Egocentric Scene Understanding via Multimodal Spatial Rectifier

Tien Do[1]　Khiem Vuong[2]　Hyun Soo Park[1]

[1] University of Minnesota　[2] Carnegie Mellon University

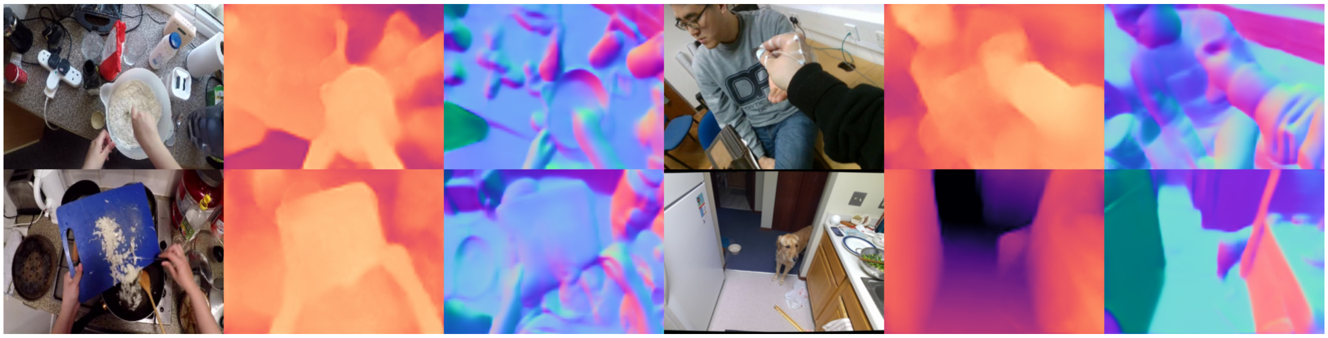


Figure 1. We study the problem of predicting geometry (depths and surface normals) from a single view egocentric image that includes dynamic objects (e.g., hand and people). We design a multimodal spatial rectifier that can effectively handle the excessively tilted images caused by head movement (e.g., nearly 90 degree pitch angle when engaging eye-hand coordination). Our method shows strong performance on unseen images from EPIC-KITCHENS [7] (left), FPHA [18] (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 [6] and NYUv2 [36], 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) [18] and EPIC-KITCHENS [7].

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 [6, 19, 36]. These trained models [14, 23] 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 [8]. 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 titled 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 through mental stabilization of visual stimuli [54]. However, the spatial rectifier shows inferior performance on predicting 3D geometry of egocentric images that involve substantial head movement (e.g., nearly 90 degree pitch), leading to excessive perspective warps. We present a *multimodal 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 supervises to learn the mapping 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

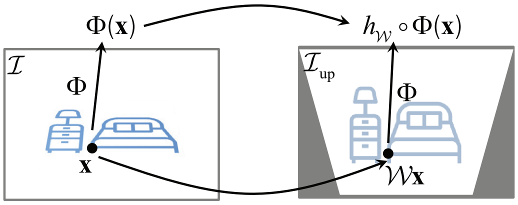


Figure 2. A spatial rectifier enforces equivariance property to learn a geometrically coherent representation. When a point is transformed by $\mathcal{W}$, its feature is expected to transformed accordingly, i.e., $\Phi(\mathcal{W}\mathbf{x}) = h_{\mathcal{W}} \circ \Phi(\mathbf{x})$.

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 provides synchronized RGBD and gravity directions captured from an egocentric viewpoint with diverse daily activities.

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 $\Phi : \mathbb{R}^2 \times \mathbb{I} \rightarrow \mathbb{R}^n$ that predicts the geometry of a pixel $\mathbf{x} \in \mathbb{R}^2$ in an image $\mathcal{I} \in \mathbb{I}$, where $\mathbb{I} = [0,1]^{3\times H\times W}$ is the image range (H and W are its height and width, respectively). We denote the prediction:

$$y = \Phi(\mathbf{x}, \mathcal{I}), \tag{1}$$

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 $\mathcal{I}$ with the gravity direction $\mathbf{g} \in \mathbb{S}^2$ in the camera coordinate system to the *upright* image $\mathcal{I}_{\text{up}}$ with the upright gravity direction $\mathbf{g}_{\text{up}}$ by explicitly enforcing an equivariant property through 3D rotation (Figure 2):

