Appendix
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

1. Computing Principle Direction $\mathbf{e}$

We describe the procedure to compute the principle direction $\mathbf{e}$ that are used in Equation (7) in the main manuscript. For the j^{th} image, we find the weight b_{ji} using the hard assignment by restricting $b_{ji} = \{0, 1\}$ where b_{ji} is one if $\mathbf{r}_i$ is closest to $\mathbf{g}$, and zero otherwise. After specifying a cluster of egocentric images based on the reference direction, we find the surface normal distribution per cluster:

$$Q_i = \frac{1}{|\mathcal{C}_i|} \sum_{j \in \mathcal{C}_i} \texttt{hist}(\mathbf{n}_j), \tag{1}$$

where $\texttt{hist}(\mathbf{n}_j)$ is the angular histogram of the surface normals of the j^{th} training image. $\mathbf{n}_j$ is the $3 \times n$ matrix that each column presents a pixel's surface normal direction and n is the total pixel in image $\mathcal{I}_j$. The optimal rotation for the j^{th} image towards the i^{th} reference direction, $\mathbf{R}^*_{ji}$, is the one that maximizes the similarity in the surface normal distributions:

$$\mathbf{R}^*_{ji} = \underset{\mathbf{R}_{ji}}{\text{argmin}}\, D_{\text{KL}}\left(\texttt{hist}(\mathbf{R}_{ji}\mathbf{n}_j) || Q_i\right), \tag{2}$$

where $\mathbf{R}_{ji}\mathbf{n}_j$ is the rotated surface normals, and D_{KL} is KL divergence [2]. We optimize Equation (2) with an initial guess of $\mathbf{R}_{ji}$ computed by the gravity and reference directions:

$$\mathbf{R}_{ji} = \mathbf{I}_3 + 2\mathbf{r}_i\mathbf{g}_j^{\mathsf{T}} - \frac{(\mathbf{r}_i + \mathbf{g}_j)(\mathbf{r}_i + \mathbf{g}_j)^{\mathsf{T}}}{1 + \mathbf{r}_i^{\mathsf{T}}\mathbf{g}_j}, \tag{3}$$

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

This optimal rotation can be parametrized by the principle direction $\mathbf{e}_j$ [1], where $\mathbf{e}$ can be computed by:

$$\mathbf{e}_j = \mathbf{R}^*_{ji}\mathbf{g}_j. \tag{4}$$

We use the optimal principle direction as a ground truth to learn the multimodal spatial rectifier.

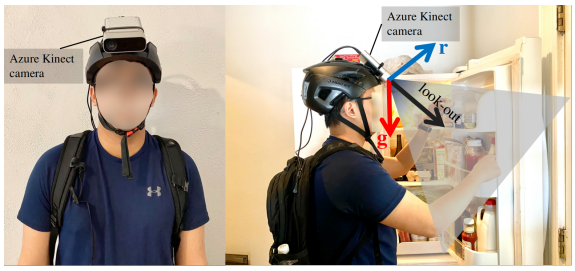

Figure 1. EDINA dataset recording setup.

2. Hardware setup

The participants were asked to wear an Azure-Kinect-mounted helmet while performing diverse daily indoor activities. The sensor was also connected to a laptop which reads and stores the raw data from the Azure Kinect device using the provided SDK. Figure 1 demonstrates the mounting configuration in which the camera is oriented to approximately 45° downward so that the captured interactions are within the field-of-view of the camera.

3. More Results

Baselines In addition to the datasets mentioned in the main manuscript, we also perform experiments on THU-READ [3]. THU-READ is an egocentric RGB-D dataset consisting of 1,920 video sequences in several different hand-action categories for a total of 171,474 RGB-D frames. We follow THU-READ's official 3/1 split for training/testing.

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 $\delta = 1.25, 1.25^2, 1.25^3$.

