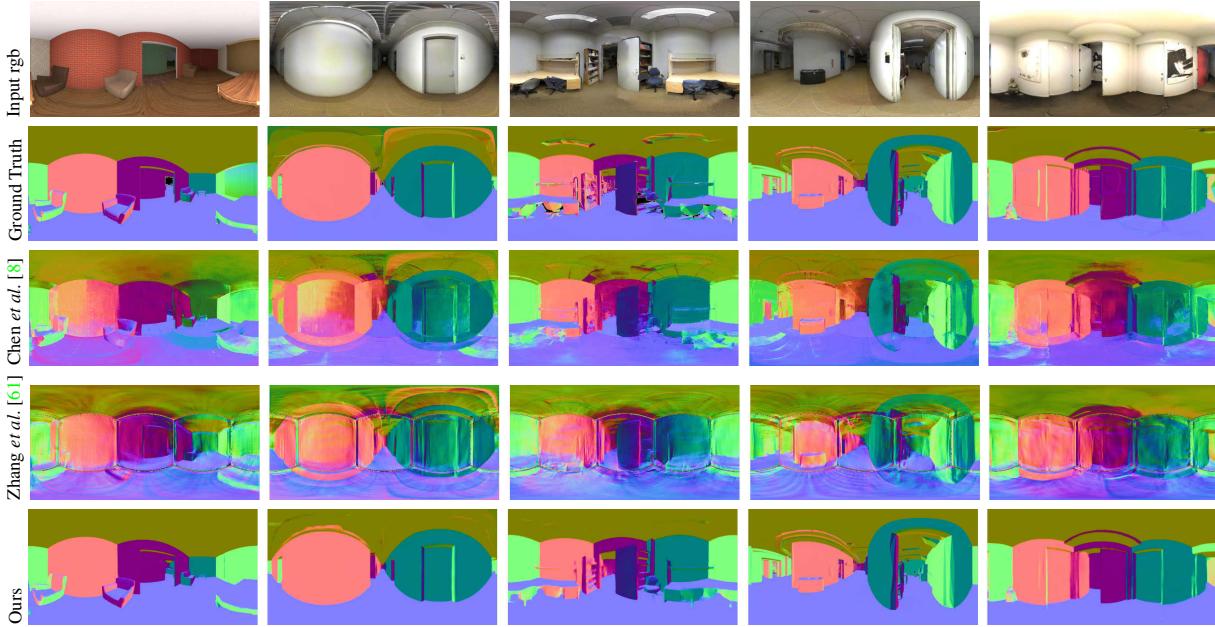


Figure 7: Qualitative comparisons against [8] and [61]. We present results for a sample of our test-split by applying the compared models to cubemaps of our dataset following the method we describe in Sec. 5.3

Additionally, we show the efficacy of our model’s generalization ability, by applying it on a subset of the Sun360 dataset containing unseen indoors scenes. Our trained model produces very promising qualitative results, even on in-the-wild data coming from considerably different distributions from our dataset’s train-split. To further evaluate its effectiveness, we experiment with an image relighting application [41]. We compare relit images using our model’s predictions to relight them, and present qualitative results on samples of our dataset and a subset of Sun360.

5.1. Training Details

All of our networks were implemented and trained using pyTorch [39] framework. Experiments were performed on a PC equipped with an NVIDIA TITAN X GPU, CUDA [37] v9.0 and CuDNN [9] v7.1.3. We used a random seed of 1337 for all of our experiments, for achieving similar training sessions and reproducibility. We initialize our network’s encoder parameters with weights pre-trained on ImageNet [11], and the remaining convolution layers with Xavier weight initialization [18]. We use ADAM [25] as the optimizer with its default parameters $[\beta_1, \beta_2, \epsilon] = [0.9, 0.999, 10^{-8}]$ and a learning rate of 0.0002, and we train all of our models for 50 epochs. We feed every network with equirectangular images at a 512×256 resolution, with the models’ predictions being of equal size. Finally, we use a loss weighting factor $\alpha = 0.025$ between the prediction and the smoothness term.