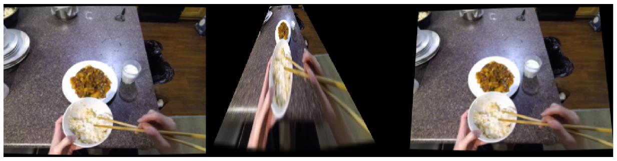


Figure 3. A unimodal spatial rectifier produces an excessive perspective warp (middle) to align the image to the gravity direction, which significantly degrade the performance of geometry prediction. We use a multimodal spatial rectifier that warps to multiple reference directions that minimizes the impact of the perspective warping (right).

$$h_{\mathcal{W}} \circ \Phi(\mathbf{x}, \mathcal{I}) = \Phi(\mathcal{W}(\mathbf{x}; \mathbf{R}_{\text{up}}), \mathcal{I}_{\text{up}}), \tag{2}$$

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 $\mathbf{R}_{\text{up}} \in SO(3)$ such that $\mathbf{g}_{\text{up}} = \mathbf{R}_{\text{up}}\mathbf{g}$. $\mathcal{I}_{\text{up}}$ is warped from the tilted image by $\mathcal{W}$, i.e., $\mathcal{I}_{\text{up}} = \mathcal{I}(\mathcal{W}(\mathbf{x}; \mathbf{R}_{\text{up}}))$. $h_{\mathcal{W}}$ 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 ($\mathbb{S}^2$), i.e., $h_{\mathcal{W}} \circ \Phi = \mathbf{R}_{\text{up}}\Phi$; (2) for the depth prediction, $h_{\mathcal{W}}$ is defined as:

$$h_{\mathcal{W}} \circ \Phi = \left(\mathbf{R}_{\text{up}}\mathbf{K}^{-1}\widetilde{\mathbf{x}}\right)_z \Phi \tag{3}$$

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

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

$$\Phi(\mathbf{x}, \mathcal{I}) = h_{\mathcal{W}}^{-1} \circ \Phi_{\text{up}}(\mathcal{W}(\mathbf{x}; \mathbf{R}_{\text{up}}), \mathcal{I}_{\text{up}}), \tag{4}$$

where $h_{\mathcal{W}}^{-1}$ is the spatial rectifier, and Φ_{up} is the geometry predictor learned from upright images. A key benefit of this function composition is that Φ_{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 $\mathcal{I}_{\text{up}}$ (e.g., 90° pitch tilted image), which in turns, significantly degrades the performance of the geometry predictor as shown in Figure 3 (middle).

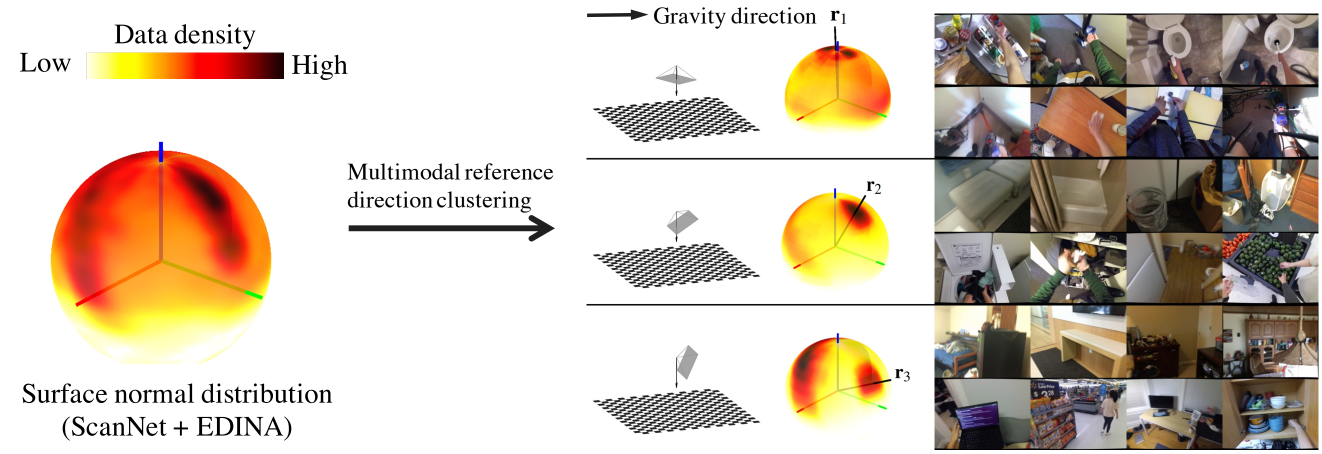


Figure 4. Unlike the spatial rectifier [8] that relies on the unimodal surface normal distribution with respect to the gravity direction (left), we present a multimodal spatial rectifier that generalizes the spatial rectifier by learning multiple reference directions (right). As a result, the surface 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:

