Surface Normal Estimation of *Tilted* Images via Spatial Rectifier

Tien Do, Khiem Vuong, Stergios I. Roumeliotis, and Hyun Soo Park

University of Minnesota
{doxxx104|vuong067|stergios|hspark}@umn.edu

Abstract. In this paper, we present a spatial rectifier to estimate surface normals of tilted images. Tilted images are of particular interest as more visual data are captured by arbitrarily oriented sensors such as body-/robot-mounted cameras. Existing approaches exhibit bounded performance on predicting surface normals because they were trained using gravity-aligned images. Our two main hypotheses are: (1) visual scene layout is indicative of the gravity direction; and (2) not all surfaces are equally represented by a learned estimator due to the structured distribution of the training data, thus, there exists a transformation for each tilted image that is more responsive to the learned estimator than others. We design a spatial rectifier that is learned to transform the surface normal distribution of a tilted image to the rectified one that matches the gravity-aligned training data distribution. Along with the spatial rectifier, we propose a novel truncated angular loss that offers a stronger gradient at smaller angular errors and robustness to outliers. The resulting estimator outperforms the state-of-the-art methods including data augmentation baselines not only on ScanNet and NYUv2 but also on a new dataset called Tilt-RGBD that includes considerable roll and pitch camera motion.

Keywords: Surface normal estimation; spatial rectifier; tilted images

1 Introduction

We live within structured environments where various scene assumptions, e.g., Manhattan world assumption [4], can be reasonably applied to model their 3D scene geometry. Nonetheless, the motion of daily-use cameras that observe these scenes is rather unrestricted, generating diverse visual data. For instance, *embodied sensor measurements*, e.g., body-mounted cameras and agile robot-mounted cameras that are poised to enter our daily spaces, often capture images from tilted camera placement (i.e., non-zero roll and pitch). Understanding scene geometry, e.g., surface normal, from such non-upright imagery can benefit numerous real-world applications such as augmented reality [1,14], 3D reconstruction [23,27,28,30,34,39], and 3D robot navigation.

Existing learning-based frameworks, however, exhibit limited capability to predict surface normals of images taken from cameras with arbitrary pointing
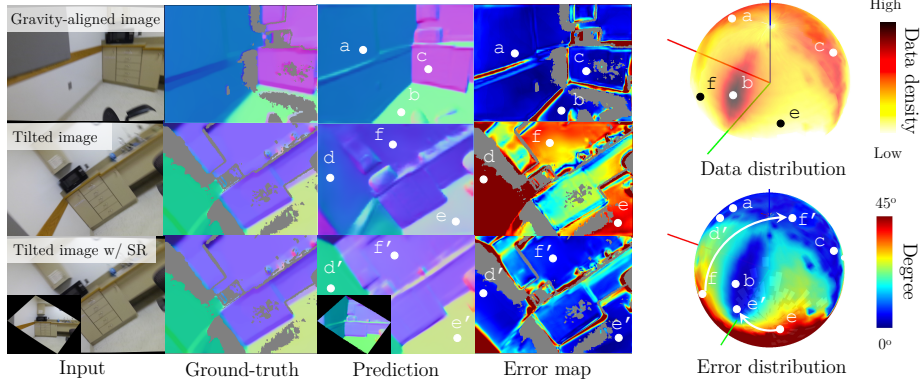
Fig. 1: We present a spatial rectifier that is designed to learn to warp a tilted image for surface normal estimation. While a state-of-the-art surface normal estimator [14] produces accurate prediction on an upright image (top row), it performs poorly on a tilted image (middle row) because of the bias in the surface normal distribution of the training data. Data distribution illustrates the density of surface normals in the ScanNet training dataset, which is highly correlated with the error distribution. We leverage the spatial rectifier to transform the surface normal in erroneous regions to nominal orientation regions, i.e., (d,e,f)→(d',e',f') in error distribution, allowing accurate surface normal prediction (bottom row).

directions because not all scene surfaces are equally represented in deep neural networks. Figure 1 shows the error distribution of surface normals predicted by a state-of-the-art network FrameNet [14] after being trained on the ScanNet dataset [5]. Albeit accurate on an upright image (regions `a`, `b`, and `c` in the top row), the prediction on a tilted image (regions `d`, `e`, and `f` in the middle row) is highly erroneous. This performance degradation is mainly caused by a domain gap between the training and testing data, i.e., the training data is predominantly composed of gravity-aligned images while the testing tilted images have large roll and pitch angles. This poses a major limitation for surface normal estimation on embodied sensor measurements using the state-of-the-art estimator.

We address this challenge by leveraging a *spatial rectifier*. Our main hypothesis is that it is possible to learn a global transformation that can transform a tilted image such that its surface normal distribution can be matched to that of gravity-aligned images. For instance, a homography warping induced by a pure 3D rotation can warp image features such that the surface normals in the underrepresented regions are transformed to the densely distributed regions (d,e,f)→(d',e',f') as shown in Figure 1. We design the spatial rectifier parameterized by the gravity direction and a principle direction of the tilted image that maximizes a distribution match, i.e., minimizes the domain gap between

the tilted image and the upright training data. The spatial rectifier is learned by synthesizing tilted images using a family of homographies induced by the gravity and principle directions in an end-to-end fashion. We demonstrate that the spatial rectifier is highly effective for surface normal estimation in embodied sensor data, outperforming the state-of-the-art methods including data augmentation baselines not only on ScanNet [5] and NYUv2 [25] but also a new dataset called Tilt-RGBD with a large roll and pitch camera motion. The bottom row of Figure 1 illustrates surface normal estimation with the spatial rectifier applied to a tilted image where the estimation error is substantially lower than the one without the spatial rectifier (the middle row).

Further, we integrate two new design factors to facilitate embodied sensor measurements in practice. (1) Robustness: there exists a mismatch between the loss used in existing approaches and their evaluation metric. On the one hand, due to the noisy nature of ground truth surface normals [13], the evaluation metric relied on measures in the angle domain. On the other hand, L_2 loss or cosine similarity between 3D unit vectors is used to train the estimators for convergence and differentiability. This mismatch introduces sub-optimal performance of their normal estimators. We address this mismatch by introducing a new loss function called *truncated angular loss* that is robust to outliers and accurate in the region of small angular error below 7.5°. (2) Resolution vs. context: Learning surface normal requires accessing both global scene layout and fine-grained local features. Hierarchical representations, such as the feature pyramid network [11, 17], have been used to balance the representational expressibility and computational efficiency while their receptive fields are bounded by the coarsest feature resolution. In contrast, we propose a new design that further increases the receptive fields by incorporating dilated convolutions in the decoding layers. The resulting network architecture significantly reduces the computation (70% FLOPS reduction as compared to the state-of-the-art DORN [11]) while maintaining its accuracy. This allows us to use a larger network (e.g., ResNeXt-101; 57% FLOPS reduction) that achieves higher accuracy and faster inference (45 fps).

To the best of our knowledge, this is the first paper that addresses the challenges in surface normal estimation of tilted images without external sensors. Our contributions include: (1) A novel spatial rectifier that achieves better performance by learning to warp a tilted image to match the surface normal distribution of the training data; (2) A robust angular loss for handling outliers in the ground truth data; (3) An efficient deep neural network design that accesses both global and local visual features, outperforming the state-of-the-art methods in computational efficiency and accuracy; and (4) A new dataset called Tilt-RGBD that includes a large variation of roll and pitch camera angles measured by body-mounted cameras.

Related Work	Training data	Testing data	Loss	Backbone	Median (°) ↓	11.25° ↑
Li et al. [21]	NYUv2	NYUv2	L_2	AlexNet	27.8	19.6
Chen et al. [3]			L_2	AlexNet	15.78	39.17
Eigen and Fergus [7]			L_2	AlexNet/VGG	13.2	44.4
SURGE [31]			L_2	AlexNet/VGG×4	12.17	47.29
Bansal et al. [1]			L_2	VGG-16	12.0	47.9
GeoNet [27]			L_2	Deeplabv3	11.8	48.4
FrameNet [14]	ScanNet		L_2	DORN	11.0	50.7
VPLNet [32]			AL	UPerNet	9.83	54.3
FrameNet [14]	ScanNet	ScanNet	L_2	DORN	7.7	62.5
Zeng et al. [37]			L_1	VGG-16	7.47	65.65
VPLNet [32]			AL	UPerNet	6.68	66.3

Table 1: Surface normal estimation methods and performance comparison. The median measures the median of angular error and 11.25° denotes the percentage of pixels whose angular errors are less than 11.25° (↓ / ↑: the lower/higher the better). Note that Zeng et al. [37] uses RGBD as an input. AL stands for angular loss $(\cos^{-1}(\cdot))$.

2   Related Work

This paper includes three key innovations to enable accurate and efficient surface normal estimation for tilted images: Spatial rectification, robust loss, and network design. We briefly review most related work regarding datasets, accuracy, loss function, and network designs summarized in Table 1.

Gravity Estimation and Spatial Rectification Estimating gravity is one the fundamental problems in visual scene understanding. Visual cues such as vanishing points in indoor scene images [20,24] can be leveraged to estimate the gravity from images without external sensors such as an IMU. In addition, learning-based methods have employed visual semantics to predict gravity [9,26,35]. This gravity estimate is, in turn, beneficial to recognize visual semantics, e.g., single view depth prediction [8,29]. In particular, a planar constraint enforced by the gravity direction and visual semantics (segmentation) is used to regularize depth prediction on KITTI [8]. Furthermore, a monocular SLAM method is used to correct in-plane rotation of images, allowing accurate depth prediction [29]. Unlike existing methods that rely on external sensors or offline gravity estimation, we present a spatial rectifier that can be trained in an end-to-end fashion in conjunction with the surface normal estimator. Our spatial rectifier, inspired by the spatial transformer networks [15], transforms an image by a 3D rotation parameterized by the gravity and the principle directions.

Robust Loss There is a mismatch between the training loss and the evaluation metric for surface normal estimation. Specifically, while the cosine similarity or L_2 loss is used for training [1, 3, 7, 13, 14, 21, 27, 31, 37–39][1], the evaluation metric is, in contrast, based on the absolute angular error, e.g., L_1 or L_0. This

---

[1] Classification-based surface normal learning is not considered as its accuracy is bounded by the class resolution [19,33].

mismatch causes not only slow convergence but also sub-optimality. Notably, Zeng et al. [37] and Liao et al. [22] proposed to use a L_1 measure on unit vectors and spherical regression loss, respectively, to overcome the limitations of L_2 loss. More recently, UprightNet [35] and VPLNet [32] employed an angular loss (AL), and showed its high effectiveness in both gravity and surface normal predictions, respectively. In this work, we propose a new angular loss called the truncated angular loss that increases robustness to outliers in the training data.

Efficient Network Design Surface normal estimation requires to access both global and local features. Modern designs of surface normal estimators leverage a large capacity to achieve accurate predictions at high resolution. For instance, FrameNet [14] employs the DORN [11] architecture, a modification of DeepLabv3 [2] that removes multiple spatial reductions (2×2 max pool layers) to achieve high resolution surface normal estimation at the cost of processing a larger number of activations. Recently, new hierarchical designs (e.g., panoptic feature pyramid networks [17]) have shown comparable performance by accessing both global and local features while significantly reducing computational cost in instance segmentation tasks. Our design is inspired by such hierarchical designs through atrous convolution in decoding layers, which outperforms DORN on acccuracy with lower memory footprint, achieving real-time performance.

3   Method

We present a new spatial rectifier to estimate the surface normals of tilted images. The spatial rectifier is jointly trained with the surface normal estimator by synthesizing randomly oriented images.

3.1   Spatial Rectifier

Given a tilted image $\mathcal{I}$, a surface normal estimator takes as input a pixel coordinate $\mathbf{x} \in \mathcal{R}(\mathcal{I})$ and outputs surface normal direction $\mathbf{n} \in \mathbb{S}^2$, $f_\phi(\mathbf{x}; \mathcal{I}) = \mathbf{n}$, where $\mathcal{R}(\mathcal{I})$ is the coordinate range of the image $\mathcal{I}$. The surface normal estimator is parameterized by ϕ.