5.2. Model Performance

To evaluate our results, we use well-established error metrics that are described in the literature, initially introduced in [16]. We measure the mean, median and root-mean-square (RMSE) angular error between the predicted and ground truth normal maps across our dataset’s test split. Furthermore, we present precision coverage errors for three commonly used thresholds, namely 11.25° , 22.5° and 30° and additionally 5° .

Table 1 presents the results of our model evaluated on our test-set when trained under four different loss function configurations, while in Fig. 3 we provide qualitative results of our best performing model. First, we can observe that the models trained with a more intuitive loss function that incorporates geometric understanding, like the cosine similarity or the quaternion loss, have improved performance over the one trained using a generic loss function like the L_2 norm. Additionally, the model trained with our proposed quaternion error outperforms all the others, with the results getting further improved when we add a smoothness term in the loss function.

5.3. Comparison against other methods

To the best of our knowledge, there is no other similar work on monocular 360° surface normal estimation. In an effort to show the importance of training directly on the omnidirectional domain, we provide comparisons of our model with learning-based methods trained on traditional perspec-

Table 4: Quantitative results of our model trained on our dataset’s train-split and evaluated on our test-split compared to the two neural network architectures for omnidirectional monocular depth estimation presented in [64] and the method of [61] re-trained on our dataset’s train-set. We present the mean, median and root mean square angular error across our dataset’s test-set. We also provide an additional threshold of 5° along with the most commonly used thresholds (11.25° , 22.5° , 30°). ↓ means lower is better, while ↑ means higher is better.

Network	Mean↓	Median↓	RMSE↓	5° ↑	11.25° ↑	22.5° ↑	30° ↑
Ours	7.14	6.66	7.88	76.16	80.82	87.45	90.47
UResNet [64]	29.86	30.1	30.21	9.91	25.18	48.84	60.42
RecNet [64]	31.64	31.95	32.1	8.24	21.26	44.19	56.59
Zhang <i>et al.</i> [61]	11.03	10.61	11.72	62.9	73.88	82.74	86.56

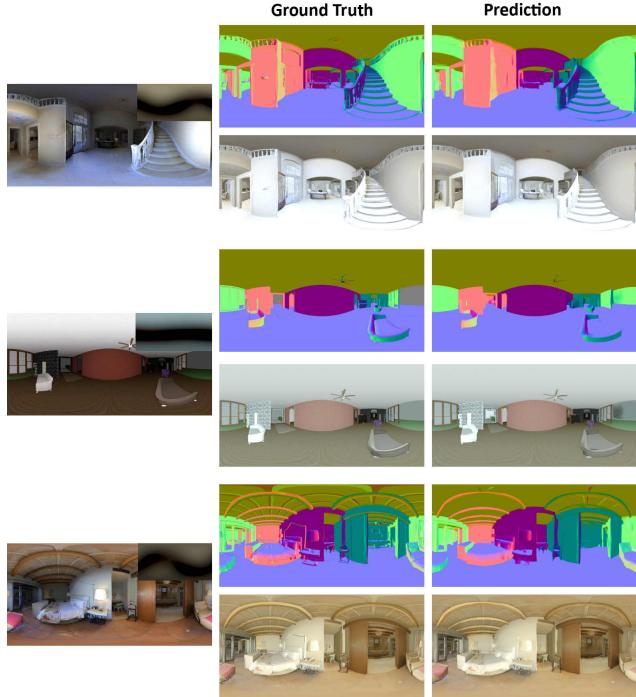


Figure 8: Qualitative comparison of images relit using ground truth normals and our model’s prediction. For each row, the second column presents the ground truth normal map and the relit image, while the second, our model’s prediction and the relit result. In addition, the irradiance map that used to relight both of the images is provided as an inset in the Input rgb image.

tive images. Specifically, we employ [61], which utilizes a similar neural network architecture, and [8] which is trained on a dataset with relative depth and surface normal annotations.