$$\Phi(\mathbf{x}, \mathcal{I}) = \frac{1}{\sum_i b_i} \sum_i b_i \left(h^{-1}_{\mathcal{W}_i} \circ \Phi_i(\mathcal{W}(\mathbf{x}; \mathbf{R}_i), \mathcal{I}_i) \right), \quad (5)$$

where $b_i \in \mathbb{R}_+$ is a non-negative weight to mix transformations, and $\mathbf{R}_i$ is the rotation that transforms the gravity of the tilted image to the i^{th} reference direction, i.e., $\mathbf{r}_i = \mathbf{R}_i\mathbf{g}$. $\mathcal{I}_i$ is warped from the tilted image by $\mathcal{W}_i$, i.e., $\mathcal{I}_i = \mathcal{I}(\mathcal{W}(\mathbf{x}; \mathbf{R}_i))$. The reference direction $\mathbf{r} \in \mathbb{S}^2$ is a generalization of the upright gravity $\mathbf{g}_{\text{up}}$, which specifies the egocentric tilted images to be warped. Φ_i is the geometry predictor designed for the i^{th} reference direction. We denote $\mathcal{W}(\mathbf{x}; \mathbf{R}_i)$ by $\mathcal{W}_i$ 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 $\{\mathbf{r}_i\}_{i=1}^K$ along the pitch angles by clustering the gravity of egocentric images with K is the predefined number of the reference directions:

$$\underset{\{\mathbf{r}_i\}_{i=1}^K}{\text{minimize}} \sum_{i=1}^{K} \sum_{j \in \mathcal{C}_i} \|\mathbf{g}_j - \mathbf{r}_i\|_2^2, \quad (6)$$

where $\mathcal{C}_i$ is the set of the indices of training instances of which gravity directions closest to the i^{th} reference direction $\mathbf{r}_i$. 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 δ 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: $(\mathbf{g}, \mathbf{e})$ are gravity and principle direction. $\mathbf{e}$ 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 $\mathbf{r}_i$ is closest to $\mathbf{g}$, and zero otherwise.

3.3. Learning Spatial Rectifier

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

Consider two learnable functions $f_{\mathbf{g}}, f_{\mathbf{e}} : \mathbb{I} \rightarrow \mathbb{S}^2$ 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:

$$\mathcal{L}_{\text{SR}}(\mathcal{I}, \mathbf{g}, \mathbf{e}) = \cos^{-1}(\mathbf{g}^{\mathsf{T}} f_{\mathbf{g}}(\mathcal{I})) + \cos^{-1}(\mathbf{e}^{\mathsf{T}} f_{\mathbf{e}}(\mathcal{I})), \quad (7)$$

Algorithm 1: Determine reference directions

```
Input  : $\delta, \{\mathbf{g}_j\}_{\mathcal{I}_j \in \mathcal{D}_{\text{train}}}$
Output: $\{\mathbf{r}_i\}_{i=1}^K$
$K = 1, t = \delta + \epsilon$;
while $t > \delta$ do
    $\{\mathbf{r}_i\}_{i=1}^K$ = K-Medoids$(\{\mathbf{g}_j\}_{\mathcal{D}_{\text{train}}}, K)$;
    $t = \sum_{i=1}^K \sum_{j \in \mathcal{C}_i} \|\mathbf{g}_j - \mathbf{r}_i\|_2^2$;
    $K \leftarrow K + 1$;
end
```

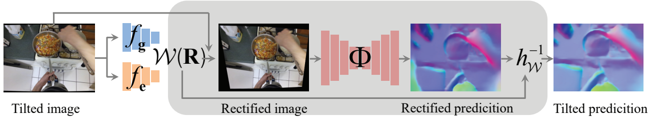


Figure 5. The multimodal spatial rectifier warps an egocentric image by predicting the gravity $\mathbf{g}$ and principle directions $\mathbf{e}$s, allowing learning a coherent geometry predictor Φ.

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

$$\mathcal{L} = \sum_{\{\mathcal{I},\mathbf{g},\mathbf{e},\mathbf{y}\}\in\mathcal{D}} \mathcal{L}_{\rm GEO}(\mathbf{y},\mathcal{I}) + \lambda\mathcal{L}_{\rm SR}(\mathcal{I},\mathbf{g},\mathbf{e}). \quad (8)$$

The geometric loss $\mathcal{L}_{\rm GEO}$ measures the geometric error between the prediction and ground truth:

$$\mathcal{L}_{\rm GEO}(\mathbf{y},\mathcal{I}) = \sum_{\mathbf{x}} d(\mathbf{y}_{\mathbf{x}}, \mathbf{\Phi}(\mathbf{x},\mathcal{I})), \text{ where}$$

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

where $\mathbf{\Phi}(\mathbf{x},\mathcal{I}) = h_{\mathcal{W}}^{-1} \circ \mathbf{\Phi}(\mathcal{W}(\mathbf{x};\mathbf{R}),\overline{\mathcal{I}})$, and $\mathbf{R}$ can be computed by the predictions of $f_{\mathbf{g}}(\mathcal{I})$ and $f_{\mathbf{e}}(\mathcal{I})$.

3.4. Network Design

The multimodal spatial rectifier is a modular predictor that can combine with a geometry predictor Φ as shown in Figure 5. It is learned to predict the gravity and principle directions from an input tilted image through $f_{\mathbf{g}}$ and $f_{\mathbf{e}}$, respectively. With the predicted direction, it computes the rotation $\mathbf{R}$ that can be used to warp the image to the reference direction $\mathcal{W}$. The geometry predictor takes as input an image and predict depths and surface normals. These predictions are unwarped by $h_{\mathcal{W}}^{-1}$.
Implementation Details Our networks take as input an RGB image of size 320 × 240 and output the same size surface normals or depths. We use a ResNet-18 architecture to estimate $f_{\mathbf{g}}$ and $f_{\mathbf{e}}$ while the geometry predictor Φ is specified in 5.2. The proposed models are implemented in PyTorch [38], trained with a batch size of 32 on a single NVIDIA Tesla V100 GPU, and optimized by Adam [26] optimizer with a learning rate of 10^{-4}. We train our models for 20 epochs.

4. EDINA Dataset

We present a new RGBD dataset called *EDINA* (Egocentric Depth on everyday INdoor Activities) that facilitates learning 3D geometry from egocentric images. Each instance in the dataset is a triplet: RGB image (1920×1080), depths and surface normals (960×540), and 3D gravity direction. The data were collected using Azure Kinect cameras [34] 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 home 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.

