Supplementary Material: Reprojection Errors as Prompts for Efficient Scene Coordinate Regression

Ting-Ru Liu[1], Hsuan-Kung Yang[2], Jou-Min Liu[1*], Chun-Wei Huang[1*], Tsung-Chih Chiang[1*], Quan Kong[2], Norimasa Kobori[2], and Chun-Yi Lee[1]

[1]National Tsing Hua University [2]Woven by Toyota, Inc.
{tingru, diasimui, alan, george, cymaxwelllee}@elsa.cs.nthu.edu.tw
{hsuan-kung.yang, quan.kong, norimasa.kobori}@woven.toyota

1 Training Process of EGFS

To further clarify our EGFS methodology, we present the pseudo-code of the training steps in Algorithm 1.

2 Ablation Study on Model Size

We adjusted the number of layers in the MLP to compare the impact of different model sizes. The results, presented in Table 1, cover a series of experiments across a range of model sizes, from 3MB to 6MB. It is observed that larger models generally improve localization performance. Nevertheless, this performance enhancement becomes less pronounced as the size increases. Therefore, a model size of 4.5MB was utilized for the experiments presented in the main manuscript.

3 A Comparison between EGFS and Semantic Masks

Utilizing semantic masks for feature selection may not be appropriate for all scenes, as mentioned in Section 4.2. To compare the performance of semantic masks and our proposed EGFS masks, we conducted further experiments and present the results in Table 3. Based on the analysis of reprojection error and semantics in Section 4.2, we selected the union of the three semantic categories with the lowest median reprojection errors in each scene of the Indoor6 dataset [1], as shown in Table 2. We then trained a scene-specific MLP for performing SCR using the areas covered by these semantic labels. The results demonstrate that utilizing semantic masks does not achieve the same level of performance as our EGFS methodology. This finding highlights the advantage of EGFS in eliminating the requirement for pre-defined semantic categories and its ability to generalize across scenes.

Algorithm 1 The training process of EGFS.

I: input image, h^*: ground truth camera pose
$\mathcal{Y}$: scene coordinates, $\mathcal{C}$: confidence map, $\mathcal{R}$: reprojection errors

function EGFS-MASK-GENERATION(f_H)
    $\mathcal{Y}, \mathcal{C} \leftarrow f(I)$
    $\mathrm{r_i} \leftarrow r(\mathrm{p_i}, \mathrm{y_i}, \mathrm{h}^*)$, where $\mathrm{r_i} \in \mathcal{R}$, $\mathrm{p_i} \in I$, $\mathrm{y_i} \in \mathcal{Y}$
    $prompts \leftarrow Q(\mathcal{R}, \tau)$
    $masks \leftarrow SAM(prompts) \cap (\mathcal{C} > \mu)$
    return $masks$
end function

function TRAINING
    `// Initial iteration`
    Sample features f from the entire image sequencce
    Use f to train scene-specific MLP f_H for k epochs
    $masks \leftarrow$ EGFS-MASK-GENERATION(f_H)

    `// Subsequence iterations`
    for $t = 2$ to T **do**
        Sample features f from $masks$
        Use f to train scene-specific MLP f_H for k epochs
        $masks \leftarrow$ EGFS-MASK-GENERATION(f_H)
    end for
end function

4 Hyperparameter Setup

Table 4 presents the hyperparameter setup employed in our experiments. Our model training involves a batch size of 5,120, with the training duration set at 20 epochs, and masks are generated every five epochs. Each iteration takes eight million features into the training buffer to train the scene coordinates. We employ the AdamW optimizer [2] with a one-cycle learning rate schedule [3], where the rates range from 5×10^{-4} to 5×10^{-3}. In addition, the point prompts are selected as the lowest 10% of points based on their reprojection errors, and the parameter that balances the confidence regularization term is set to ten.

5 Additional Visualizations and Qualitative Results

To further substantiate the effectiveness of our proposed EGFS methodology, additional visualizations and qualitative results are presented in Figs. 1 and 2.

Table 1: Analysis on the impact of model sizes. For the Cambridge dataset, the numbers reported represent the median translational and rotational errors in cm and degrees, respectively. For the Indoor6 dataset, the percentages reported represent the proportion of reprojection errors less than 5 cm in translation and 5 degrees in rotation.

Dataset	Scene	3MB	3.5MB	4MB	4.5MB	5MB	5.5MB	6MB
Cambridge	King's College	15/0.3	16/0.3	14/0.3	14/0.3	15/0.3	14/0.3	14/0.3
	Great Court	33/0.1	34/0.1	33/0.1	31/0.1	31/0.1	31/0.1	31/0.1
	Old Hospital	19/0.4	19/0.4	22/0.4	21/0.4	19/0.4	19/0.4	21/0.4
	Shop Facade	6/0.3	5/0.3	5/0.3	5/0.3	5/0.2	5/0.3	5/0.3
	St Mary's Church	16/0.5	16/0.5	15/0.4	15/0.5	15/0.4	15/0.4	15/0.5
	Average	18/0.3	18/0.3	18/0.3	17/0.3	17/0.3	17/0.3	17/0.3
Indoor6	scene1	38.4%	42.4%	44.8%	46.4%	45.6%	45.3%	46.1%
	scene2a	54.7%	56.4%	57.6%	60.6%	56.8%	60.7%	62.6%
	scene3	48.1%	52.7%	55.2%	56.4%	54.9%	57.8%	56.4%
	scene4a	68.4%	76.6%	75.3%	78.7%	75.3%	76.6%	80.6%
	scene5	23.5%	24.1%	25.5%	22.8%	25.2%	27.8%	28.0%
	scene6	63.4%	67.2%	71.5%	71.6%	73.7%	75.9%	75.9%
	Average	49.4%	53.2%	55.0%	56.1%	55.3%	57.4%	58.3%

Table 2: The top three semantic labels with the lowest reprojection errors for each scene in the Indoor6 dataset, along with the union of these labels.

