Supplementary Material for Self-Supervised Domain Adaptation for Visual Navigation with Global Map Consistency

1. Experiment Setup Details

In all our experiments, we use the agent from Active Neural SLAM [1] which can move forward by 25cm, and turn right and left by 10 degrees. We use a batch size of 4, Adam optimizer [3] and learning rate of 5×10^{-5} with an RTX 2080 GPU.

All experiments initiate with the pre-trained model which is trained with ground-truth supervision in 72 scenes from the Gibson training set with no actuation and odometry noise. While the pre-training trains all of the modules jointly, fine-tuning only trains the localization model in Figure 1 of the main manuscript. The fine-tuning and evaluation setup varies for each experiment and the details are summarized in Figure 1 and Table 1. Below we further clarify details about the conducted experiments from the main paper.

Standard Experiment In the standard experiments from Section 4.1, the pre-trained agent is fine-tuned on the 72 scenes from the training set with the presence of various odometry and actuation noise. The fine-tuning runs for 4000 iterations on a set of 160 round-trip trajectories of 100 agent timesteps. As seen in Figure 1, evaluation is performed on the validation split of the Gibson dataset [5], with noisy actuation and odometry noise. The contents of the evaluation are three-fold, namely localization, mapping, and exploration. For localization and mapping, each fine-tuned agent is tested on 140 trajectories collected from the validation scenes. The length of validation trajectories is 100 unless specified otherwise. For the exploration task, we report the result of the agent exploring the 14 validation scenes with 10 random starting points from each scene.

In all experiments, both the GT Train and our model are fine-tuned in the same setup except that the GT Train trains with ground-truth pose supervision during fine-tuning. Domain randomization is fine-tuned on about an order of magnitude more number of trajectories for larger number of iterations: we used 1500 trajectories with random odometry and actuation noise and the agent is trained for 40000 iterations, also with ground-truth supervision.

Test-time Adaptation In Section 4.2, we report the test-time adaptation results. The main difference from the standard setup is that the agent is fine-tuned on a single validation scene with odometry and actuation noise as shown in Figure 1. Another difference is the fast and light-weight adaptation. The agent fine-tunes for 2000 iterations on 20 trajectories from a single scene, compared to 4000 iterations and 160 trajectories in the standard experiment. The evaluation for localization and mapping runs on 80 trajectories from the same scene. Note that the sets of trajectories for fine-tuning and evaluation do not overlap. Evaluation on exploration is performed in the identical setup as the standard experiment, with area coverage measured against 10 random starting points for each validation scene. After being fine-tuned and evaluated on each of 14 validation scenes, the mean value of each metric across all scenes is reported, as depicted in Figure 1.

Ablation Study The ablation study follows the same setup used in the standard experiment. The pre-trained agent fine-tunes for 4000 iterations on 160 trajectories both for self-supervised learning and ground-truth supervised learning. The fine-tuned agent is evaluated on 140 trajectories on the noisy validation scenes.

2. Actuation Noise Details

We adopt the actuation noise model similar to RobustNav [2]. Three types of actuation noise are tested in Section 4.1, namely constant / stochastic motion bias and motion drift. The motion bias induces the constant and stochastic bias. For each forward step (note that the noiseless action moves by 0.25m), the constant bias δ_c is randomly selected from [±0.05m, ±0.1m, ±0.15m]. For the right or left turn action (compared to the original action command turns by 10°), the constant bias is drawn from [±1.7°, ±3.3°, ±5°]. In addition, the stochastic bias δ_s is drawn from $\mathcal{N}(\mu_s, \Sigma_s)$ where μ_s = **0** and Σ_s