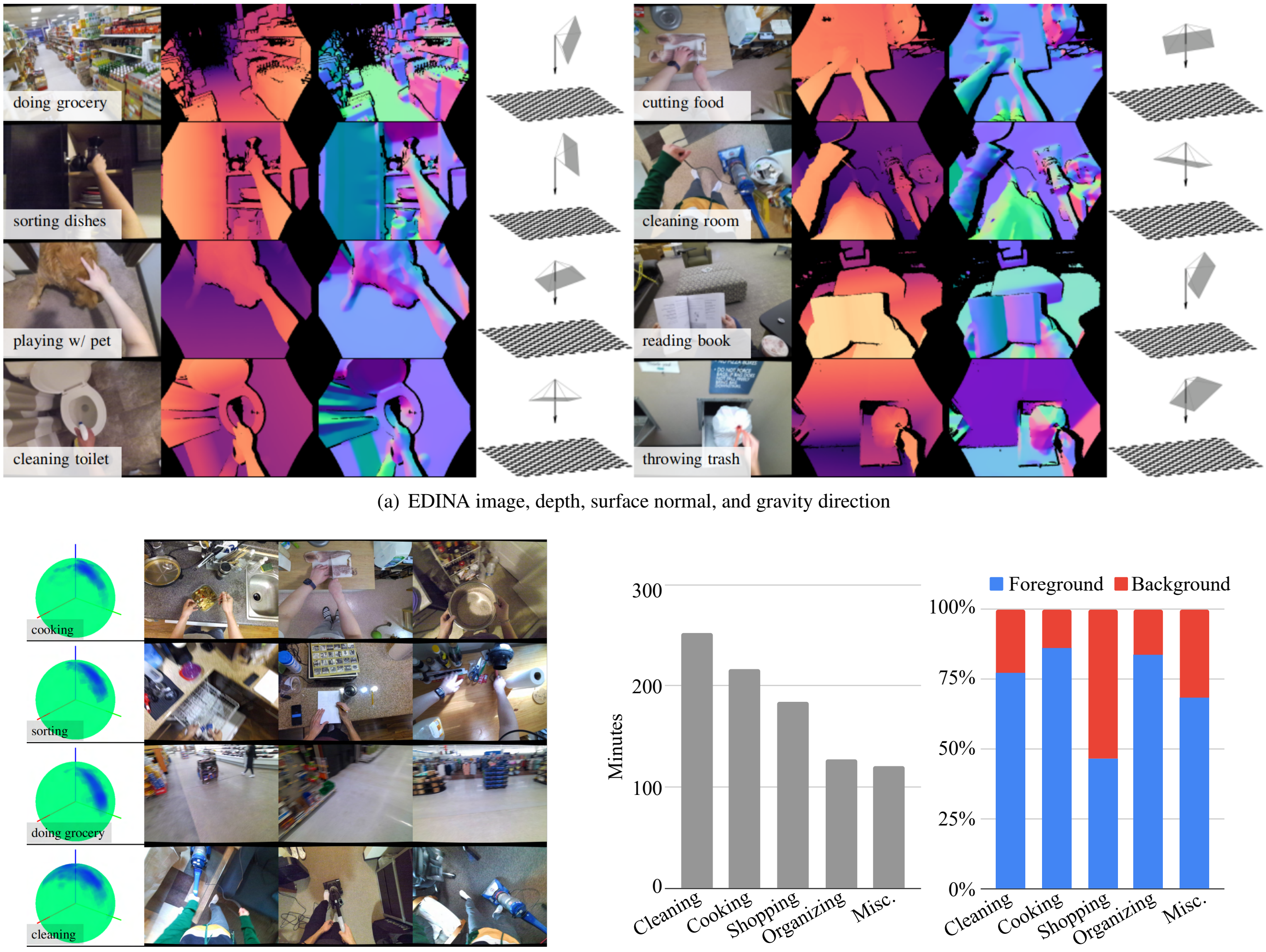
5. Experiments

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 1500 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 (Sq. 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 $\max(\frac{\hat{d}}{d*}, \frac{d*}{\hat{d}}) < \delta$, where $d*$ is the ground-truth depth and



(a) EDINA image, depth, surface normal, and gravity direction

(b) Gravity distribution

(c) Activities

Figure 6. 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.

$\delta = 1.25, 1.25^2, 1.25^3$. 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 ξ 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-

Testing	Method	Abs. Rel↓	Sq. Rel↓	log-RMSE↓	RMSE ↓	1.25↑	1.25^2 ↑	1.25^3 ↑
EDINA	MiDaS (`MIX6`)†	0.194	0.079	0.267	0.247	68.20	83.96	93.14
	DPT (`MIX6`)†	0.195	0.073	0.256	0.234	66.95	86.07	94.39
	PFPN (`ScanNet`)	0.536	0.292	0.450	0.410	28.50	63.31	84.60
	PFPN (`EDINA`)	0.173	0.052	0.210	0.181	78.81	92.97	97.06
	PFPN	0.161	0.044	0.197	0.168	81.03	94.16	97.68
	PFPN+MSR (Ours)	**0.145** (-9.7%)	**0.041** (-8.5%)	**0.182** (-7.7%)	**0.155** (-7.9%)	**84.06**	**94.54**	**97.87**
FPHA	PFPN (`ScanNet`)	1.252	0.893	0.788	0.580	10.36	28.07	48.87
	PFPN (`EDINA`)	1.229	4.114	0.802	1.483	25.98	46.38	62.70
	PFPN	0.737	0.457	0.549	0.397	32.60	57.61	75.14
	PFPN+MSR (Ours)	**0.657** (-10.8%)	**0.369** (-19.2%)	**0.508** (-7.3%)	**0.337** (-15.2%)	**37.70**	**62.50**	**78.30**

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.