Depth Prediction Table 1 summarizes the performance of

Testing	Method	Abs. Rel↓	Sq. Rel↓	log-RMSE↓	RMSE ↓	1.25↑	1.25^2 ↑	1.25^3 ↑
EDINA	PFPN (THU-READ)	0.405	0.210	1.044	0.431	28.71	45.72	61.97
	PFPN (FPHA)	0.314	0.167	0.500	0.378	42.82	66.48	78.98
	PFPN (ScanNet)	0.536	0.292	0.450	0.410	28.50	63.31	84.60
	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 (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+SR($\mathbf{e_2}$)	1.573	3.145	0.938	1.155	5.58	19.75	42.32
	PFPN+SR($\mathbf{e_3}$)	0.381	0.333	0.475	0.416	51.75	73.37	84.62
	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 (THU-READ)	0.439	0.150	4.629	**0.279**	32.03	54.92	70.76
	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**
	PFPN (FPHA)	0.119	0.023	0.139	0.075	91.29	97.31	98.75

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.

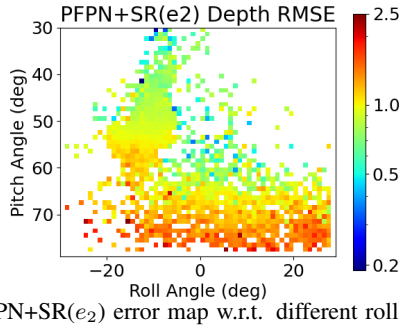

Figure 2. PFPN+SR(e_2) error map w.r.t. different roll and pitch angles (on a subset of test images).

our multimodal spatial rectifier and the effectiveness of our EDINA dataset. A baseline network with our multimodal spatial rectifier (PFPN+MSR) outperforms other baselines on nearly all evaluation metrics, not only on our EDINA dataset but also on FPHA dataset. 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 addition, due to this substantial roll and pitch angles, it results in significant performance degradation for SR($\mathbf{e}_2$) or SR($\mathbf{e}_3$) (which motivates our multimodal spatial rectifier). This is also shown in the performance SR($\mathbf{e}_2$) on depth RMSE with respect to camera roll and pitch angle in Figure 2. Furthermore, the 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 on ScanNet+EDINA datasets. We conclude that our MSR module is highly beneficial for learning egocentric scene geometry.

A baseline PFPN trained on THU-READ, FPHA, and ScanNet performs poorly on EDINA. In addition, the baseline PFPN trained on THU-READ tends to generalize relatively well on FPHA because both datasets include hand-object interactions. 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. Our baseline PFPN trained on ScanNet and EDINA outperforms PFPN trained on other datasets on FPHA. This indicates that learning can greatly benefit from a large amount of high quality ground truth geometry from ScanNet, together with our EDINA.

Comparison of the clustered reference distributions of different datasets. We show in Figure 3 the comparison between ScanNet+EDINA and FPHA surface normal distribution. Note that FPHA has stronger distribution on the tilted modes due to the shoulder mounted camera, which shows the strong generalization capacity of our proposed MSR.

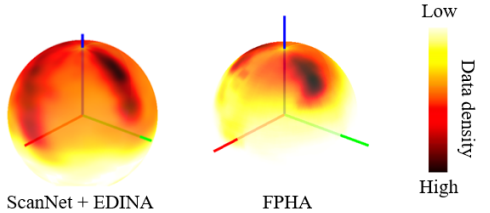

Figure 3. ScanNet+EDINA vs FPHA surface normal distribution.

Qualitative Comparison We show the qualitative comparison on depths and surface normals estimation with and without the multimodal spatial rectifier on Figure 4 and Fig-

ure 5, respectively.

Qualitative Results on EPIC-KITCHENS Figure 6 illustrates the depths, surface normals and gravity prediction on the EPIC-KITCHENS dataset using our multimodal spatial rectifier trained on the ScanNet and our EDINA dataset.

References

[1] 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] Solomon Kullback and Richard Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 1951. 1

[3] Yansong Tang, Yi Tian, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Action recognition in rgb-d egocentric videos. In *ICIP*, 2017. 1

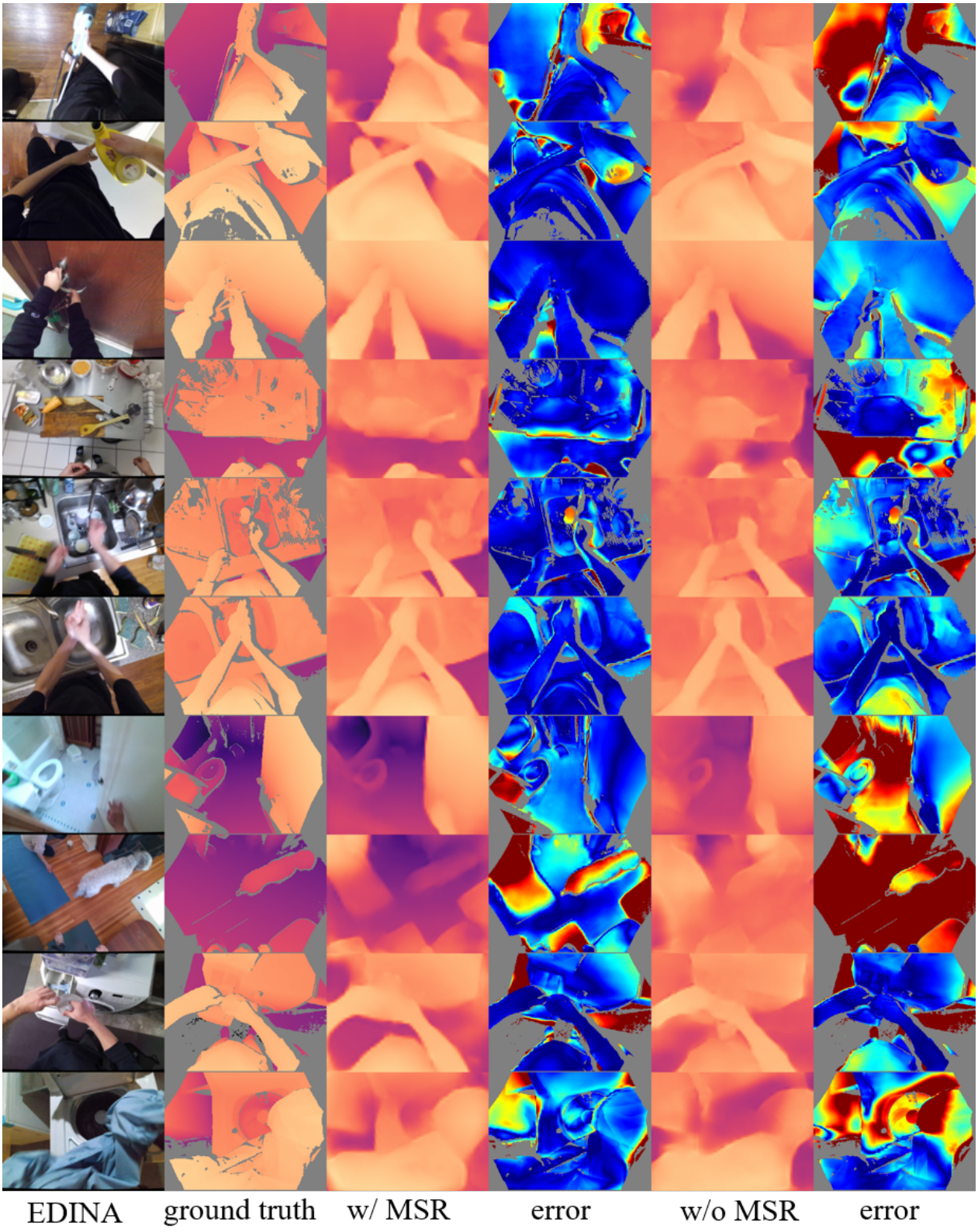

Figure 4. Qualitative results for depth prediction on EDINA dataset. From left to right: (1) RGB image, (2) ground truth depths, (3) depths prediction using PFPN+MSR and its error (the hotter the higher error), and (4) depths prediction using PFPN and its error.

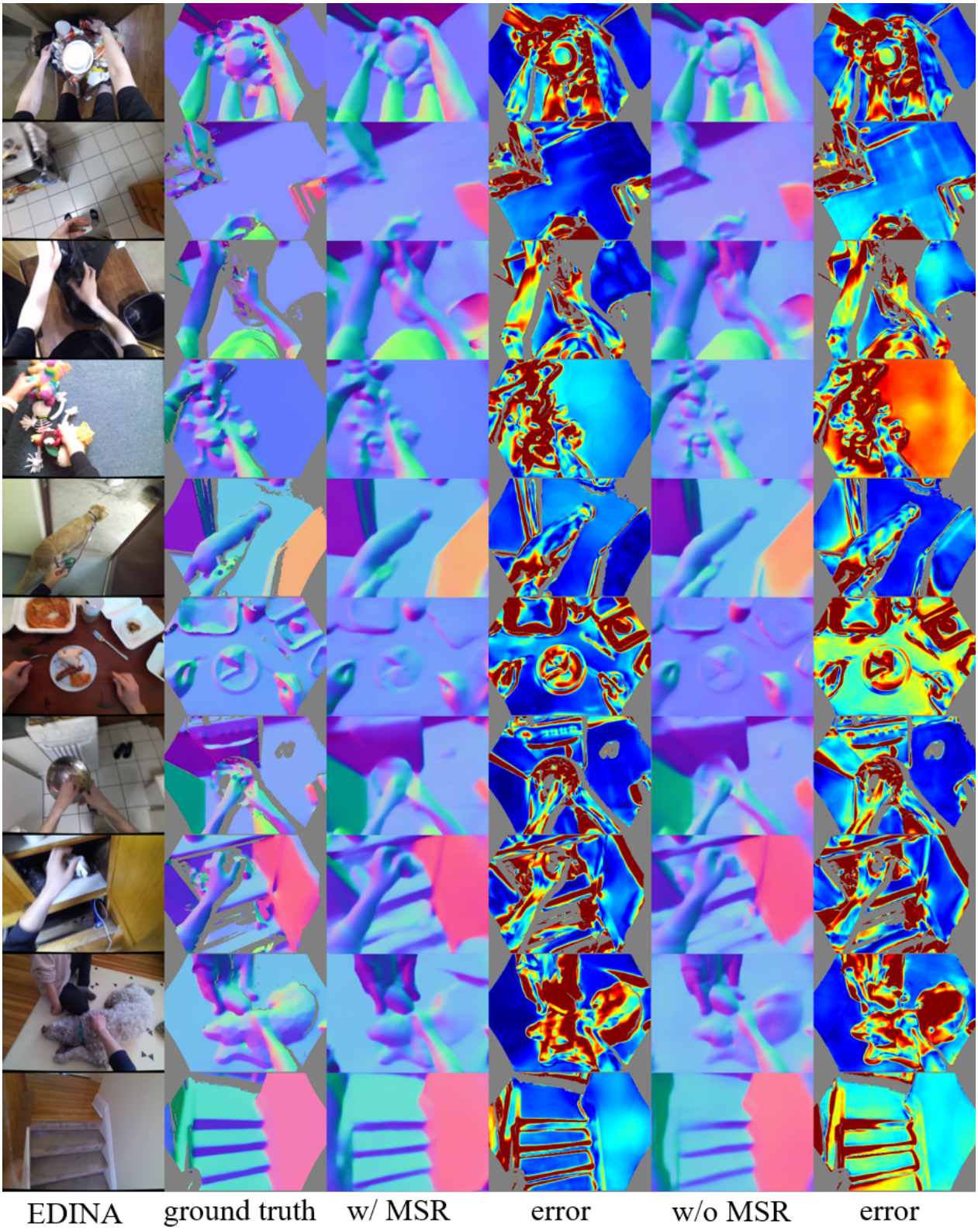

Figure 5. Qualitative results for surface normal prediction on EDINA dataset. From left to right: (1) RGB image, (2) ground truth surface normals, (3) surface normals prediction using PFPN+MSR and its error (the hotter the higher error), and (4) surface normals prediction using PFPN and its error.

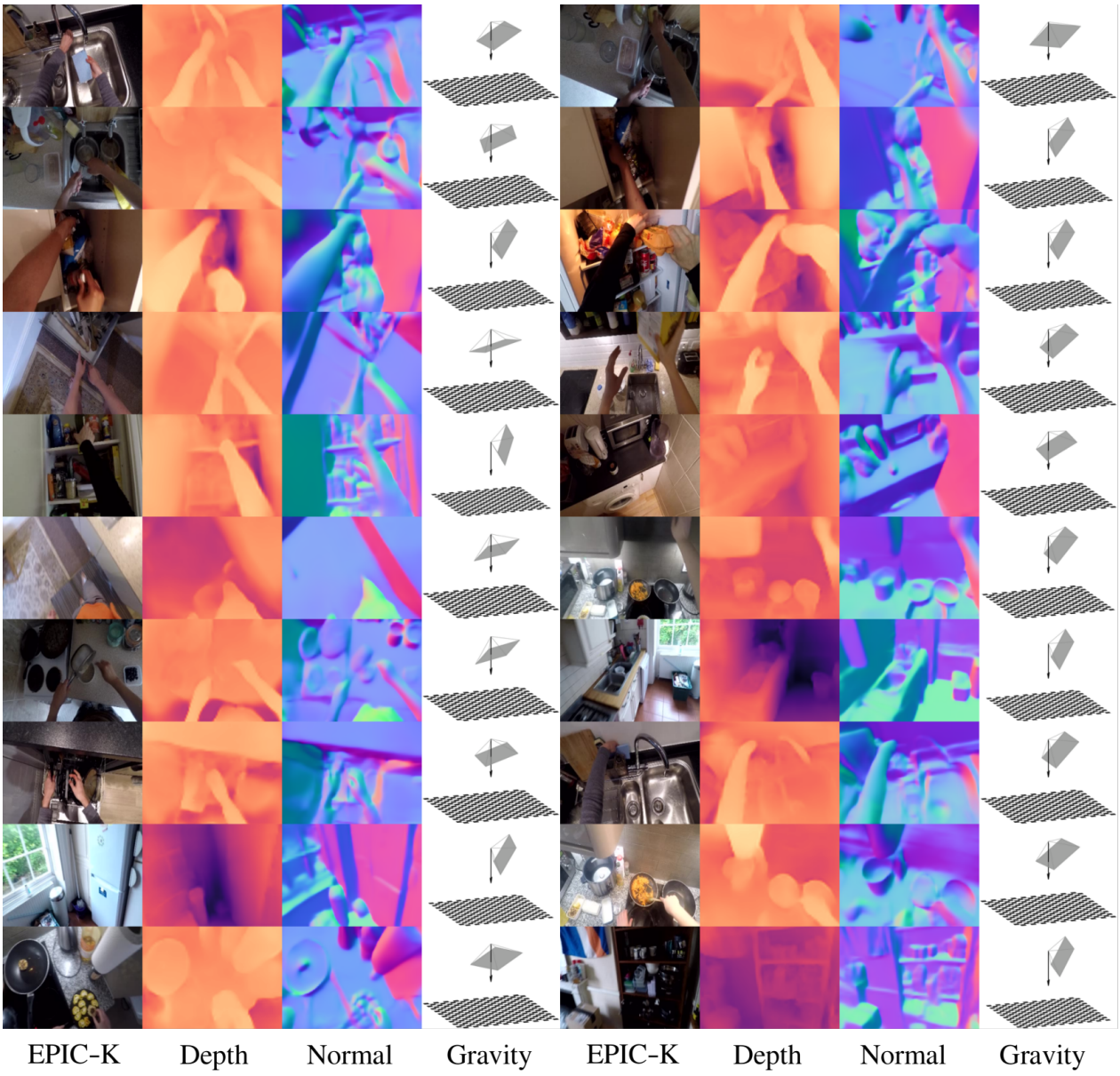

Figure 6. Qualitative results for depth, surface normal, and gravity prediction on EPIC-KITCHENS dataset. In each column, from left to right: (1) RGB image, (2) depth prediction using PFPN+MSR, (3) surface normals prediction using PFPN+MSR, and (4) gravity prediction.