To accomplish this in a fair manner, we follow two schemes. First, we run the predictions of the compared models directly on the equirectangular images of our

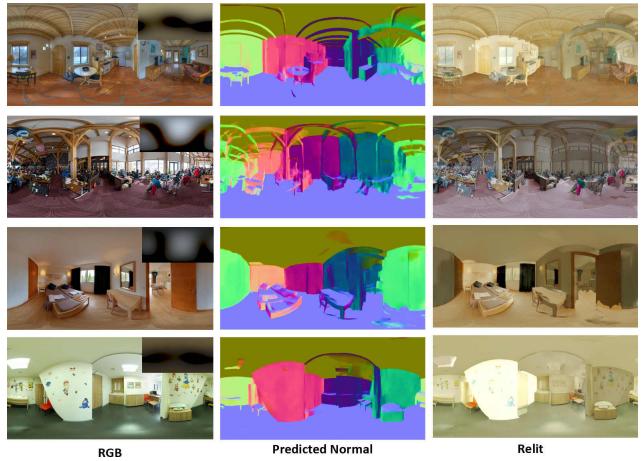


Figure 9: Samples from the Sun360 dataset, relit using our model’s predictions. For every sample we provide the input rgb, our model’s surface prediction and the relit image. The irradiance map used to relight the images is provided as an inset in the input rgb image.

dataset’s test-split, to evaluate how well 2D learned features cope with the distortion on the spherical domain. Moreover, we feed them cubemap projections of spherical images, and exploiting the known rotations between the cube’s faces, we rotate the predicted normal vectors accordingly when back-projecting them to equirectangular. We should note that because our dataset is composed of indoors scenes, the top and bottom faces of the cubemap projections depict only portions of ceiling and floor content respectively. These mostly contain equally textured areas, not sufficient for detecting features in an image. Thus, we do not consider these areas when measuring each model’s performance by masking them in the final error computation.

In addition, our dataset contains floor-aligned camera poses, which is in contradiction with the datasets used to train the compared methods. These datasets contain scenes captured by arbitrary camera poses not necessarily aligned