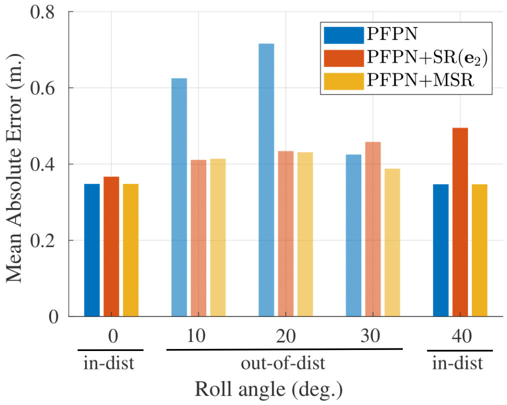


Figure 7. Performance of PFPN, PFPN+SR($\mathbf{e}_2$), and PFPN+MSR on HM3D test set. The dark and light color indicates the in- (at 0° and 40°) and out-of-distribution (at 10°, 20°, 30°), respectively.

tifier [8] (SR) with a unimodal reference direction $\mathbf{e}_2 = \begin{bmatrix}0 & 1 & 0\end{bmatrix}^\mathsf{T}$. (3) PFPN+SR($\mathbf{e}_3$): we train PFPN using the SR with a unimodal reference direction $\mathbf{e}_3 = \begin{bmatrix}0 & 0 & 1\end{bmatrix}^\mathsf{T}$. (4) PFPN+MSR: we train PFPN with our multimodal spatial rectifer (MSR) described in Section 3.2. (5-8) DORN: DORN [14] is a high-capacity network architecture that is recently utilized in state-of-the-art surface normal estimation methods [8, 23]. Similar to PFPN, we also train DORN with the unimodal spatial rectifier on two reference directions $\mathbf{e}_2$, $\mathbf{e}_3$ and with our multimodal spatial rectifier, denoted by DORN+SR($\mathbf{e}_2$), DORN+SR($\mathbf{e}_3$), and DORN+MSR, respectively. (9) MiDaS [44], (10) DPT [43]: state-of-the-art depth prediction model trained on a large scale dataset MIX6 [43]. Since the depth prediction from MiDaS and DPT is ambiguous up to a scale factor, we scale the predicted depth maps with a common factor computed from the least-squares method [3,55] using the ground truth depth on the train set. We denote a network trained on a dataset by METHOD (`DATASET`), e.g., PFPN (`EDINA`) is the PFPN network that is trained on EDINA dataset. By default, all networks are trained on `ScanNet+EDINA`.

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 $\mathbf{e}_3$ axis with respect to the upright orientation (roll). Figure 7 illustrates the performance between PFPN, PFPN+SR($\mathbf{e}_2$), 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°, and 30°. We can observe that for the in-distribution case, the baseline and MSR performs similarly while the PFPN+SR($\mathbf{e}_2$) 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

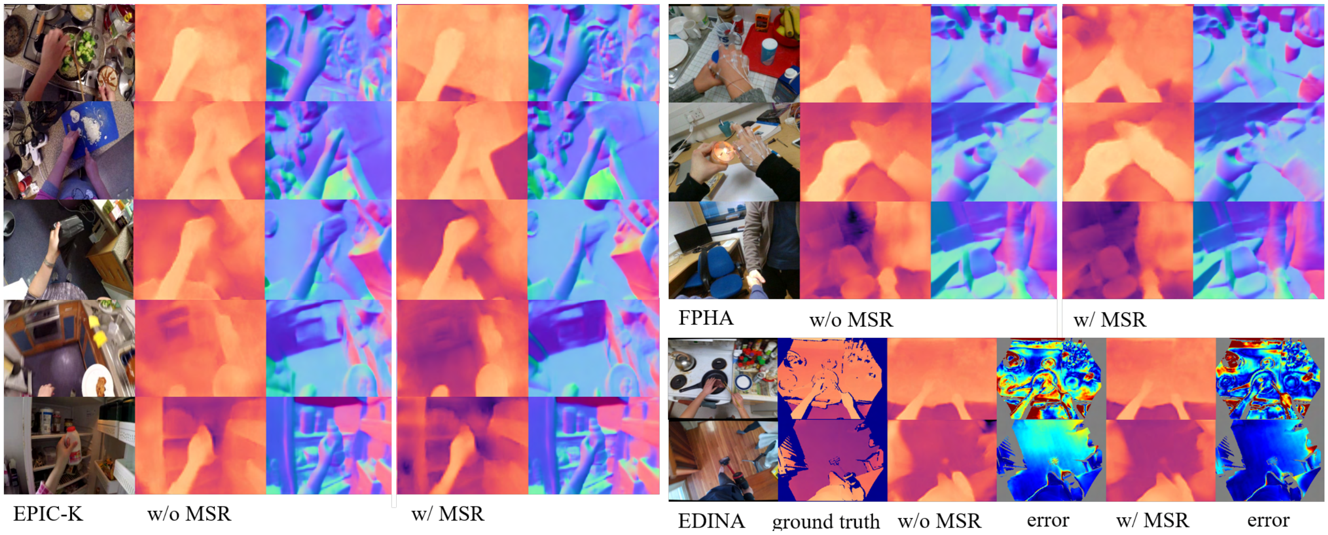


Figure 8. Qualitative results for EPIC-KITCHENS (Left), FPHA (top right), and EDINA (bottom right). For EPIC-KITCHENS and FPHA, from left to right: (1) RGB image, (2) depths and surface normals using PFPN trained on ScanNet and Edina, and (3) depths and surface normals using PFPN+MSR trained on ScanNet and Edina. For EDINA, from left to right: (1) RGB image, (2) ground truth depths, (3) estimated depths (w/o and w/ MSR), (4) the corresponding depth error (w/o and w/ MSR).

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`), performs 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 ($\xi = 5°, 7.5°$), the unimodal spatial rectifier with $\mathbf{e}_2$ as the reference direction (PFPN+SR ($\mathbf{e}_2$)) shows notable improvements compared to the baseline PFPN while inferior in terms of RMSE and mean. Moreover, this issue further escalates when $\mathbf{e}_3$ is used as the only reference direction (PFPN+SR ($\mathbf{e}_3$)). 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.

Method	Mean↓	Median↓	RMSE↓	5°↑	7.5°↑	11.25°↑
PFPN	20.24	13.61	27.51	15.46	26.93	42.63
PFPN+SR ($\mathbf{e}_2$)	20.27	13.41	28.47	25.10	34.00	44.81
PFPN+SR ($\mathbf{e}_3$)	39.20	31.19	50.63	16.29	23.47	30.06
PFPN+MSR	**19.30**	**12.54**	**27.37**	**26.00**	**35.49**	**46.74**
DORN	19.57	12.92	27.07	17.42	29.01	44.66
DORN+SR ($\mathbf{e}_2$)	19.96	12.68	28.46	25.53	35.00	46.35
DORN+SR ($\mathbf{e}_3$)	21.99	14.83	30.46	21.33	29.83	40.87
DORN+MSR	**18.56**	**11.55**	**26.83**	**26.58**	**37.04**	**49.18**

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

[1] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2d-3d alignment via surface normal prediction. In *CVPR*, 2016. 2, 6

[2] Gedas Bertasius, Aaron Chan, and Jianbo Shi. Egocentric basketball motion planning from a single first-person image. In *CVPR*, 2018. 3

[3] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. 7

[4] Weifeng Chen, Donglai Xiang, and Jia Deng. Surface normals in the wild. In *CVPR*, 2017. 2

[5] Xiaotian Chen, Xuejin Chen, and Zheng-Jun Zha. Structure-aware residual pyramid network for monocular depth estimation. *IJCAI*, 2019. 2

[6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1, 2, 3, 5

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 1, 2, 3

[8] Tien Do, Khiem Vuong, Stergios I. Roumeliotis, and Hyun Soo Park. Surface normal estimation of tilted images via spatial rectifier. In *ECCV*, 2020. 1, 2, 3, 4, 7

[9] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *CVPR*, 2015. 2

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 2

[11] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding egocentric activities. In *ICCV*, 2011. 3

[12] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012. 3

[13] David F. Fouhey, Abhinav Gupta, and Martial Hebert. Data-driven 3d primitives for single image understanding. In *ICCV*, 2013. 6

[14] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 1, 2, 7

[15] Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. Personal-location-based temporal segmentation of egocentric videos for lifelogging applications. *JVCIR*, 2018. 3

[16] Antonino Furnari, Giovanni Maria Farinella, and Sebastiano Battiato. Recognizing personal locations from egocentric videos. *THMS*, 2016. 3

[17] Antonino Furnari, Giovanni Maria Farinella, and Sebastiano Battiato. Temporal segmentation of egocentric videos to highlight personal locations of interest. In *ECCV*, 2016. 3

[18] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018. 1, 2, 3, 6

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. “Are we ready for autonomous driving? the kitti vision benchmark suite”. In *CVPR*, 2012. 1, 2

[20] Zhixiang Hao, Yu Li, Shaodi You, and Feng Lu. Detail preserving depth estimation from a single image using attention guided networks. In *3DV*, 2018. 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[22] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *WACV*, 2019. 2

[23] Jingwei Huang, Yichao Zhou, Thomas Funkhouser, and Leonidas J. Guibas. Framenet: Learning local canonical frames of 3d surfaces from a single rgb image. In *ICCV*, 2019. 1, 2, 5, 7

[24] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä. Guiding monocular depth estimation using depth-attention volume. In *ECCV*, 2020. 2

[25] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *ECCV*, 2018. 2

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 5

[27] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019. 6

[28] L Ladickỳ, Bernhard Zeisl, and Marc Pollefeys. Discriminatively trained dense surface normal estimation. In *ECCV*, 2014. 6

[29] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016. 2

[30] Jae-Han Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *CVPR*, 2019. 2

[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

[32] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *CVPR*, 2019. 2

[33] Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293, 2014. 4

[34] Microsoft. Azure Kinect. https://azure.microsoft.com/en-us/services/kinect-dk/. 5

[35] Mohammad Moghimi, Pablo Azagra, Luis Montesano, Ana C Murillo, and Serge Belongie. Experiments on an rgb-d wearable vision system for egocentric activity recognition. In *CVPR Workshops*, 2014. 3

[36] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 2, 3

[37] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 2009. 4

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*. 2019. 5

[39] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 3

[40] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *CVPR*, 2018. 2

[41] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. In *arXiv*, 2021. 5

[42] Michael Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In *ICCV Workshops*, 2019. 2

[43] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *arXiv*, 2021. 2, 7

[44] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 2, 7

[45] Haoyu Ren, Mostafa El-Khamy, and Jungwon Lee. Deep robust single image depth estimation neural network using scene understanding. In *CVPR Workshops*, 2019. 2

[46] Nicholas Rhinehart and Kris M Kitani. First-person activity forecasting with online inverse reinforcement learning. In *ICCV*, 2017. 3

[47] Grégory Rogez, James S Supancic, and Deva Ramanan. Understanding everyday hands in action from rgb-d images. In *ICCV*, 2015. 3

[48] Yuki Saito, Ryo Hachiuma, Masahiro Yamaguchi, and Hideo Saito. In-plane rotation-aware monocular depth estimation using slam. In *IW-FCV*, 2020. 2

[49] Kourosh Sartipi, Tien Do, Tong Ke, Khiem Vuong, and Stergios I Roumeliotis. Deep depth estimation from visual-inertial slam. In *IROS*, 2020. 2

[50] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 2

[51] Yansong Tang, Zian Wang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Multi-stream deep neural networks for rgb-d egocentric action recognition. *TCSVT*, 2018. 3

[52] Peng Wang, Xiaohui Shen, Bryan Russell, Scott Cohen, Brian Price, and Alan L Yuille. Surge: Surface regularized geometry estimation from a single image. In *NeurIPS*, 2016. 2

[53] Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, and Jan-Michael Frahm. Vplnet: Deep single view normal estimation with vanishing points and lines. In *CVPR*, 2020. 2

[54] Robert H. Wurtz, Wilsaan M. Joiner, and Rebecca A. Berman. Neuronal mechanisms for visual stability: progress and problems. *Philosophical Transactions of The Royal Society B*, 2011. 2

[55] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 7

[56] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, 2013. 2

[57] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *ICCV*, 2019. 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

[59] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, 2019. 2

[60] Yunhan Zhao, Shu Kong, and Charless Fowlkes. Camera pose matters: Improving depth prediction by mitigating pose distribution bias. In *CVPR*, 2021. 2