A. Appendix

In this appendix for Single View Scene Scale Estimation using Scale Field, we describe detailed dataset configuration (Appendix A.1), additional implementation details (Appendix A.2), in-depth ablation studies (Appendix A.3) and more qualitative results (Appendix A.4).

A.1. Dataset Configuration

Panorama Dataset. As was described in the main manuscript, we utilize our own panoramic images for panorama dataset for diversity of scenes. Fig. 1 shows some examples of our outdoor panorama images used in scale field dataset generation. Tab. 1 shows the number of training and testing samples from each panorama dataset.

Figure 1. Examples of custom panorama images.

Table 1. **Panorama-based dataset configuration.** We do not utilize custom panorama dataset for evaluation.

Dataset	Train Split		Test Split (#)
	Panos (#)	Crops (#)	
Stanford2D3D	1010	40400	2132
Matterport3D	7608	304320	2012
Custom	2179	87160	-

Web Image Dataset. As was described in the main manuscript, our web image dataset consists of three categories of object, indoor and outdoor, where each of the categories contains 120, 981 and 271 images, respectively. For both training and testing, we square-cropped the image to three crops, so that every crops can cover the whole image. For example, if the image is tall, as in image height is bigger than image width, then three crops would be the top-most crop, center crop and the bottom-most crop. This finally yields 3237 training samples and 852 testing samples.

We further report the distribution of annotated camera heights of our web image dataset in Tab. 2. Since many of the web images are taken by human, very big portion of both training and testing samples have camera heights in range of 1.0∼10m.

Table 2. **Camera height distribution.** Camera height range denoted in meters (m).

Split	Camera Height Range				
	Range1 (∼0.1)	Range2 (0.1∼1.0)	Range3 (1.0∼10)	Range4 (10∼100)	Range5 (100∼)
Train	14	502	2838	71	12
Test	3	101	710	35	3

A.2. Implementation Details

When training all three variants of the networks, *i.e.*, G2H+SF, G2H+CamH and CamParams, each of ground2horizon and scale field was normalized using its mean and variance values, retrieved from Stanford2D3D and Matterport3D datasets. The outputs of FC layers have 256 channels, and were softmaxed and weighted summed by predefined bin values. The bin ranges and distributions for each parameter follow those of [11, 35] with modifications to fit our datasets, and are summarized in Tab. 3.

Table 3. **Bin ranges and distributions for global parameters estimation.** $\mathcal{U}$ and $\mathcal{N}$ refer to uniform and normal distributions, respectively. Horizon line offset is the vertical distance of the horizon line from the center of the image, with the upper left corner set as origin.

Param.	Range	Distribution
Cam H	[0.05m, 300m]	Logscale $\mathcal{U}$
Cam Roll	[-30°, 30°]	$\mathcal{N}$ (0, 20°)
HRZ Offset	[-0.5, 1.0]	$\mathcal{N}$ (0.5, 0.5)
FoV	[15°, 120°]	$\mathcal{U}$

Image augmentation of random scaling with the factor in the range of $[1.0, 1.5]$ and random cropping was applied when training G2H+SF and G2H+CamH models. Since model CamParams requires the principal point of the image to be at the center of the image when converting predicted camera parameters into ground2horizon and scale field, we do not apply image augmentation on CamParams training.

A.3. Ablation Study

In this section, we provide in-depth analysis of our three model variants. Additional to root mean squared error (RMSE) reported in Tab. 3 in the main paper, we provide more metrics for scale field estimation, which are masked root mean squared error (Masked RMSE) and masked relative error (Masked REL). Term 'masked' stands for evaluation only applied on ground pixels, where ground truth scale field value exists. Our scale field is purposely defined only on the ground pixels. So, while RMSE also checks the ability to predict scale field only on valid region, Masked RMSE focuses more on how accurate the predicted scale field is.

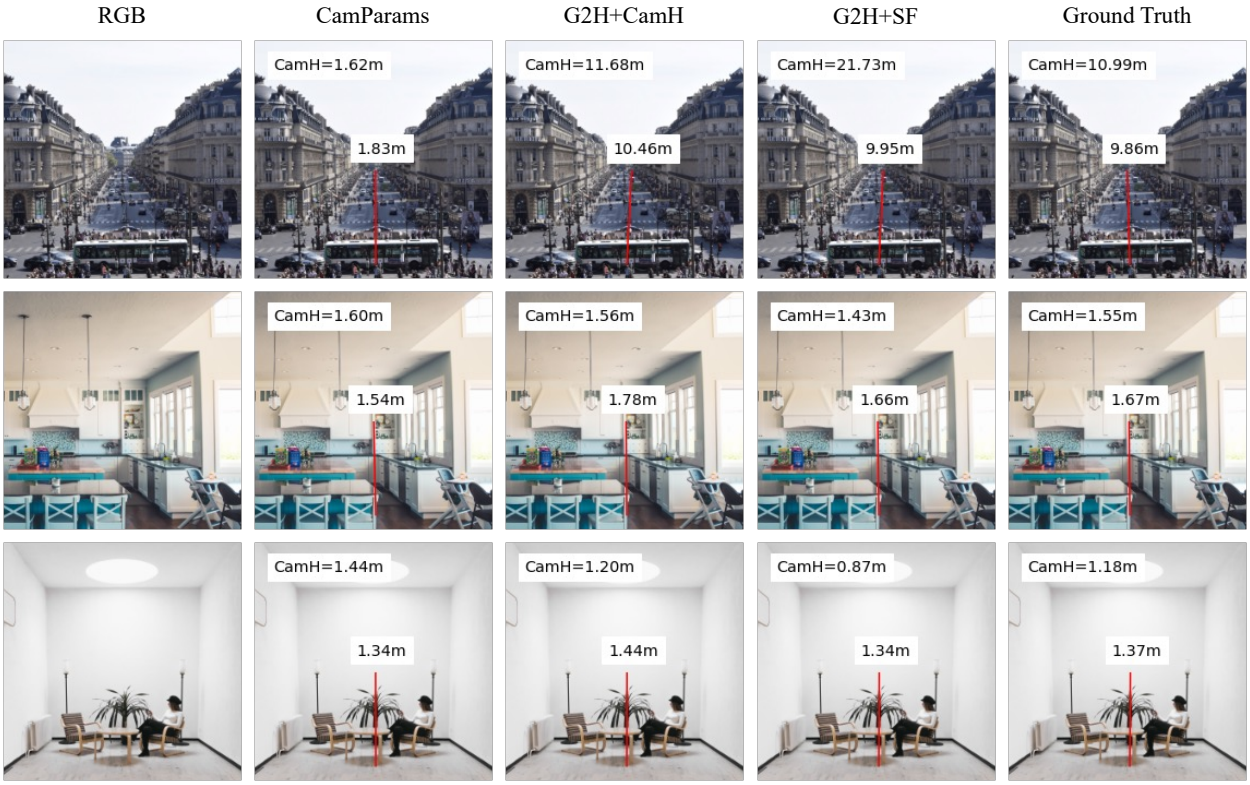

Figure 2. **Qualitative examples showing the robustness of predicting scale field.** CamParams was only trained on P train set. G2H+CamH and G2H+SF were trained on P+W train sets. Predicted metric heights for 100 pixels at position (130, 240) are visualized. Note that while G2H+SF does not predict the best camera height, scale field values are closest to ground truths.

Table 4. **Quantitative evaluation on Stanford2D3D dataset.** Both RMSE and REL metrics are only measured in valid area.

Model	Train Set	Scale RMSE (e^{-2})	REL	Height
CamParams	P	1.553	0.080	**0.039**
G2H+CamH	P	1.486	**0.060**	0.040
G2H+SF	P	**1.474**	0.067	0.043
G2H+CamH	P + W	1.325	0.063	**0.036**
G2H+SF	P + W	**1.274**	**0.062**	0.049

Table 5. **Quantitative evaluation on Matterport3D dataset.** Both RMSE and REL metrics are only measured in valid area.

Model	Train Set	Scale RMSE (e^{-2})	REL	Height
CamParams	P	2.046	0.113	0.082
G2H+CamH	P	1.915	0.105	**0.080**
G2H+SF	P	**1.867**	**0.102**	0.086
G2H+CamH	P + W	1.935	0.106	**0.081**
G2H+SF	P + W	**1.933**	**0.101**	0.085

Table 6. **Quantitative evaluation on Web Image dataset.** Both RMSE and REL metrics are only measured in valid area.

Model	Train Set	Scale RMSE (e^{-2})	REL	Height
CamParams	P	14.18	0.271	**0.555**
G2H+CamH	P	13.96	0.239	0.606
G2H+SF	P	**13.33**	**0.220**	0.563
G2H+CamH	P + W	7.271	0.160	0.331
G2H+SF	P + W	**3.825**	**0.118**	**0.180**

We also show Masked REL, since the scale field values may vary in big margin, according to the camera height. For example, ground2horizon vector with same pixel height may convert to scale field value of $\times 1000$, when substituting camera height from 100m to 0.1m. Following same reason, we also evaluate camera height in REL metric as well.

A.3.1 Additional Model Analysis

Overall results showed in Tab. 4, Tab. 5 and Tab. 6 do not differ from Tab. 3 in the main manuscript. In Stanford2D3D and Matterport3D datasets, camera height is dis-

tributed inside 0.15m range. While model G2H+SF shows the best performance in predicting scale field in all three scale metrics, the difference between three models are relatively small. Nonetheless, in web image dataset, G2H+SF model performs significantly better than other models, as shown in Tab. 6.

Interestingly, while G2H+SF shows comparable but not the best results on camera height prediction, resulted scale field outperforms all the other models. Fig. 2 is a visualization of this observation. For example, in the first row, predicted camera height of G2H+SF is far worse than that of G2H+CamH. However, converted metric height using scale field is much more closer to the ground truth.

In order to retrieve accurate scale field using the outputs of model CamParams and G2H+CamH, all the predicted values must be accurate as well. Our main model G2H+SF directly predicts scale field. It only predicts G2H for orientation information, thus its scale field does not depend on G2H prediction itself. On the other hand, the retrieved camera height from G2H+SF is the averaged value of $SF/\|\text{ground2horizon}\|$. It will be more sensitive to the accuracy of all the outputs, compared to CamParams and G2H+CamH, which predict camera heights directly.

We conclude the analysis by stating that our scale field formulation does not depend on series of relevant parameter predictions. Thereby we argue that it is a more robust way to both represent and train scale information for deep learning based approaches.

A.3.2 Performance Analysis in Different Camera Height Ranges

We further analyze two of our models, G2H+SF and G2H+CamH, by evaluating both models on each of the camera height ranges described in Tab. 2. Overall results are shown in Tab. 7. It is shown that G2H+SF shows better performance in the most of the metrics, in all the camera height ranges. Similar tendency can be found as in Appendix A.3.1, where the difference between two models are relatively small in range of 1.0∼10m of camera heights, while in other ranges, G2H+SF performs significantly better. Unfortunately, in Range5 (100m∼), both models fail to predict reasonable scale field, therefore evaluations were not able.

Camera height estimation heads in both G2H+CamH and CamParams are classification-based. It is inevitable for these heads to be more fitted to certain camera height range that they encounter most frequently. 1.0∼10m of camera height in our web image dataset possess nearly 88% of the whole dataset distribution. Adding panorama dataset into the statistics, the data bias becomes even more significant. This explains why CamParams and G2H+CamH models work poorly on other camera height ranges. For example, in Fig. 2, model CamParams, which was only trained on panorama-based dataset, only outputs camera heights within certain narrow range.

Table 7. **Camera height range-wise quantitative evaluation on Web Image dataset.** Both G2H+CamH and G2H+SF models were trained on P+W train set. Both RMSE and REL metrics are only measured in valid area. Camera height range denoted in meters (m).

Range	Model	Scale RMSE	Scale REL	Height
Range1 (∼0.1)	G2H+CamH	$1.984e^{-2}$	0.460	1.256
	G2H+SF	$\mathbf{0.624e^{-2}}$	**0.259**	**0.187**
Range2 (0.1∼1.0)	G2H+CamH	$2.723e^{-3}$	0.475	0.487
	G2H+SF	$\mathbf{1.472e^{-3}}$	**0.259**	**0.173**
Range3 (1.0∼10)	G2H+CamH	$2.574e^{-4}$	0.120	0.150
	G2H+SF	$\mathbf{2.214e^{-4}}$	**0.103**	**0.129**
Range4 (10∼100)	G2H+CamH	$3.271e^{-5}$	0.035	**0.412**
	G2H+SF	$\mathbf{1.205e^{-5}}$	**0.013**	0.902
Range5 (100∼)	G2H+CamH	-	-	-
	G2H+SF	-	-	-

However, we argue that this is not due to unfair comparison, since many of the arbitrary image found in web are human-taken. Therefore, this kind of camera height distribution do not deviate from the regular statistics. Moreover, while G2H+SF was also trained under the same configuration using the same datasets as G2H+CamH, it is much more robust on less-seen camera height ranges. We once again stress that our scale field is a more robust way to train neural network for scale estimation.

A.4. Additional Qualitative Results

We provide more qualitative results in Fig. 3, Fig. 4 and Fig. 5, categorized by camera height ranges, to show that our single view scale estimation network (G2H+SF) can robustly predict scale field in various types of images, compared to other model variants.

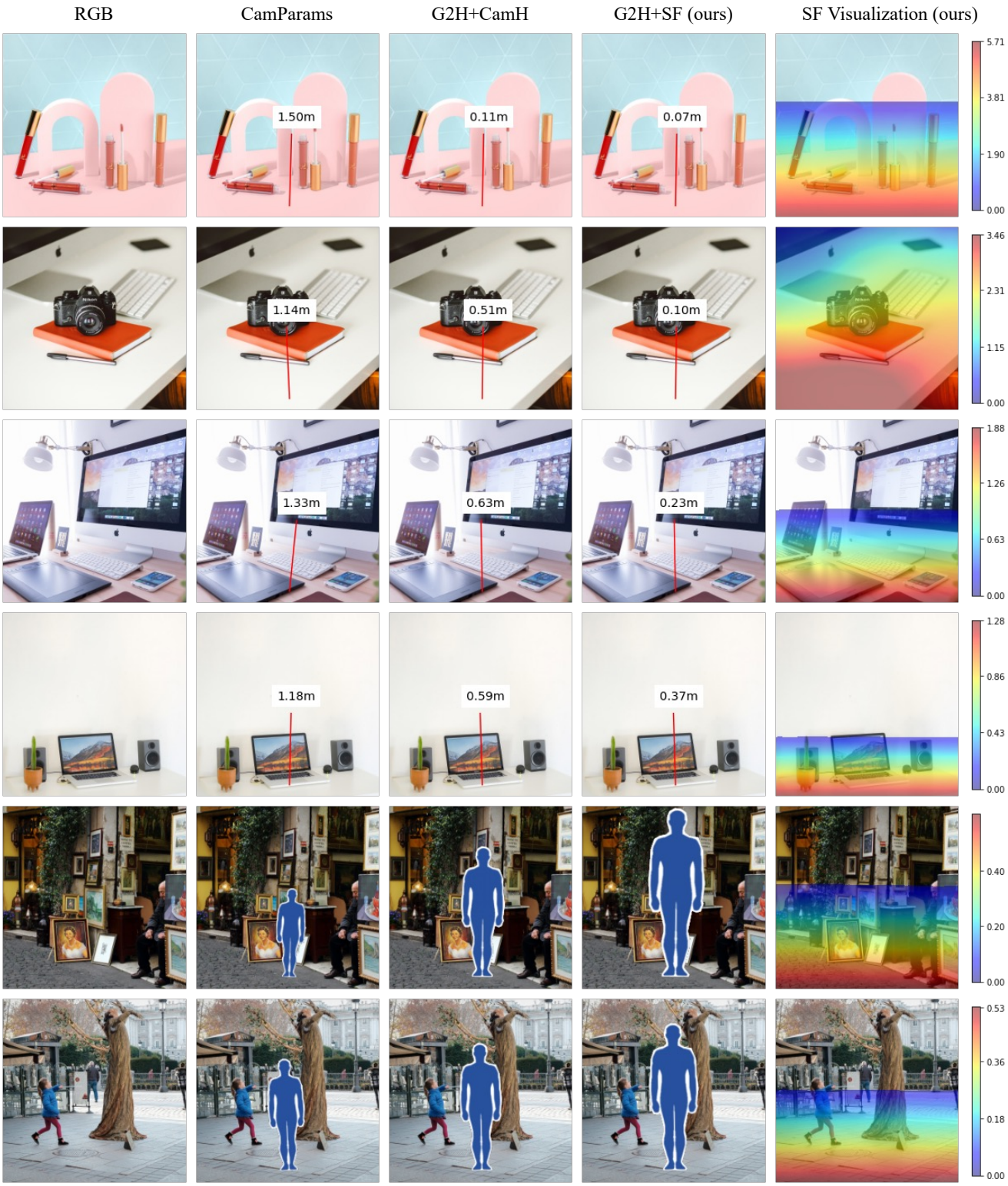

Figure 3. **Qualitative comparison on Range1 and Range2 images.** Inserted human silhouette set to have height of 1.7m. Metric height measured at pixel coordinate (130, 240) for all examples.

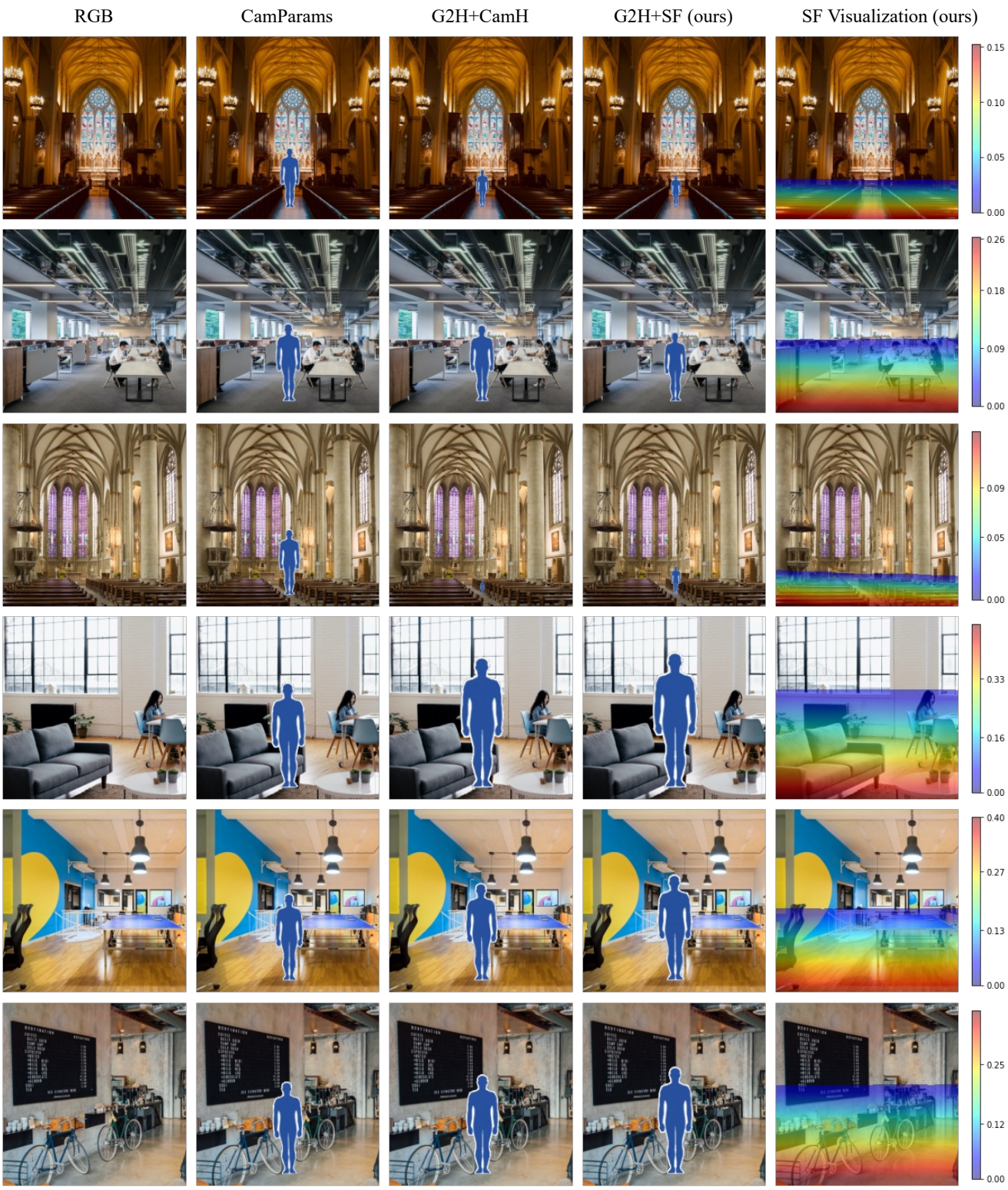

Figure 4. **Qualitative comparison on Range3 images.** Inserted human silhouette set to have height of 1.7m.

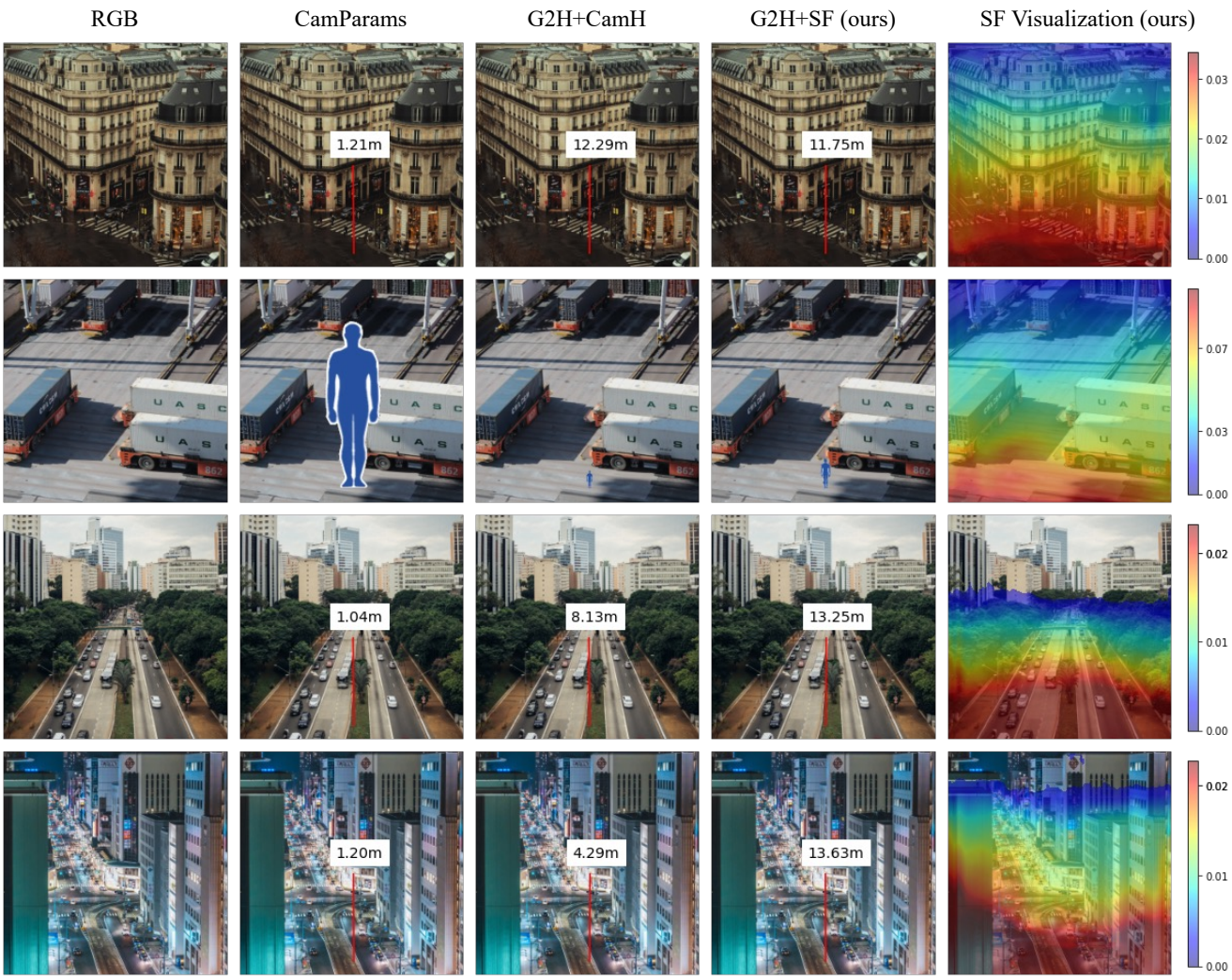

Figure 5. **Qualitative comparison on Range4 images.** Inserted human silhouette set to have height of 1.7m. Metric height measured at pixel coordinate (130, 240) for all examples.