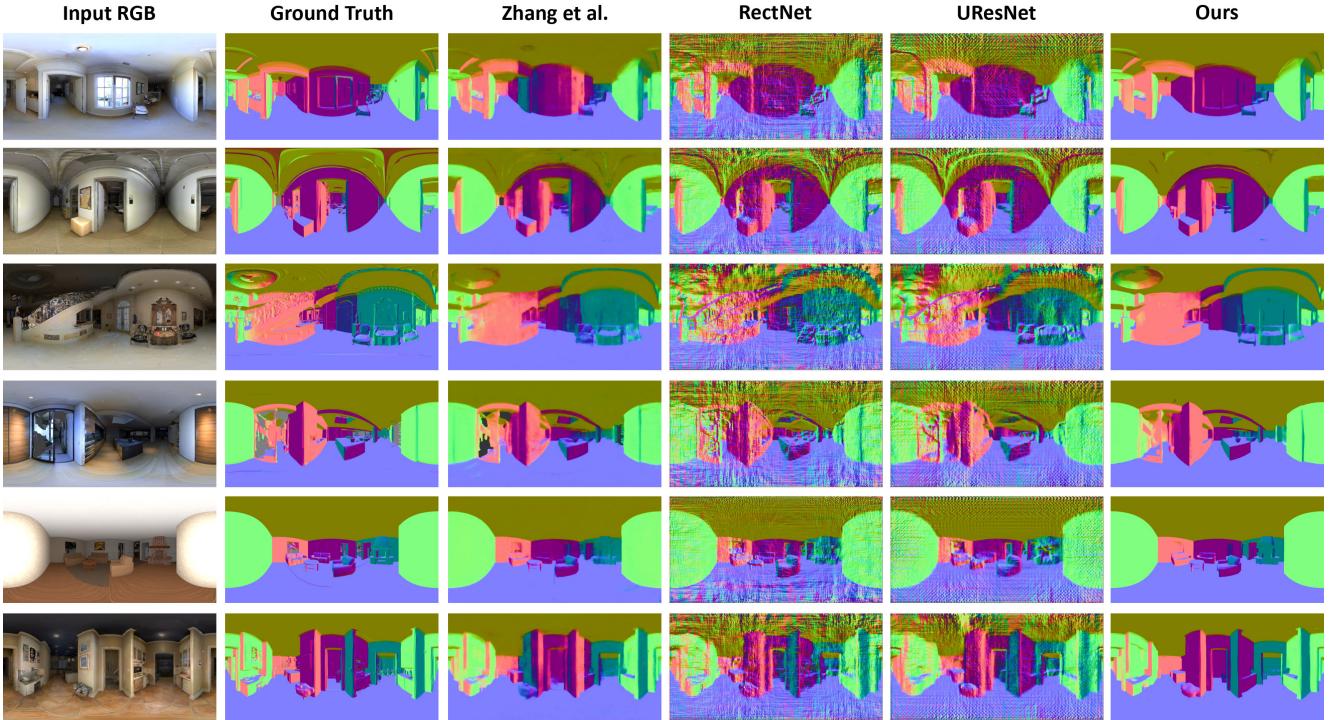


Figure 10: Qualitative comparison on a sample of our test-set. From left to right: Input RGB, ground truth surface normal, Zhang *et al.* [61] re-trained on our test-set, RectNet [64], UResNet [64], our model.

to the floor. Thus, models trained on them would possibly make rightful predictions but unaligned to our dataset’s global orientation. To account for that, we perform singular value decomposition (SVD) between the prediction and the ground truth, and apply the resulting rotation to the prediction, before we calculate their error.

Results of both our evaluation methods are presented in Table 2, and qualitative samples in Fig. 7. When we run the compared models on cubemaps instead of directly on spherical images, both of the networks’ performance is superior. This is expected, as these models are trained on 2D datasets and cannot produce effective features from the characteristics of the distorted equirectangular images. However, we can clearly see discontinuities and inconsistency between each cubemap face. We associate this to the fact that a 90° FOV camera cannot capture global context information required for the models to make consistent predictions.

Finally, we compare with the omnidirectional depth estimation models presented in [64], by running their predictions on our test-set and then converting the resulted 360° depth maps to surface maps. Additionally, we re-train the model of [61] on our train set. Qualitative results are presented in Table 4, with additional qualitative results presented in Fig. 10.

5.4. Surface normal estimation and 360° scene re-lighting

To further evaluate the performance of our model, we additionally experiment with spherical image re-lighting. We examine [41], in which the authors focus on rendering diffuse objects lit from a given environment map. They show that the scene’s irradiance, being a function of the the scene surface normal only, can be approximated in terms of a quadratic polynomial incorporated in the cartesian coordinates of the normal vector, by only 9 spherical harmonic coefficients with an error of only 1%. Specifically, the final relit image is composed of a sum of spherical harmonic basis functions, scaled by the lighting coefficients of the given environment map.

To extract natural spherical harmonic coefficients, we use a dataset of HDR indoors environment maps introduced in [17]. We utilize 9 lighting coefficients for relighting our images, which are later used for estimating an analytic approximation. Finally, the irradiance scaled by each pixel’s intensity produces the output relit image.

We provide qualitative results of images sampled from our test-set in Fig 8, and additionally, in Fig 9, we present samples from Sun360. The first, are relit using both the ground truth normals and predictions of our model, while the second only with our model’s output. Again our network shows promising results, as the differences between the two

reilit images are almost imperceptible, and manifest mostly in highly detailed regions of the image.

6. Conclusion & Discussion

In conclusion, we address the task of monocular 360° surface estimation as a learning problem. To overcome the lack of sufficient training data, we resolve to leveraging 3D rendering to generate spherical images of synthetic (CG) as well as realistic 3D datasets, along with their respective ground truth normal maps and make this dataset publicly available online. In addition, we train a deep CNN to estimate spherical surface normal given a single equirectangular image as input, by employing a simple to implement novel loss function. Our results show better network performance when it is trained with our proposed error function. Furthermore, they demonstrate that when 3D perception is assimilated in the learning objective, neural networks that tackle 3D geometry problems achieve better results. Additionally, we qualitatively present the generalization ability of our trained model via running its predictions on in-the-wild data and using them for an image re-lighting application.

3D perception on spherical media is still considerably unexplored despite their wide utilization. Synthesizing data to circumvent the lack of spherical datasets can be a solution for training neural network models. However, these data will be product of rendering CG or large-scale scanned 3D models, that contain inaccuracies and invalid information. Additionally, it is very difficult to cover a large amount of real-life indoors or outdoors scenes. Accounting for the disadvantages of synthetic data, in the future, we would like to experiment with 3D perception on arbitrary 360° video sequences, employing self-supervised deep neural network models, and additionally model the spherical distortion in the neural network's architecture.

Acknowledgements

This work was supported and received funding from the European Union Horizon 2020 Framework Programme project Hyper360, under Grant Agreement no. 761934. We are also grateful and acknowledge the support of NVIDIA for a hardware donation

References

- [1] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. [4324](#)
- [2] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543, 2016. [4324](#)
- [3] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3762–3769, 2014. [4323](#)
- [4] A. Avetisyan, M. Dahnert, A. Dai, M. Savva, A. X. Chang, and M. Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2019. [4321](#)
- [5] A. Bansal, B. Russell, and A. Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5965–5974, 2016. [4323, 4324, 4326](#)
- [6] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. [4324](#)
- [7] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*, pages 730–738, 2016. [4324](#)
- [8] W. Chen, D. Xiang, and J. Deng. Surface normals in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1557–1566, 2017. [4324, 4326, 4328, 4329](#)
- [9] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer. cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, 2014. [4328](#)
- [10] B. Coors, A. Paul Condurache, and A. Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–533, 2018. [4323](#)
- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. [4328](#)
- [12] T. Dharmasiri, A. Spek, and T. Drummond. Joint prediction of depths, normals and surface curvature from rgb images using cnns. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1505–1512. IEEE, 2017. [4324](#)
- [13] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. [4321, 4323, 4324](#)
- [14] C.-L. Fan, J. Lee, W.-C. Lo, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu. Fixation prediction for 360° video streaming in head-mounted virtual reality. In *Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video, NOSSDAV'17*, pages 67–72, New York, NY, USA, 2017. ACM. [4322](#)
- [15] C. Fernandez-Labrador, J. M. Fcil, A. Perez-Yus, C. Demonceaux, J. Civera, and J. J. Guerrero. Corners for layout: End-to-end layout recovery from 360 images. *arXiv:1903.08094*, 2019. [4323](#)
- [16] D. F. Fouhey, A. Gupta, and M. Hebert. Data-driven 3d primitives for single image understanding. In *Proceedings of the*