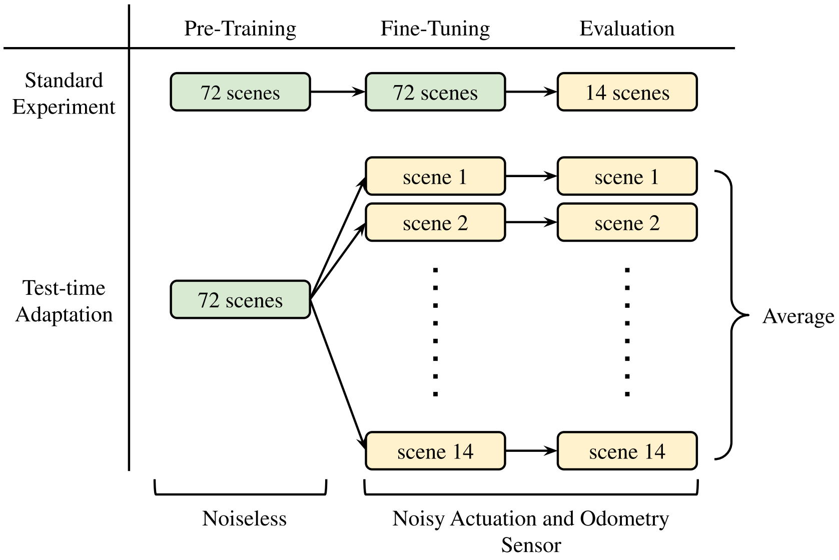

Figure 1. **Training setup for standard experiment and test-time adaptation.** Scenes shaded in green indicate the train split and yellow is the val split from the Gibson dataset [5, 4]. In the standard experiment, the pre-trained agent is fine-tuned in the train scenes. Within the same train scenes, it learns to adapt to the added actuation and odometry noise which it has never seen before. The fine-tuned agent is then evaluated on unseen validation scenes where the same noise is present. In test-time adaptation, the pre-trained agent is fine-tuned on 20 trajectories, and evaluated on 80 trajectories from a single scene. As a result, we compute the average performance across the 14 validation scenes.

	Standard Experiment (Section 4.1)	Test-time Adaptation (Section 4.2)	Ablation Study (Section 4.3)
Pre-Training			
Dataset split	Train	Train	Train
Actuation and Odometry Noise	Noiseless	Noiseless	Noiseless
Fine-Tuning			
Dataset split	Train	Val	Train
Actuation and odometry noise	Noisy	Noisy	Noisy
No. of trajectories	160	20	160
Trajectory length	100	100	100
Training iterations	4000	2000	4000
Evaluation			
Dataset Split	Val	Val	Val
Actuation and odometry noise	Noisy	Noisy	Noisy
No. of trajectories	140	80	140
Trajectory length	100	100	100

Table 1. Training details for standard experiment, test-time adaptation, and ablation study.

$= \left(\begin{smallmatrix}0.1&0&0\\0&0.1&0\\0&0&3.3\end{smallmatrix}\right)$. Each diagonal value corresponds to position (x, y), and orientation ϕ. Lastly, the motion drift is uniformly drawn from [No drift, Drift right by 10°, Drift left by 10°]. The aforementioned magnitudes of noise parameters are chosen considering the action space based on the values suggested in RobustNav [2].

Metric	BCE	MSE
Mean	1.02e-05	3.74e-07
Max	0.0044	0.0002
Variance	4.18e-09	5.63e-12

Table 2. Gradient Analysis on BCE and MSE

3. Gradient Analysis on BCE and MSE

In ablation study, we discussed the analysis of the use of BCE and MSE to compare binary occupancy maps. In Table 2, we report the mean, max, and variance of gradient magnitudes collected from training in the standard experiment. Compared to MSE, the higher variance and range in BCE loss lead to unstable fine-tuning, and therefore the loss is difficult to converge.

4. Additional Qualitative Results

We display additional qualitative results of mapping in Figure 2 and localization in Figure 3. Similar to Figure 4 from the original manuscript, we show global maps and trajectories generated from the same sequence of actions across various adaptation methods. As observed before, the qualitative results on mapping and trajectory estimation are close to the results with GT supervision, whereas agents without adaptation or domain randomization still suffer from the accumulated error.

References

[1] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155*, 2020.

[2] Prithvijit Chattopadhyay, Judy Hoffman, R. Mottaghi, and Aniruddha Kembhavi. Robustnav: Towards benchmarking robustness in embodied navigation. *ArXiv*, abs/2106.04531, 2021.

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[4] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[5] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9068–9079, 2018.

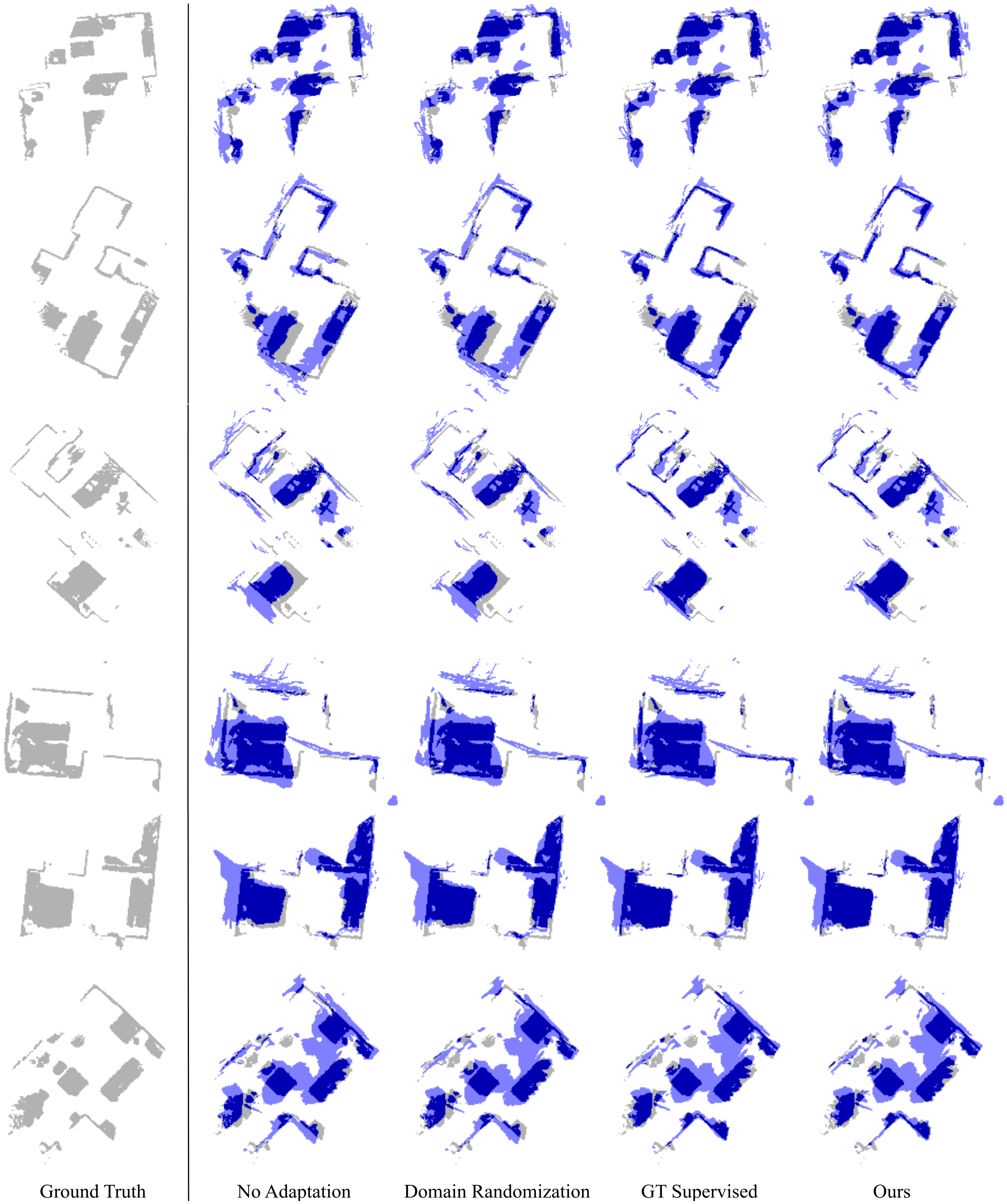

Figure 2. **Qualitative Result of Mapping.** We show more qualitative result of the global maps. