Egocentric Scene Understanding via Multimodal Spatial Rectifier

Tien Do\textsuperscript{1} \quad Khiem Vuong\textsuperscript{2} \quad Hyun Soo Park\textsuperscript{1}

\textsuperscript{1} University of Minnesota \quad \textsuperscript{2} Carnegie Mellon University

Figure 1. We study the problem of predicting geometry (depths and surface normals) from a single view egocentric image that includes dynamic objects (e.g., hand and people). We design a multimodal spatial rectifier that can effectively handle the excessively tilted images caused by head movement (e.g., nearly 90 degree pitch angle when engaging eye-hand coordination). Our method shows strong performance on unseen images from EPIC-KITCHENS \cite{Sarlin17} (left), FPHA \cite{Wang18} (top right), and our EDINA (bottom right) datasets.

Abstract

In this paper, we study a problem of egocentric scene understanding, i.e., predicting depths and surface normals from an egocentric image. Egocentric scene understanding poses unprecedented challenges: (1) due to large head movements, the images are taken from non-canonical viewpoints (i.e., tilted images) where existing models of geometry prediction do not apply; (2) dynamic foreground objects including hands constitute a large proportion of visual scenes. These challenges limit the performance of the existing models learned from large indoor datasets, such as ScanNet \cite{ScanNet} and NYUv2 \cite{Silberman12}, which comprise predominantly upright images of static scenes. We present a multimodal spatial rectifier that stabilizes the egocentric images to a set of reference directions, which allows learning a coherent visual representation. Unlike unimodal spatial rectifier that often produces excessive perspective warp for egocentric images, the multimodal spatial rectifier learns from multiple directions that can minimize the impact of the perspective warp. To learn visual representations of the dynamic foreground objects, we present a new dataset called EDINA (Egocentric Depth on everyday Indoor Activities) that comprises more than 500K synchronized RGBD frames and gravity directions. Equipped with the multimodal spatial rectifier and the EDINA dataset, our proposed method on single-view depth and surface normal estimation significantly outperforms the baselines not only on our EDINA dataset, but also on other popular egocentric datasets, such as First Person Hand Action (FPHA) \cite{Wang18} and EPIC-KITCHENS \cite{Sarlin17}.

1. Introduction

We interact with surrounding objects in structured yet rather complex, unorganized, and dynamic environments, enabled by our robust egocentric perception that facilitates understanding 3D scene geometry around us. Such innate perceptual ability shows in stark contrast with that of existing computer vision systems, trained to operate on images depicting static and well-organized scenes recorded by carefully controlled cameras \cite{ScanNet, NYUv2, SILBERMAN201212}. These trained models \cite{DIMARCO20182844, Moulin2017} are, despite their remarkable performance, shown to be highly brittle when predicting the scene geometry of egocentric images that observe unscripted everyday activities, including diverse hand-object interactions, captured by in situ embodied sensors such as head/body-mounted cameras \cite{Allievi19}. This requires additional sensors such as IMU and depth sensors in augmented/mixed reality devices (e.g., Hololens and Magic Leap One) to deliver interactive and immersive experiences in our daily spaces.

In this paper, we study a problem of egocentric 3D
scene understanding—predicting depths and surface normals from a single view egocentric image. In addition to challenges of classic scene understanding problems [6], egocentric scene understanding poses two more challenges: (1) Images are no longer upright. Head movements induce significant roll and pitch motions where the scene is often depicted in a tilted way. In particular, by the nature of hand-eye coordination, egocentric images inherently are affected by severe pitch motion when manipulating objects, which is substantially different from the existing data distribution, e.g., ScanNet [6], NYUv2 [36], and KITTI [19]. (2) Images include not only background objects, e.g., furniture, room layout, and walls, but also dynamic foreground objects, e.g., humans and arms/hands (see Figure 1). Classic scene understanding mainly focuses on reconstructing the overall geometric layout made of such background objects while the foreground ones are considered as outliers. In contrast, these foregrounds are more salient in egocentric scenes as they are highly indicative of evolving activities.

We conjecture that the challenges of egocentric scene understanding can be addressed by an image stabilization method that incorporates the fundamentals of equivariance, called spatial rectifier [8]—an image warping that transforms a tilted image to a canonical orientation (i.e., gravity-aligned) such that a prediction model can learn from the upright images. This is analogous to our robust perception aligned) such that a prediction model can learn from the upright images. This is analogous to our robust perception.

Modal spatial rectifier by generalizing the canonical direction, i.e., instead of unimodal gravity-aligned direction, we learn multiple reference directions from the orientations of the egocentric images, which allows minimizing the impact of excessive perspective warping. Our multimodal spatial rectifier makes use the clusters of egocentric images based on the distribution of surface normals into multiple pitch modes, where we learn a geometric predictor (surface normals or depths) that is specialized for each mode to rectify associated roll angles.

To facilitate learning the visual representation of dynamic egocentric scenes, we present a new dataset called EDINA (Egocentric Depth on everyday INdoor Activities). Our dataset comprises 16 hours RGBD recording of indoor activities including cleaning, cooking, eating, and shopping. Our dataset provides a synchronized RGB, depth, surface normal, and the 3D gravity direction to train our multimodal spatial rectifier and geometry prediction models. Our depth and surface normal predictors learned from the EDINA outperform the baseline predictors not only on EDINA dataset but also other datasets, such as EPIC-KITCHENS [7] and First Person Hand Action (FPHA) [18].

Our contributions include: (1) a multimodal spatial rectifier; (2) a large dataset of egocentric RGBD with the gravity that is designed to study egocentric scene understanding, by capturing diverse daily activities in the presence of dynamic foreground objects; (3) comprehensive experiments to highlight the effectiveness of our multimodal spatial rectifier and our EDINA dataset towards depth and surface normal prediction on egocentric scenes.}

2. Related Works

Our egocentric scene understanding lies in the intersection between single view geometry and equivariant spatial rectifier. We briefly review the related work.

**Single View Depth and Surface Normal** Single view scene understanding approaches have shown great progress by leveraging a large amount of data such as ScanNet [6] that supersedes the learning from an image to a 3D scene geometry such as depths [5, 9, 10, 14, 20, 22, 24, 25, 29, 30, 32, 40, 42–45, 57, 59, 60] or surface normals [1, 4, 8, 9, 23, 31, 40, 52, 53, 58]. Existing methods that show remarkable performance on scene understanding tasks have focused on either: (1) designing deep neural network architectures [8, 14, 24]; or (2) exploiting useful 2D visual cues for learning 3D geometry, including textures [23], vanishing points [53], planar surfaces [32, 52], and depth-surface normal consistency [40]. Nevertheless, they are only enabled by large-scale indoor RGBD datasets, such as ScanNet [6], NYUv2 [36], Sun3D [56], and Sun RGBD [50]. However, due to the nature of data collection methods, a model trained on such datasets show a notable performance degradation when applying it to egocentric images because of two reasons: (1) The model has not exposed to the tilted images that have substantially different visual patterns from that of upright images. (2) The model has limited capability of learning dynamic foreground objects that are abundant in egocentric scenes, e.g., hands, pots, pans, vacuums, brooms, pets, and humans. To address these challenges in egocentric images, we make use of a multimodal spatial rectifier, which allows using large existing datasets in conjunction with egocentric datasets.

**Rotation Equivariance** Equivariance is a geometric property of a visual representation: the visual representation in an image must be transformed, according to the transformation of the scene. Enforcing equivariance in learning a scene geometry allows geometrically coherent learned models. To achieve this, camera poses [60] and gravity directions [8] can be employed. For instance, equivariance is used to learn the geometry of scenes by augmenting transformations, i.e., spatial rectifier [8] that rectifies a tilted image to an upright (gravity-aligned) image [8, 48, 49]. Despite the substantial improvement on the tilted images, the spatial rectifier with a unimodal gravity-aligned direction shows poor performance on egocentric images. It is mainly caused
by excessive warping of egocentric images due to a large variation of camera angle, e.g., nearly 90 degree pitch angle when engaging eye-hand coordination. Our multimodal spatial rectifier prevents such excessive perspective warp by predicting multiple reference directions, which significantly improves the egocentric scene understanding task.

**Egocentric Scene Datasets** Egocentric scene datasets have been used for a wide range of tasks such as action recognition [11, 12, 39], action anticipation [2, 46], and many others [15–17]. Notably, Damen et al. [7] proposed EPIC-KITCHENS, a large-scale egocentric benchmark with densely annotated actions and object interactions in the kitchen environment. A few egocentric RGBD datasets that exist were designed for activity recognition [18, 35, 47, 51]. With a few exception, such datasets do not include the 3D gravity direction that is critical for learning an equivariant representation. Our EDINA dataset does not include the 3D gravity direction which in turns, significantly degrades the performance of geometry prediction. We use a multimodal spatial rectifier that warps to multiple reference directions that minimizes the impact of the perspective warping (right).

3. Method

We present a multimodal spatial rectifier that stabilizes tilted images into multiple transformation modes. This method minimizes the impact of perspective warping while retaining equivariance property.

3.1. Equivariant Spatial Rectifier

Consider a function $\Phi : \mathbb{R}^2 \times I \rightarrow \mathbb{R}^n$ that predicts the geometry of a pixel $x \in \mathbb{R}^2$ in an image $I \in I$, where $I = [0, 1]^{H \times W}$ is the image range ($H$ and $W$ are its height and width, respectively). We denote the prediction:

$$y = \Phi(x, I),$$

where $y \in \mathbb{R}^n$ and $n$ is the dimension of the geometry, e.g., $n = 1$ for depth, and $n = 3$ for surface normal.

A spatial rectifier [8] is learned to transform a tilted image $I$ with the gravity direction $g \in S^2$ in the camera coordinate system to the upright image $I_{up}$ with the upright gravity direction $g_{up}$ by explicitly enforcing an equivariant property through 3D rotation (Figure 2):

$$h_{\mathcal{W}} \circ \Phi(x, I) = \Phi(\mathcal{W}(x; R_{up}), I_{up}),$$

where $\mathcal{W} : \mathbb{R}^2 \times SO(3) \rightarrow \mathbb{R}^2$ is a 2D transformation that maps a point in the tilted image to the upright image based on the 3D gravity direction. That is, the transformation can be determined by a homography induced by camera pure rotation $R_{up} \in SO(3)$ such that $\mathcal{W} = R_{up}g$. $I_{up}$ is warped from the tilted image by $I$, i.e., $I_{up} = \mathcal{I}(\mathcal{W}(x; R_{up}))$. $h_{\mathcal{W}}$ is the geometry transformation parametrized by $\mathcal{W}$, e.g., (1) for the surface normal prediction, $h_{\mathcal{W}}$ is equivalent to rotating the surface normal vector ($S^2$), i.e., $h_{\mathcal{W}} \circ \Phi = R_{up}\Phi$; (2) for the depth prediction, $h_{\mathcal{W}}$ is defined as:

$$h_{\mathcal{W}} \circ \Phi = \left( R_{up}K^{-1}\tilde{x} \right)_x \tilde{v}$$

where $(\tilde{v})_x$ denote the 3rd coordinate of a vector $v \in \mathbb{R}^3$, and $K$ is the camera intrinsic matrix, and $\tilde{x} \in \mathbb{P}^2$ is the homogeneous representation of $x$.

Predicting the geometry of a tilted image can be modeled as a function composition:

$$\Phi(x, I) = h_{\mathcal{W}}^{-1} \circ \Phi_{up}(\mathcal{W}(x; R_{up}), I_{up}),$$

where $h_{\mathcal{W}}^{-1}$ is the spatial rectifier, and $\Phi_{up}$ is the geometry predictor learned from upright images. A key benefit of this function composition is that $\Phi_{up}$ can be trained solely by the large training dataset made of the upright images (e.g., ScanNet [6] and NYUv2 [36]), which can be, in turn, used to predict the surface normals of a tilted image.

**Limitation** Despite of its strong performance on tilted images, the spatial rectifier exhibits a major limitation towards egocentric scene understanding due to its single modal rectification. The spatial rectifier is designed to warp a tilted image with respect to a single upright direction, which applies to roll and mild pitch camera rotations. In constrast, egocentric images often have substantial head orientation due to the hand-eye coordination, resulting in severe perspective warped image $I_{up}$ (e.g., $90^\circ$ pitch tilted image), which in turns, significantly degrades the performance of the geometry predictor as shown in Figure 3 (middle).
Surface normal distribution (ScanNet + EDINA)

Multimodal reference direction clustering

Figure 4. Unlike the spatial rectifier [8] that relies on the unimodal surface normal distribution with respect to the gravity direction (left), we present a multimodal spatial rectifier that generalizes the spatial rectifier by learning multiple reference directions (right). As a result, the surface normal distribution of the scene datasets can be decomposed into multiple clusters, which allows minimizing the impact of image warping and more importantly, learning a geometrically coherent representation.

3.2. Multimodal Spatial Rectifier

We generalize the spatial rectifier model by leveraging a mixture of expert models [33] called multimodal spatial rectifier where each expert model predicts the geometry corresponding to a spatial rectification mode:

\[
\Phi(x, I) = \frac{1}{\sum_i b_i} \sum_i b_i (h_{W_i} \circ \Phi_i(W(x; R_i), I)) ,
\]

where \( b_i \in \mathbb{R}_+ \) is a non-negative weight to mix transformations, and \( R_i \) is the rotation that transforms the gravity of the tilted image to the \( i \)-th reference direction, i.e., \( R_i g, I_i \) is warped from the tilted image by \( W_i \), i.e., \( I_i = W_i(I(x; R_i)) \). The reference direction \( r \in S^2 \) is a generalization of the upright gravity \( g_{up} \), which specifies the egocentric images to be warped. \( \Phi_i \) is the geometry predictor designed for the \( i \)-th reference direction. We denote \( W(x; R_i) \) by \( W_i \) by abuse of notation. The key benefit of the multimodal spatial rectifier is the flexibility of image warping. The severe head orientation of an egocentric image can be warped to the closest reference direction, which prevents excessive perspective warping (see Figure 3).

We find the set of reference directions \( \{r_i\}_{i=1}^K \) along the pitch angles by clustering the gravity of egocentric images with \( K \) is the predefined number of the reference directions:

\[
\min_{\{r_i\}_{i=1}^K} \sum_{i=1}^K \sum_{j \in C_i} \|g_j - r_i\|^2_2 ,
\]

where \( C_i \) is the set of the indices of training instances of which gravity directions closest to the \( i \)-th reference direction \( r_i \). In practice, we design an iterative algorithm inspired by K-Medoids algorithm [37] by increasing the number of cluster numbers \( K \) until the total deviation reaches below a threshold \( \delta \) indicating the data is well-fitted (see Algorithm 1). Figure 4 illustrates gravity cluster centers and images as well as their surface normal map belonging to each cluster. Similar to spatial rectifier [8], we represent a 3D rotation by two unit vectors: \( (g, e) \) are gravity and principle direction. \( e \) is the unit vector that is a mode of surface normals distribution in an image (see details in Appendix).

In practice, we use one-hot encoding for \( \{b_i\} \), i.e., \( b_i = 1 \) if \( r_i \) is closest to \( g \), and zero otherwise.

3.3. Learning Spatial Rectifier

We learn a spatial rectifier given a set of ground truth directions \( \{(I, g, e, y)\}_{D} \) where \( D \) is the training dataset. \( y \in \mathbb{R}^{n \times H \times W} \) is the ground truth geometry \( n = 1 \) for depth and \( n = 3 \) for surface normal.

Consider two learnable functions \( f_g, f_e : \mathbb{I} \rightarrow S^2 \) that predict the gravity and principle directions from an image, respectively. These two functions constitute a spatial rectifier that can be learned by minimizing the following loss:

\[
L_{SR}(I, g, e) = \cos^{-1}(g^T f_g(I)) + \cos^{-1}(e^T f_e(I)) ,
\]

Algorithm 1: Determine reference directions

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta, {g_j}<em>{I \in D</em>{train}} )</td>
<td>( {r_i}_{i=1}^K )</td>
</tr>
<tr>
<td>( K = 1, t = \delta + \epsilon )</td>
<td>( K )</td>
</tr>
<tr>
<td>while ( t &gt; \delta ) do</td>
<td>( K \rightarrow K + 1 )</td>
</tr>
<tr>
<td>( {r_i}<em>{i=1}^K = \text{K-Medoids}({g_j}</em>{D_{train}}, K) )</td>
<td></td>
</tr>
<tr>
<td>( t = \sum_{i=1}^K \sum_{j \in C_i} |g_j - r_i|^2_2 )</td>
<td></td>
</tr>
<tr>
<td>end</td>
<td></td>
</tr>
</tbody>
</table>
The multimodal spatial rectifier warps an egocentric image by predicting the gravity $g$ and principle directions $e$, allowing learning a coherent geometry predictor $\Phi$.

We jointly learn the multimodal spatial rectifier together with the geometry predictor by minimizing the following loss:

$$L = \sum_{(I, g, e, y) \in \mathcal{D}} L_{\text{GEO}}(y, I) + \lambda L_{\text{SR}}(I, g, e).$$

(8)

The geometric loss $L_{\text{GEO}}$ measures the geometric error between the prediction and ground truth:

$$L_{\text{GEO}}(y, I) = \sum_x d(y_x, \Phi(x, I)), \quad \text{where}$$

$$d(y, \Phi) = \begin{cases} |y - \Phi| & \text{for depth} \\ \cos^{-1}(y^T \Phi) & \text{for surface normal} \end{cases}$$

where $\Phi(x, I) = h_{\Phi}^{-1} \circ \Phi(\mathcal{W}(x; R), \tilde{I})$, and $R$ can be computed by the predictions of $f_g(I)$ and $f_e(I)$.

3.4. Network Design

The multimodal spatial rectifier is a modular predictor that can combine with a geometry predictor $\Phi$ as shown in Figure 5. It is learned to predict the gravity and principle directions from an input tilted image through $f$.

$$\hat{\mathcal{W}} = \max(\hat{\mathcal{W}}, \tilde{\mathcal{W}})$$

The rectified image $\hat{\mathcal{W}}$ is the output of the network.

**Implementation Details**

Our networks take as input an RGB image of size $320 \times 240$ and output the same size surface normals or depths. We use a ResNet-18 architecture to estimate $f_g$ and $f_e$ while the geometry predictor $\Phi$ is based on the Coarse-Details [3] that provide RGBD images (depth range: 0.5~5.46m) with inertial signals (rotational velocity and linear acceleration). Eighteen participants were asked to perform diverse daily indoor activities, e.g., cleaning, sorting, cooking, eating, doing laundry, training/playing with pet, walking, shopping, vacuuming, making bed, exercising, throwing trash, watering plants, sweeping, wiping, while wearing a head-mounted camera. The camera is oriented to approximately 45° downward to ensure observing hand-object interactions. Total number of data instances is 550K images (16 hrs). Figure 6(a) illustrates the representative examples of EDINA dataset that include substantially tilted egocentric images depicting diverse activities.

The gravity direction is correlated with activities. For instance, the majority of cooking and cleaning activities are performed while facing down, whereas the shopping and interacting with others are performed while facing front as shown in Figure 6(b). Figure 6(c) illustrates the amount of data of four major indoor activities of cleaning, cooking, shopping, and organizing. Unlike existing scene datasets such as ScanNet, a large proportion of pixels of egocentric scenes belong to the foreground. Our dataset is available at https://github.com/tien-d/EgoDepthNormal.

5. Experiments

We evaluate our two main contributions: accuracy of multimodal spatial rectifier and effectiveness on multiple datasets including EDINA.

5.1. Evaluation Datasets

**HM3D** [41] To facilitate more controlled experiments, we use HM3D, a large-scale dataset containing 1,000 distinctive building-scale, real-world 3D reconstructions. The data are composed of textured 3D mesh reconstruction with high visual fidelity, which allows us to render the photo-realistic scenes from diverse viewpoints with known camera orientations. We render the RGB-D frames from each viewpoint and only retain the views that are complete (no missing surfaces or reconstruction artifacts). **ScanNet** [6] ScanNet is a large RGB-D indoor datasets with 1,500 sequences, spanning a wide variety of scenes. We use the standard dataset split used in FrameNet [23] that comprises 199,720 frames for training and 64,319 frames for validating. In addition, we utilize FrameNet’s high-quality ground-truth surface normals to augment with our EDINA for training.

**Evaluation Metrics**

We assess the accuracy of the predicted depths using multiple standard metrics, including:

(a) mean absolute relative error (Abs. Rel), (b) mean square relative error (Sqr. Rel), (c) logarithmic root mean square error (log-RMSE), (f) root mean square error (RMSE), and (g) the percentage of the estimated depths $\hat{d}$ for which $\max(\frac{d - \hat{d}}{d}, \frac{\hat{d} - d}{\hat{d}}) < \delta$, where $d$ is the ground-truth depth and
We present EDINA (Egocentric Depth on everyday INdoor Activities) dataset. (a) We show egocentric images of diverse activities with depths, surface normals, and gravity direction (black). (b) Gravity direction is highly correlated with egocentric activities. The images of cooking and cleaning activities have nearly 90° pitch angle, which is different from shopping activities. (c) EDINA includes four major indoor activities of cleaning, cooking, shopping, and home organizing. Unlike existing scene datasets such as ScanNet, a large proportion of pixels of egocentric scenes belong to the foreground.

In terms of surface normal error metrics, we also employ standard metrics originally used in [1, 13]: (a) mean absolute of the error (Mean), (b) median of absolute error (Median), (c) root mean square error (RMSE), and (d) the percentage of pixels with angular error below a threshold \( \xi \) with \( \xi = 5°, 7.5°, 11.25° \). EDINA (ours) We use EDINA dataset to train and evaluate our models on surface normal and depth estimation. With a total of 550K RGB-D images and IMU measurements, we include 500K images collected by 15 participants in the training set and use the remaining 50K images collected by the rest of the three participants as the testing set. We also follow the approach of [28] to generate ground truth surface normals from the depth images. FPHA [18] We use FPHA that is an egocentric RGB-D dataset consisting of 1,175 video sequences in several different hand-action categories for a total of 105,459 RGB-D frames and follow its official train/test split.

5.2. Baselines

We construct various baseline algorithms using the state-of-the-art scene understanding approaches. (1) PFPN: Panoptic FPN [27] is a lightweight network architecture which has been used in various high-resolution prediction tasks. We employ PFPN with the ResNet-101 [21] backbone as our baseline network architecture for both depth and surface normal estimation tasks. (2) PFPN+SR\(e_2\): we train PFPN using the spatial rec-
<table>
<thead>
<tr>
<th>Testing</th>
<th>Method</th>
<th>Abs. Rel</th>
<th>Sq. Rel</th>
<th>log-RMSE</th>
<th>RMSE</th>
<th>1.25↑</th>
<th>1.25↑</th>
<th>1.25↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDINA</td>
<td>MiDaS (MIX6)</td>
<td>0.194</td>
<td>0.079</td>
<td>0.267</td>
<td>0.247</td>
<td>68.20</td>
<td>83.96</td>
<td>93.14</td>
</tr>
<tr>
<td></td>
<td>DPT (MIX6)</td>
<td>0.195</td>
<td>0.073</td>
<td>0.256</td>
<td>0.234</td>
<td>66.95</td>
<td>86.07</td>
<td>94.39</td>
</tr>
<tr>
<td></td>
<td>PFPN (ScanNet)</td>
<td>0.536</td>
<td>0.292</td>
<td>0.450</td>
<td>0.410</td>
<td>28.50</td>
<td>63.31</td>
<td>84.60</td>
</tr>
<tr>
<td></td>
<td>PFPN (EDINA)</td>
<td>0.173</td>
<td>0.052</td>
<td>0.210</td>
<td>0.181</td>
<td>78.81</td>
<td>92.97</td>
<td>97.06</td>
</tr>
<tr>
<td></td>
<td>PFPN</td>
<td>0.161</td>
<td>0.044</td>
<td>0.197</td>
<td>0.168</td>
<td>81.03</td>
<td>94.16</td>
<td>97.68</td>
</tr>
<tr>
<td></td>
<td>PFPN+MSR (Ours)</td>
<td>0.145</td>
<td>0.041</td>
<td>0.182</td>
<td>0.155</td>
<td>84.06</td>
<td>94.54</td>
<td>97.87</td>
</tr>
<tr>
<td></td>
<td>MiDaS+SR (Ours)</td>
<td>0.041</td>
<td>0.182</td>
<td>0.155</td>
<td>0.123</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DPT+SR (Ours)</td>
<td>0.041</td>
<td>0.182</td>
<td>0.155</td>
<td>0.123</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PFPN+SR (Ours)</td>
<td>0.039</td>
<td>0.174</td>
<td>0.149</td>
<td>0.119</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. We compare the performance of depth prediction of our method (MSR) with baselines on EDINA and FPHA testing data. The † indicates methods that predict scale-ambiguous depth and thus require a scale correction step. The numbers in the parenthesis show the percentage of the reduction in error metrics of PFPN+MSR (Ours) with respect to the baseline PFPN, where the green highlight denote this improvement in percentage.

5.3. Performance Benchmark

Depth Prediction We first show the effectiveness of our MSR through a controlled experiment using the HM3D dataset. Specifically, we render from HM3D a training set containing 82,941 RGB-D frames respectively at upright (tilt 0°) and tilt 40° orientation and a testing set containing 3,944 RGB-D frames respectively at upright and tilt angles at 10°, 20°, 30°, and 40°. The tilted images are rendered with the rotation around e_z axis with respect to the upright orientation (roll). Figure 7 illustrates the performance between PFPN, PFPN+SR(e_2), and PFPN+MSR (at two distribution modes 0° and 40°) in 2 cases: (i) in-distribution: 0° and 40°, and (ii) out-of-distribution: 10°, 20°, 30°, and 40°. We can observe that for the in-distribution case, the baseline and MSR performs similarly while the PFPN+SR(e_2) slightly underperforms the former ones due to its excessive warping. On the other hand, for the out-of-distribution case, while the baseline method degenerates at 10°, 20°, and 30°, both SR and MSR generalize reasonably well with the SR slightly degenerates when the tilt angle is further from its central mode (upright).

Table 1 demonstrates the performance of our multimodal spatial rectifier and the effectiveness of our EDINA dataset. A baseline network equipped with our spatial rectifier (PFPN+MSR) outperforms other baselines on all evaluation metrics, not only on our EDINA dataset but also on FPHA dataset. While the performance margin for the network equipped with and without MSR is narrow on EDINA, it is significant when generalizing to FPHA. We conjecture that EDINA dataset that comprises a large variation in pitch angles can be overfitted by a large capacity network such as PFPN. In contrast, FPHA dataset is taken from a shoulder-mounted camera, imposing more roll motion on the image, thus it causes a strong degradation for PFPN trained.

Figure 7. Performance of PFPN, PFPN+SR(e_2), and PFPN+MSR on HM3D test set. The dark and light color indicates the in- (at 0° and 40°) and out-of-distribution (at 10°, 20°, 30°), respectively.
on ScanNet+EDINA datasets. We conclude that our MSR module is highly beneficial for learning egocentric scene geometry. Figure 8 illustrates the qualitative results of our method on EPIC-KITCHENS and FPHA. More qualitative results can be found in Supplementary Materials.

In addition, baselines that do not employ egocentric data, i.e., MiDaS (MIX6), DPT (MIX6), PFPN (ScanNet), perform poorly on both EDINA and FPHA. On the other hand, the network trained only on EDINA performs strongly on its own test set while lacking generalizability towards to other dataset such as FPHA. This indicates that learning can greatly benefit from a large amount of high quality ground truth geometry from ScanNet, together with our EDINA.

Surface Normal Prediction In Table 2, we compare our method with the baselines on EDINA dataset and demonstrate the effectiveness of our proposed multimodal spatial rectifier on surface normal prediction. On median and tight thresholds ($\xi = 5^\circ, 7.5^\circ$), the unimodal spatial rectifier with $e_2$ as the reference direction (PFPN+SR ($e_2$)) shows notable improvements compared to the baseline PFPN while inferior in terms of RMSE and mean. Moreover, this issue further escalates when $e_3$ is used as the only reference direction (PFPN+SR ($e_3$)). This is mainly caused by the excessive warping that is very common on egocentric data. In contrast, by predicting the multiple reference directions, our PFPN+MSR can generalize to diverse viewpoints, thus outperforms other baselines on all metrics. Note that this also applies for DORN+MSR, suggesting that it is highly flexible and can be easily integrated into other networks. See Figure 8 for qualitative results.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Median</th>
<th>RMSE</th>
<th>$5^\circ$</th>
<th>$7.5^\circ$</th>
<th>$11.25^\circ$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFPN</td>
<td>20.24</td>
<td>13.61</td>
<td>27.51</td>
<td>15.46</td>
<td>26.93</td>
<td>42.63</td>
</tr>
<tr>
<td>PFPN+SR ($e_2$)</td>
<td>20.27</td>
<td>13.41</td>
<td>28.47</td>
<td>25.10</td>
<td>34.00</td>
<td>44.81</td>
</tr>
<tr>
<td>PFPN+SR ($e_3$)</td>
<td>39.20</td>
<td>31.19</td>
<td>50.63</td>
<td>16.29</td>
<td>23.47</td>
<td>30.06</td>
</tr>
<tr>
<td>PFPN+MSR</td>
<td>19.30</td>
<td>12.54</td>
<td>27.37</td>
<td>26.00</td>
<td>35.49</td>
<td>46.74</td>
</tr>
<tr>
<td>DORN</td>
<td>19.57</td>
<td>12.92</td>
<td>27.07</td>
<td>17.42</td>
<td>29.01</td>
<td>44.66</td>
</tr>
<tr>
<td>DORN+SR ($e_2$)</td>
<td>19.96</td>
<td>12.68</td>
<td>28.46</td>
<td>25.53</td>
<td>35.00</td>
<td>46.35</td>
</tr>
<tr>
<td>DORN+SR ($e_3$)</td>
<td>21.99</td>
<td>14.83</td>
<td>30.46</td>
<td>21.33</td>
<td>29.83</td>
<td>40.87</td>
</tr>
<tr>
<td>DORN+MSR</td>
<td>18.56</td>
<td>11.55</td>
<td>26.83</td>
<td>26.58</td>
<td>37.04</td>
<td>49.18</td>
</tr>
</tbody>
</table>

Table 2. We compare the performance of surface normal prediction of our method (MSR) with baselines including the unimodal spatial rectifier (SR) on EDINA testing data.

6. Summary

In this paper, we present a new multimodal spatial rectifier for egocentric scene understanding, i.e., predicting depths and surface normals from a single view egocentric image. The multimodal spatial rectifier identifies multiple reference directions to learn a geometrically coherent representation from tilted egocentric images. This rectifier enables warping the image to the closest mode such that the geometry predictor in this mode can accurately estimate the geometry of the rectified scene. To facilitate the learning of our multimodal spatial rectifier, we introduce a new dataset called EDINA that comprises 550K synchronized RGBD and gravity data of diverse indoor activities. We show that EDINA is complementary to ScanNet, allowing us to learn a strong multimodal spatial rectifier. We evaluate our method on egocentric datasets including our EDINA, FPHA and EPIC-KITCHENS, which outperforms the baselines.

Acknowledgements This work is partially supported by NSF CAREER IIS-1846031.
References


[31] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In CVPR, 2015. 2


[58] Jin Zeng, Yanfeng Tong, Yunmu Huang, Qiong Yan, Wenxiu Sun, Jing Chen, and Yongtian Wang. Deep surface normal estimation with hierarchical rgb-d fusion. In CVPR, 2019. 2