- IEEE International Conference on Computer Vision*, pages 3392–3399, 2013. 4321, 4328
- [17] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambarotto, C. Gagné, and J.-F. Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 9(4), 2017. 4330
- [18] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010. 4328
- [19] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017. 4321, 4326
- [20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4326
- [21] P. Heise, S. Klose, B. Jensen, and A. Knoll. Pm-huber: Patchmatch with huber regularization for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2360–2367, 2013. 4326
- [22] H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, and M. Sun. Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1396–1405. IEEE, 2017. 4322
- [23] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4326
- [24] R. Khasanova and P. Frossard. Graph-based classification of omnidirectional images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 869–878, 2017. 4323
- [25] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4328
- [26] P. Krähenbühl and V. Koltun. Efficient nonlocal regularization for optical flow. In *European Conference on Computer Vision*, pages 356–369. Springer, 2012. 4323
- [27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 4323
- [28] J. B. Kuipers et al. *Quaternions and rotation sequences*, volume 66. Princeton university press Princeton, 1999. 4325
- [29] L. Ladický, B. Zeisl, and M. Pollefeys. Discriminatively trained dense surface normal estimation. In *ECCV*, volume 2, page 4, 2014. 4323
- [30] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 4321
- [31] H.-A. Le, A. S. Baslamisli, T. Mensink, and T. Gevers. Three for one and one for three: Flow, segmentation, and surface normals. *arXiv preprint arXiv:1807.07473*, 2018. 4324
- [32] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127, 2015. 4323, 4324
- [33] W.-C. Lo, C.-L. Fan, J. Lee, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu. 360° video viewing dataset in head-mounted virtual reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference, MMSys’17*, pages 211–216, New York, NY, USA, 2017. ACM. 4322
- [34] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4326
- [35] K. Mo, H. Li, Z. Lin, and J.-Y. Lee. The adobeindoornav dataset: Towards deep reinforcement learning based real-world indoor robot visual navigation. 02 2018. 4322
- [36] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 4326
- [37] J. Nickolls, I. Buck, and M. Garland. Scalable parallel programming. In *2008 IEEE Hot Chips 20 Symposium (HCS)*, pages 40–53. IEEE, 2008. 4328
- [38] B. Nuernberger, E. Ofek, H. Benko, and A. D. Wilson. Snapreality: Aligning augmented reality to the real world. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1233–1244. ACM, 2016. 4321
- [39] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 4328
- [40] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 4324
- [41] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’01*, pages 497–500, New York, NY, USA, 2001. ACM. 4328, 4330
- [42] L. Ran, Y. Zhang, Q. Zhang, and T. Yang. Convolutional neural network-based robot navigation using uncalibrated spherical images. *Sensors*, 17(6):1341, 2017. 4322
- [43] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4323, 4326
- [44] E. B. Saff and A. B. Kuijlaars. Distributing many points on a sphere. *The mathematical intelligencer*, 19(1):5–11, 1997. 4323
- [45] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008. 4322, 4324
- [46] D. Scaramuzza and R. Siegwart. Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles. *IEEE transactions on robotics*, 24(5):1015–1026, 2008. 4322

- [47] Y. Shi, K. Xu, M. Nießner, S. Rusinkiewicz, and T. Funkhouser. Planematch: Patch coplanarity prediction for robust rgbd reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–766, 2018. [4323](#)
- [48] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. [4322](#), [4323](#), [4324](#)
- [49] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [4323](#), [4326](#)
- [50] J. P. Snyder and P. M. Voxland. *An album of map projections*. Number 1453. US Government Printing Office, 1989. [4322](#), [4323](#)
- [51] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017. [4323](#), [4324](#)
- [52] Y.-C. Su and K. Grauman. Learning spherical convolution for fast features from 360 imagery. In *Advances in Neural Information Processing Systems*, pages 529–539, 2017. [4323](#)
- [53] K. Tateno, N. Navab, and F. Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 707–722, 2018. [4323](#)
- [54] F. Wang, H. Hu, H. Cheng, J. Lin, S. Yang, M. Shih, H. Chu, and M. Sun. Self-supervised learning of depth and camera motion from 360° videos. *CoRR*, abs/1811.05304, 2018. [4323](#)
- [55] P. Wang, X. Shen, B. Russell, S. Cohen, B. Price, and A. L. Yuille. Surge: Surface regularized geometry estimation from a single image. In *Advances in Neural Information Processing Systems*, pages 172–180, 2016. [4323](#)
- [56] X. Wang, D. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 539–547, 2015. [4321](#), [4323](#)
- [57] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702. IEEE, 2012. [4321](#), [4322](#), [4323](#)
- [58] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. [4321](#)
- [59] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. [4323](#)
- [60] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. [4321](#)
- [61] Y. Zhang, S. Song, E. Yumer, M. Savva, J.-Y. Lee, H. Jin, and T. Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5287–5295, 2017. [4323](#), [4324](#), [4326](#), [4328](#), [4329](#), [4330](#)
- [62] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. [4323](#)
- [63] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE, 2017. [4321](#)
- [64] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 448–465, 2018. [4323](#), [4324](#), [4326](#), [4329](#), [4330](#)