Scene	Semantic Labels with Low Errors
scene1	Fireplace, Sofa, Armchair
scene2a	Painting, Wall, Door
scene3	Rug, Sofa, Wall
scene4a	Painting, Shelf, Door
scene5	Painting, Door, Shelf
scene6	Refrigerator, Counter, Cabinet
Union	Fireplace, Sofa, Armchair, Painting, Wall, Door, Rug, Shelf, Refrigerator, Counter, Cabinet

Table 3: Comparison of the effectiveness between semantic masks and our proposed EGFS masks for feature selection on Indoor6.

Scene	*Semantic Masks*	*EGFS Masks*
scene1	40.9%	46.4%
scene2a	54.5%	60.6%
scene3	54.6%	56.4%
scene4a	69.6%	78.7%
scene5	25.5%	22.8%
scene6	67.2%	73.7%
Average	52.1%	56.1%

Table 4: The hyperparameters utilized in our experiments.

Hyperparameters	EGFS
Training buffer size	8M
Batch size	5,120
Epochs	20
Learning rate	$[5 \cdot 10^{-4}, 5 \cdot 10^{-3}]$
Optimizer	AdamW
Proportion of point prompts	10%
Confidence regularization parameter	10

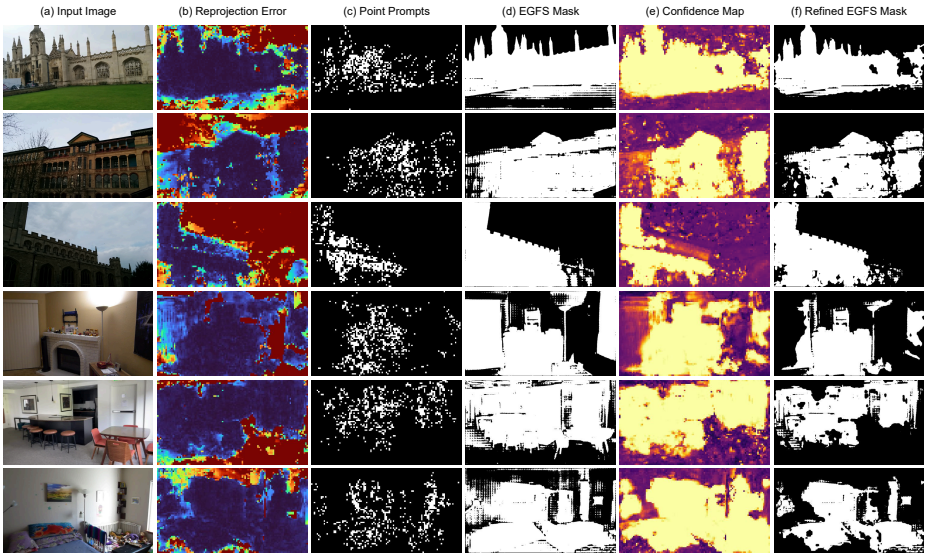

Fig. 1: Additional visualizations of the reprojection errors, point prompts, confidence maps, and EGFS masks before and after confidence refinement.

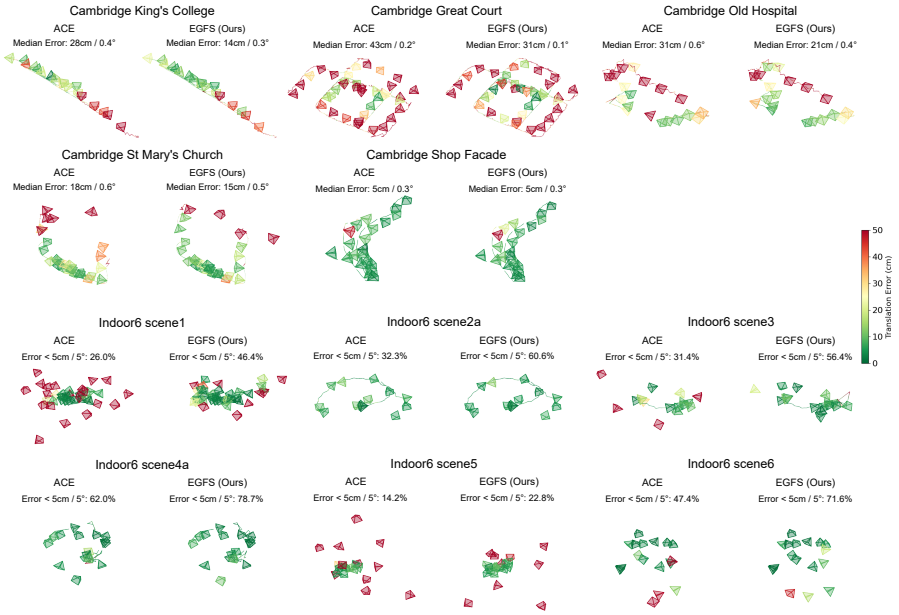

Fig. 2: Additional visualizations of the estimated camera pose trajectories from testing sequences with the camera frustums colored based on translational errors. Pose errors denoted.

References

1. Tien Do, Ondrej Miksik, Joseph DeGol, Hyun Soo Park, and Sudipta N. Sinha. Learning to detect scene landmarks for camera localization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1
2. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2019. 2
3. Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, 2019. 2