Consider a spatial rectifier:

$$\overline{\mathcal{I}}(\mathbf{x}) = \mathcal{I}(W_{\mathbf{g},\mathbf{e}}(\mathbf{x})),$$

where $W_{\mathbf{g},\mathbf{e}}$ is the spatial rectifier that warps a tilted image $\mathcal{I}$ to the rectified image $\overline{\mathcal{I}}$, parameterized by the unit vectors $\mathbf{g}$ and $\mathbf{e}$ expressing the gravity and the principle directions, respectively, in the tilted image coordinate system (Figure 2). We define the spatial rectifier using a homography induced by a pure rotation mapping from $\mathbf{g}$ to $\mathbf{e}$, i.e., $W_{\mathbf{g},\mathbf{e}} = \mathbf{KRK}^{-1}$ where $\mathbf{K}$ is the matrix comprising the intrinsic parameters of the image $\mathcal{I}$. The rotation $\mathbf{R}$ can be
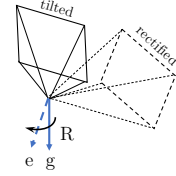

Fig. 2: Geometry.

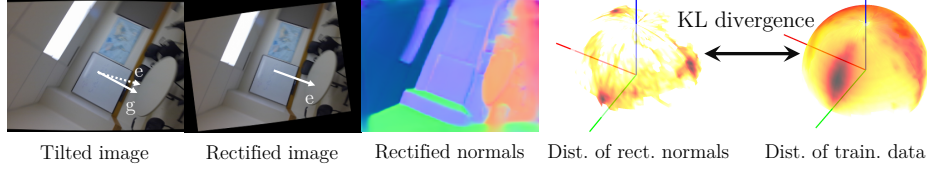| Tilted image | Rectified image | Rectified normals | Dist. of rect. normals | Dist. of train. data |

Fig. 3: Spatial rectifier uses the tilted image to synthesize the rectified image by maximizing the surface normal distribution match. The rectified surface normal distribution is matched with that of the gravity aligned images.

written as:

$$\mathbf{R}(\mathbf{g}, \mathbf{e}) = \mathbf{I}_3 + 2\mathbf{e}\mathbf{g}^\mathsf{T} - \frac{1}{1 + \mathbf{e}^\mathsf{T}\mathbf{g}} (\mathbf{e} + \mathbf{g})(\mathbf{e} + \mathbf{g})^\mathsf{T}, \tag{1}$$

where $\mathbf{I}_3$ is the 3×3 identity matrix.

The key innovation of our spatial rectifier is a learnable $\mathbf{e}$ that transforms the surface normal distribution from an underrepresented region to a densely distributed region in the training data, i.e., (d,e,f)→(d',e',f') in Figure 1. In general, the principle direction determines the minimum amount of rotation of the tilted image towards the rectified one whose surface normals follow the distribution of the training data. A special case is when $\mathbf{e}$ is aligned with one of the image's principle directions, e.g., the y-axis or $\mathbf{e} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^\mathsf{T}$, which produces a gravity-aligned image. We hypothesize that this principle direction and the gravity can be predicted by using the visual semantics of the tilted image:

$$h_{\boldsymbol{\theta}}(\mathcal{I}) = \begin{bmatrix} \mathbf{g}^\mathsf{T} & \mathbf{e}^\mathsf{T} \end{bmatrix}, \tag{2}$$

where $h_{\boldsymbol{\theta}}$ is a learnable estimator parameterized by $\boldsymbol{\theta}$.

We compute the ground truth of the principle direction $\mathbf{e}$ that maximizes the area of visible pixels in the rectified image while minimizing the KL divergence [18] of two surface normal distributions:

$$\mathbf{e}^* = \underset{\mathbf{e}}{\arg\min} \; D_{KL}\left(P(\mathbf{e})\|Q\right) + \lambda_{\mathbf{e}} V\left(\overline{\mathcal{I}}\left(W_{\mathbf{g},\mathbf{e}}\right)\right), \tag{3}$$

where Q is the discretized surface normal distribution from all training data and $P(\mathbf{e})$ is the surface normal distribution of the rectified image given $\mathbf{e}$ (Figure 3). Minimizing the KL divergence allows us to find a transformation from the tilted image to the rectified image ($\mathbf{g} \rightarrow \mathbf{e}$), the surface normal distribution of which optimally matches the training data. V counts the number of invisible pixels in the rectified image:

$$V(\overline{\mathcal{I}}) = \int_{\mathcal{R}(\overline{\mathcal{I}})} \delta\left(\overline{\mathcal{I}}(\mathbf{x})\right) \mathrm{d}\mathbf{x}, \tag{4}$$

where $\delta(\cdot)$ is the Kronecker delta function that outputs one if the pixel at $\mathbf{x}$ is invisible, i.e., $\overline{\mathcal{I}}(\mathbf{x}) = 0$, and zero otherwise. This ensures a minimal transformation
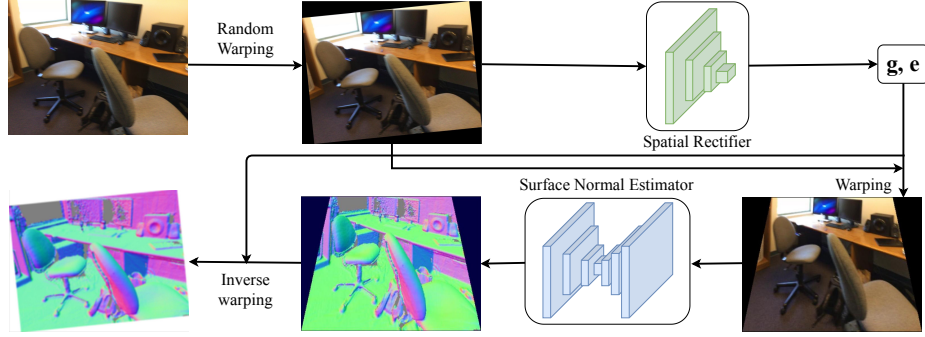
Fig. 4: We jointly train the spatial rectifier and surface normal estimator by randomly synthesizing tilted images. A training image is warped to produce a tilted image. The rectifier is learned to predict $\mathbf{g}$ and $\mathbf{e}$ in the tilted image coordinate system. This allows us to warp the tilted image to the rectified image whose surface normal distribution matches that of the training data. The surface normal estimator is used to predict the rectified surface normals and warp back to the tilted surface normals.

of the tilted image, facilitating learning surface normals from the rectified image. The scalar factor $\lambda_{\mathbf{e}}$ balances the relative importance of distribution matching and visibility. Figure 3 illustrates the optimal $\mathbf{e}$ that maximizes the surface normal distribution match between the rectified image and the gravity-aligned training images (e.g., ScanNet [5]).

3.2   Surface Normal Estimation by Synthesis

We learn the parameters of the neural networks ($\boldsymbol{\phi}$ and $\boldsymbol{\theta}$) by minimizing the following loss:

$$\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta}) = \sum_{\mathcal{D}} \sum_{\mathbf{x}\in\mathcal{R}(\mathcal{I})} \mathcal{L}_{\text{TAL}}(\mathbf{n}_{\mathbf{x};\mathcal{I}}, \widehat{\mathbf{n}}_{\mathbf{x};\mathcal{I}}) + \lambda\mathcal{L}_{\text{TAL}}(h_{\boldsymbol{\theta}}(\mathcal{I}), \left[\widehat{\mathbf{g}}^{\mathsf{T}}\ \widehat{\mathbf{e}}^{\mathsf{T}}\right]),$$

where $\mathcal{D}$ is the training dataset. For each training image, we synthesize a tilted image $\mathcal{I}$ and its corresponding ground truth surface normals $\widehat{\mathbf{n}}_{\mathbf{x};\mathcal{I}}$, gravity $\widehat{\mathbf{g}}$, and principle axis $\widehat{\mathbf{e}}$. λ controls the importance between the surface normal and the spatial rectification. $\mathcal{L}_{\text{TAL}}$ is the truncated angular loss (Section 3.3). The predicted surface normals of the tilted image $\mathbf{n}_{\mathbf{x};\mathcal{I}}$ is obtained by transforming the rectified one $f_{\boldsymbol{\phi}}(\mathbf{x};\overline{\mathcal{I}})$ as follows:

$$\mathbf{n}_{\mathbf{x};\mathcal{I}} = \mathbf{R}_{\mathbf{g},\mathbf{e}}^{\mathsf{T}} f_{\boldsymbol{\phi}}(\mathbf{W}_{\mathbf{g},\mathbf{e}}^{-1}(\mathbf{x});\overline{\mathcal{I}}),$$

where $\mathbf{R}_{\mathbf{g},\mathbf{e}}$ and $\mathbf{W}_{\mathbf{g},\mathbf{e}}$ are defined in Section 3.1. Note that since $f_{\boldsymbol{\phi}}$ takes rectified images as inputs, it does not require a large capacity network to memorize all possible tilted orientations. Figure 4 illustrates our approach that synthesizes a

tilted image with a random rotation where its predicted parameters ($\mathbf{g}$ and $\mathbf{e}$) are used to rectify the synthesized image. We measure the loss of surface normals by transforming back to the tilted image coordinates.

3.3   Truncated Angular Loss

L_2 loss on 3D unit vectors was used to train a surface normal estimator in the literature [1, 3, 7, 13, 14, 21, 27, 31, 37–39]:

$$\mathcal{L}_{L_2}(\mathbf{n}, \widehat{\mathbf{n}}) = \|\mathbf{n} - \widehat{\mathbf{n}}\|_2^2 = 1 - \mathbf{n}^\mathsf{T}\widehat{\mathbf{n}}$$

where $\mathbf{n}$ and $\widehat{\mathbf{n}}$ are the estimated surface normals and its ground truth, respectively. This measures the Euclidean distance between two unit vectors, which is equivalent to the cosine similarity. A key limitation of this loss is the vanishing gradient around small angular error, i.e., $\partial\mathcal{L}_{L_2}/\partial\xi \approx 0$ if $\xi \approx 0$, where $\xi = \cos^{-1}(\mathbf{n}^\mathsf{T}\widehat{\mathbf{n}})$. As a result, the estimator is less sensitive to small errors.

On the other hand, the angular error or the cardinality of the pixel set with an angular threshold has been used for their evaluation metric to address considerable noise in the ground truth surface normals [13], i.e.,

$$E_{\text{angular}} = \cos^{-1}(\mathbf{n}^\mathsf{T}\widehat{\mathbf{n}}), \quad \text{or}$$
$$E_{\text{card}}(\xi) = |\{\cos^{-1}(\mathbf{n}^\mathsf{T}\widehat{\mathbf{n}}) < \xi\}|, \quad \xi = 11.25°, 22.5°, 30°$$

This mismatch between the evaluation metric and the training loss introduces sub-optimality in the trained estimator. To address this issue, we propose the truncated angular loss (TAL) derived from the angular loss:

$$\mathcal{L}_{\text{TAL}}(\mathbf{n}, \widehat{\mathbf{n}}) = \begin{cases} 0, & 1 - \epsilon \leq \mathbf{n}^\mathsf{T}\widehat{\mathbf{n}} \\ \cos^{-1}(\mathbf{n}^\mathsf{T}\widehat{\mathbf{n}}), & 0 \leq \mathbf{n}^\mathsf{T}\widehat{\mathbf{n}} < 1 - \epsilon \\ \frac{\pi}{2} - \mathbf{n}^\mathsf{T}\widehat{\mathbf{n}}, & \mathbf{n}^\mathsf{T}\widehat{\mathbf{n}} < 0 \end{cases}$$

where $\epsilon = 10^{-6}$. When $\mathbf{n}^\mathsf{T}\widehat{\mathbf{n}} \approx 1$, i.e., as the angular error is close to zero, the loss is clamped zero to avoid the infinite gradient of inverse cosine. More importantly, when $\mathbf{n}^\mathsf{T}\widehat{\mathbf{n}} < 0$, i.e., where the angular error is large (outlier), TAL is linear, i.e., it assigns the constant weight for each $\mathbf{n}^\mathsf{T}\widehat{\mathbf{n}}$ in this region, similar to L_2, whereas AL assigns larger weight for larger $\mathbf{n}^\mathsf{T}\widehat{\mathbf{n}}$, as shown in Figure 5. This makes the training with TAL less sensitive to outliers than AL.
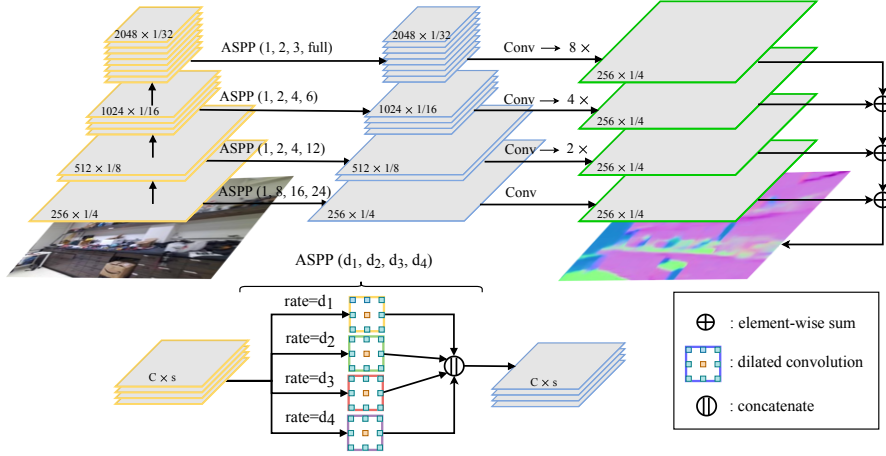

Fig. 5: Truncated angular loss.

Fig. 6: Illustration of the surface normal estimation network architecture.

3.4 Surface Normal Estimator Design

For the surface normal estimator module, we design a new network architecture inspired by the Panoptic FPN [17] that has shown strong performance on high-resolution prediction tasks while maintaining its compact representation. We employ an asymmetric encoder-decoder structure with the following modifications: (i) We do not suppress the number of channels of each pyramid feature immediately to $C = 128$; and (ii) We make use of the atrous spatial pyramid pooling (ASPP) module similar to DeepLabv3 [2].

The network encodes an image with multi-scale feature maps ranging from 1/4 to 1/32 for spatial resolution and 256 to 2048 for channel dimension. For the decoder, we assemble these multi-scale features with ASPP to increase the receptive fields as shown in Figure 6. The decoded feature maps are combined and upsampled to produce a quarter resolution feature maps and summed to predict surface normals.

4 Results

We evaluate the performance of our surface normal estimation quantitatively and qualitatively. All networks used in the evaluation, including ours, take as input 320×240 resolution images and output the same size surface normals. $h_{\boldsymbol{\theta}}$ (Section 3.1) makes use of ResNet-18 backbone. The networks are trained with a batch size of 16 and optimized by the Adam [16] optimizer with a learning rate 10^{-4}. We use an NVIDIA Tesla V100 GPU (with 32GB of memory) to train for 20 epochs and report the best epoch on the corresponding dataset's validation set. Our code, dataset and pretrained networks are available at `https://github.com/MARSLab-UMN/TiltedImageSurfaceNormal`.
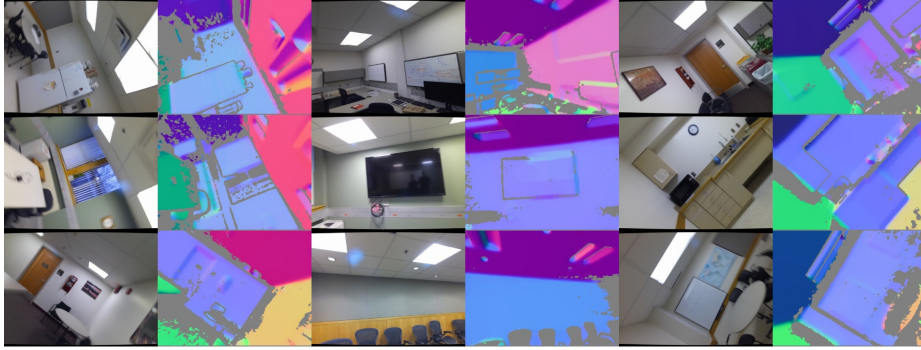
Fig. 7: Selected RGB images and ground truth surface normals from the Tilt-RGBD dataset.

Evaluation Metrics For the quantitative evaluation of surface normal prediction, we employ standard metrics used in [1, 10]: (a) Mean absolute of the error (mean), (b) Median of absolute error (median), (c) Root mean square error (RMSE), and (d) Percentage of pixels with angular error below a threshold ξ with $\xi = 5°$, $7.5°$, $11.25°$, $22.5°$, $30.0°$.

4.1   Evaluation Dataset

We evaluate our method using ScanNet [5], NYUv2 [25], and a new dataset from body-mounted cameras called *Tilt-RGBD* that includes tilted images. All networks are trained and validated by ScanNet and tested on NYUv2 and Tilt-RGBD. No domain adaptation is applied.

ScanNet ScanNet [5] is an indoor RGB-D video dataset that spans a large variation of scenes. We use the images from their standard training/validation/testing splits by extracting every 10th frame: 189,916 images for training, 53,193 images for validation, and 20,942 images for testing. For the ground truth surface normals, we make use of the ground truth generated by FrameNet [14]. For comparison with FrameNet, we make use of the training (199,720)/testing (64,319) split that was originally used in the paper.

NYUv2 NYUv2 [25] was captured by MS Kinect, which includes 654 testing images. The ground truth surface normals are generated by Ladicky et al. [19]. Note that we evaluate only on the valid pixels following Zhang et al. [40].

Tilt-RGBD We collect a new dataset with body-mounted cameras (Azure Kinect) that includes 24 different scenes. Each image instance is associated with the gravity measured by an IMU following [6]. Two types of scenes are collected. (i) *Gravity-aligned images*: We control the camera orientation to be aligned with gravity. 14 scenes and 2,403 images are used. (ii) *Tilted images*: We capture images without controlling the camera's pose; this process generates a large variation of tilted images that includes 14 scenes and 2,428 images. We follow the approach of [19] to generate ground truth surface normals from

Gravity-aligned images	Mean	Median	RMSE	5°	7.5°	11.25°	22.5°	30.0°
FrameNet	11.57	6.59	18.19	37.68	55.69	71.23	86.85	90.93
DORN+TAL	9.54	4.96	16.60	50.33	67.56	79.74	90.24	92.96
DORN+TAL+SR	9.40	5.49	15.76	45.09	65.75	80.71	91.31	93.75
DFPN+TAL	9.74	5.40	16.30	46.05	65.45	78.90	90.40	93.14
DFPN+TAL+AUG	10.19	5.62	17.57	43.89	64.39	79.07	89.95	92.62
DFPN+TAL+IMU	8.62	**4.26**	15.57	**56.90**	**72.42**	82.03	91.17	93.68
DFPN+TAL+SR	**8.59**	4.42	**15.24**	55.80	72.37	**82.26**	**91.39**	**94.01**
Tilted images	Mean	Median	RMSE	5°	7.5°	11.25°	22.5°	30.0°
FrameNet	17.65	10.51	25.38	24.62	38.35	52.25	73.65	81.17
DORN+TAL	14.33	7.09	22.76	37.15	52.08	65.03	80.92	85.79
DORN+TAL+SR	11.77	6.59	18.95	37.00	56.05	72.35	87.29	90.72
DFPN+TAL	14.53	7.84	22.47	32.01	48.30	62.92	81.41	86.43
DFPN+TAL+AUG	12.89	6.86	21.28	34.81	54.28	71.20	85.75	89.10
DFPN+TAL+IMU	11.46	**5.22**	19.88	**48.30**	**62.92**	74.24	86.29	89.75
DFPN+TAL+SR	**11.22**	5.82	**18.88**	43.03	61.19	**75.02**	**87.53**	**90.80**

Table 2: Comparison between all baselines in term of generalization capacity on Tilt-RGBD on gravity-aligned and tilted images.

the depth images. Figure 7 shows representative RGB images along with their ground truth surface normals.

4.2 Baseline

FrameNet FrameNet [14] uses the DORN design that jointly predicts surface normal, tangent, and bitangent directions. We obtain the provided pre-trained network on ScanNet to perform our experiments; **VPLNet** VPLNet [32] incorporates lines and vanishing point directions to improve surface normals prediction, and is currently the state-of-the-art surface normal estimator on ScanNet and NYUv2. Note that we are only able to compare with VPLNet on ScanNet and NYUv2 using the reported results [32] since neither the pretrained network nor the implementation is provided. **DORN+TAL** DORN [11] is a high-capacity network based on the ResNet-101 [12] backbone. We use our truncated angular loss (TAL) to train the network; **DORN+TAL+SR** We train DORN with our proposed spatial rectifier (SR) described in Section 3.2; **DFPN+TAL** Dilated feature pyramid network (DFPN) is our proposed network described in Section 3.4 with our truncated angular loss (TAL). We use DFPN with ResNet-101 backbone; **DFPN+TAL+AUG** We train our DFPN with data augmentation (AUG) by synthesizing tilted images with random rotations. This network does not take rectified images and therefore, is expected to memorize all the tilted representations; **DFPN+TAL+IMU** We train our DFPN with rectified images by assuming the gravity is given (from the ground plane normal direction for ScanNet); **DFPN+TAL+SR (ours)** We train our DFPN with our spatial rectifier (SR) described in Section 3.2.
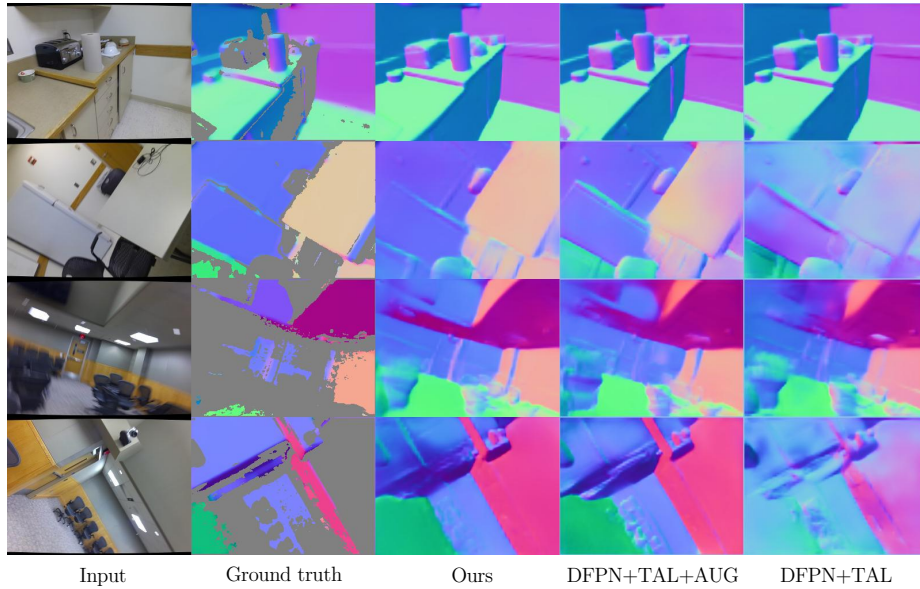
| Input | Ground truth | Ours | DFPN+TAL+AUG | DFPN+TAL |

Fig. 8: Qualitative results: We compare ours (DFPN+TAL+SR) with DFPN+TAL+AUG and DFPN+TAL.

4.3   Surface Normal Estimation on Tilt-RGBD

We compare our method with the baselines on Tilt-RGBD on both gravity-aligned and tilted images summarized in Table 2. We observe that all networks trained on ScanNet perform excellent on unseen gravity-aligned frames indicating that the ScanNet dataset contains sufficient scene diversity. Nevertheless, the best performance in terms of median and tight thresholds ($\xi = 5°, 7.5°$) belongs to DFPN+TAL+IMU. This is due to the fact that gravity direction estimated from the IMU is highly accurate and is employed directly as the surface normal direction of the dominating part of the scene, e.g., floors, ceilings. Our method achieves on-par performance with DFPN+TAL+IMU since the gravity estimate from the network is less accurate.

For tilted images, there is a significant performance degradation in all metrics, e.g., the percentage drop in 11.25° for FrameNet (by 19%), DFPN+TAL (by 16%), DORN+TAL (by 14%) while DORN+TAL+SR, DFPN+TAL+AUG, DFPN+TAL+IMU, and DFPN+TAL+SR show less degradation due to either data augmentation or external sensor information. DFPN+TAL+SR performs significantly better than augmenting the data DFPN+TAL+AUG and on par with DFPN+TAL+IMU while it does not require an external sensor. In addition, DFPN+TAL+SR also outperforms DORN+TAL+SR although DORN has higher capacity than our DFPN (Section 4.4), which suggests that the SR is more compatible with our DFPN. Figure 8 shows our qualitative results where we compare our method with the baselines.

Network	Backbone	ScanNet						NYUv2					
		Mean	Median	RMSE	11.25°	22.5°	30°	Mean	Median	RMSE	11.25°	22.5°	30°
P-FPN+TAL	ResNet-101	15.8	8.0	24.8	61.8	78.8	84.2	17.0	8.9	26.0	57.2	75.6	82.1
DORN+TAL	ResNet-101	15.1	**7.4**	24.1	**63.8**	79.6	84.9	16.6	8.5	25.7	58.5	76.4	82.6
DFPN+TAL	ResNet-101	15.4	7.5	24.6	63.3	79.4	84.7	16.9	8.6	26.2	58.2	76.1	82.2
DFPN+TAL	ResNeXt-101	**15.0**	7.5	**23.9**	63.8	**80.0**	**85.2**	**16.2**	**8.2**	**25.3**	**59.5**	**77.1**	**83.2**

Table 4: Accuracy comparison between different network architectures.

4.4   Network Efficiency

We compare our proposed DFPN+TAL with DORN+TAL in terms of number of floating operations (FLOPS), number of parameters, actual memory consumption, and inference time summarized in Table 3 as well as surface normal estimation accuracy in Table 4. With the ResNet-101 backbone, our DFPN +TAL shows improvement in terms of accuracy over the Panoptic FPN (P-FPN) [17] and performs comparably to DORN+TAL while highly efficient [less than 1/3 of DORN+TAL in terms of FLOPs, memory, and inference time

Network	Backbone	# params	FLOPS	Memory	FPS
P-FPN+TAL	ResNet-101	53.5M	17.3G	6.9GB	68.5
DORN+TAL	ResNet-101	99.5M	97.0G	13.2GB	18.6
DFPN+TAL	ResNet-101	93.0M	29.2G	7.8GB	43.7
DFPN+TAL	ResNeXt-101	137.3M	42.6G	11.5GB	30.0

Table 3: Comparison between different network architectures in terms of FLOPS, number of parameters, memory consumption (in training), and inference time (FPS with batch size of 4).

(FPS)]. With efficient DFPN+TAL, we further increase the network capacity to ResNeXt-101 [36], producing more accurate prediction. It outperforms DORN+TAL in terms of accuracy both on ScanNet and NYUv2 with smaller FLOPS (2.4× faster training time), memory consumption, and realtime inference (30 FPS on NVIDIA GTX 1660).

4.5   Surface normal training loss

We compare three losses used for surface normal estimation: L_2, angular loss (AL), and truncated angular loss (TAL). As shown in Figure 9 and Table 5 (Top), changing from L_2 to AL or TAL leads to faster convergence and significantly lower error metrics. Furthermore, the TAL loss performs on-par with AL at loose thresholds such as 22.5°and 30°and shows improvement in particular at tight thresholds such as 5°and 7.5°, both on ScanNet and NYUv2. In addition, we also show in the Table 5 (Bottom) that our method outperforms FrameNet and VPLNet in all of the metrics both on ScanNet and NYUv2. Note that in order to ensure the fairness in comparison with FrameNet [14] and VPLNet [32], we train and evaluate our method on their modified ScanNet data split. Our results suggest that TAL outperforms the L_2 and AL losses, which allows us to
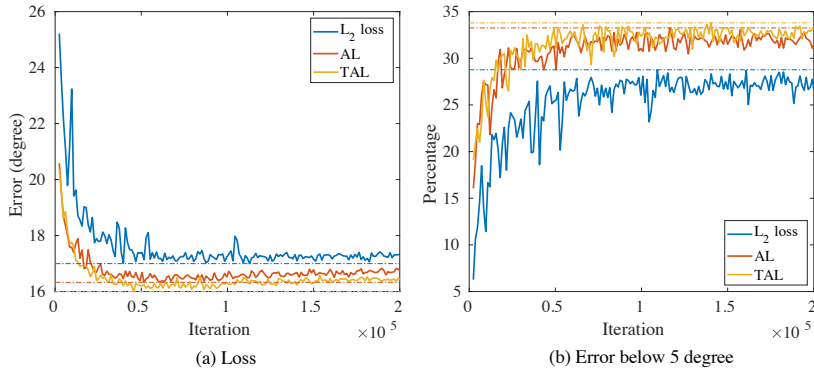
(a) Loss

(b) Error below 5 degree

Fig. 9: (a) Loss values and (b) percentage of pixels with angular error less than 5° over training iterations evaluated on NYUv2. Truncated angular loss (TAL) outperforms L_2 and Angular loss (AL).

Loss/Method	ScanNet								NYUv2							
	Mean	Median	RMSE	5°	7.5°	11.25°	22.5°	30°	Mean	Median	RMSE	5°	7.5°	11.25°	22.5°	30°
L_2	15.7	8.2	24.3	30.4	46.7	61.2	78.7	84.4	17.1	9.4	25.7	28.0	42.1	55.7	75.2	82.0
AL	15.1	7.5	24.0	33.9	50.1	63.8	79.9	85.2	16.7	8.6	25.6	30.9	45.3	58.3	76.6	82.8
TAL	**15.0**	**7.5**	**23.9**	**34.2**	**50.2**	**63.8**	**80.0**	**85.2**	**16.2**	**8.2**	**25.3**	**33.4**	**47.0**	**59.5**	**77.1**	**83.2**
FrameNet [14]	14.7	7.7	22.8	33.4	48.9	62.5	80.1	85.8	18.6	11.0	26.8	23.2	36.6	50.7	72.0	79.5
VPLNet [32]	13.8	6.8	-	-	-	66.3	81.8	87.0	18.0	9.8	-	-	-	54.3	73.8	80.7
DFPN+TAL	**12.6**	**6.0**	**21.1**	**42.8**	**57.5**	**69.3**	**83.9**	**88.6**	**16.1**	**8.1**	**25.1**	**33.6**	**47.3**	**59.8**	**77.4**	**83.4**

Table 5: (Top) Comparison between different loss functions. (Bottom) Comparison with the state-of-the-art methods.

achieve the state-of-the-art performance across different train/test splits while maintaining its generalization capability.

5 Summary

In this paper, we present a new spatial rectifier to estimate surface normals from a tilted image. The spatial rectifier is learned to warp the tilted image to the rectified image such that its surface normal distribution matches that of the training data. We train the spatial rectifier jointly with the surface normal estimator by synthesizing tilted images from the ScanNet dataset. To facilitate practical deployment of surface normal estimation, we design a new efficient network that produces the state-of-the-art accuracy while maintaining low computational burden. Further, we propose a truncated angular loss that overcomes a key limitation of the L_2 loss, resulting in accurate estimation, especially in the region of small angular error. Our method outperforms the state-of-the-art baselines not only on ScanNet and NYUv2 but also on a new Tilt-RGBD dataset that includes large roll and pitch camera motions.

References

1. Bansal, A., Russell, B., Gupta, A.: Marr revisited: 2d-3d alignment via surface normal prediction. In: CVPR (2016)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI (2017)
3. Chen, W., Xiang, D., Deng, J.: Surface normals in the wild. In: CVPR (2017)
4. Coughlan, J.M., Yuille, A.L.: The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In: NeurIPS (2001)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
6. Do, T., Neira, L., Yang, Y., Roumeliotis, S.I.: Attitude tracking from a camera and an accelerometer on gyro-less devices. In: ISRR (2019)
7. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: CVPR (2015)
8. Fei, X., Wong, A., Soatto, S.: Geo-supervised visual depth prediction. RA-L (2019)
9. Fischer, P., Dosovitskiy, A., Brox, T.: Image orientation estimation with convolutional networks. In: GCPR (2015)
10. Fouhey, D.F., Gupta, A., Hebert, M.: Data-driven 3d primitives for single image understanding. In: ICCV (2013)
11. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: CVPR (2018)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
13. Hickson, S., Raveendran, K., Fathi, A., Murphy, K., Essa, I.: Floors are flat: Leveraging semantics for real-time surface normal prediction. arXiv:1906.06792 (2019)
14. Huang, J., Zhou, Y., Funkhouser, T., Guibas, L.J.: Framenet: Learning local canonical frames of 3d surfaces from a single rgb image. In: ICCV (2019)
15. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: NeurIPS (2015)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR. vol. 1412 (2014)
17. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: CVPR (2019)
18. Kullback, S., Leibler, R.: On information and sufficiency. Annals of Mathematical Statistics (1951)
19. Ladický, L., Zeisl, B., Pollefeys, M.: Discriminatively trained dense surface normal estimation. In: ECCV (2014)
20. Lee, J.K., Yoon, K.J.: Real-time joint estimation of camera orientation and vanishing points. In: CVPR (2015)
21. Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: CVPR (2015)
22. Liao, S., Gavves, E., Snoek, C.G.: Spherical regression: Learning viewpoints, surface normals and 3d rotations on n-spheres. In: CVPR (2019)
23. Liu, C., Kim, K., Gu, J., Furukawa, Y., Kautz, J.: Planercnn: 3d plane detection and reconstruction from a single image. In: CVPR (2019)
24. Mirzaei, F.M., Roumeliotis, S.I.: Optimal estimation of vanishing points in a manhattan world. In: CVPR (2011)

25. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
26. Olmschenk, G., Tang, H., Zhu, Z.: Pitch and roll camera orientation from a single 2d image using convolutional neural networks. In: CRV (2017)
27. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: CVPR (2018)
28. Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B., Pollefeys, M.: Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In: CVPR (2019)
29. Saito, Y., Hachiuma, R., Yamaguchi, M., Saito, H.: In-plane rotation-aware monocular depth estimation using slam. In: IW-FCV (2020)
30. Tang, J., Folkesson, J., Jensfelt, P.: Sparse2dense: From direct sparse odometry to dense 3-d reconstruction. RA-L (2019)
31. Wang, P., Shen, X., Russell, B., Cohen, S., Price, B., Yuille, A.L.: Surge: Surface regularized geometry estimation from a single image. In: NeurIPS (2016)
32. Wang, R., Geraghty, D., Matzen, K., Szeliski, R., Frahm, J.M.: Vplnet: Deep single view normal estimation with vanishing points and lines. In: NeurIPS (2020)
33. Wang, X., Fouhey, D., Gupta, A.: Designing deep networks for surface normal estimation. In: CVPR (2015)
34. Weerasekera, C.S., Latif, Y., Garg, R., Reid, I.: Dense monocular reconstruction using surface normals. In: ICRA (2017)
35. Xian, W., Li, Z., Fisher, M., Eisenmann, J., Shechtman, E., Snavely, N.: Uprightnet: Geometry-aware camera orientation estimation from single images. In: CVPR (2019)
36. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR (2017)
37. Zeng, J., Tong, Y., Huang, Y., Yan, Q., Sun, W., Chen, J., Wang, Y.: Deep surface normal estimation with hierarchical rgb-d fusion. In: CVPR (2019)
38. Zhan, H., Weerasekera, C.S., Garg, R., Reid, I.: Self-supervised learning for single view depth and surface normal estimation. arXiv:1903.00112 (2019)
39. Zhang, Y., Funkhouser, T.: Deep depth completion of a single rgb-d image. In: CVPR (2018)
40. Zhang, Y., Song, S., Yumer, E., Savva, M., Lee, J.Y., Jin, H., Funkhouser, T.: Physically-based rendering for indoor scene understanding using convolutional neural networks. In: CVPR (2017)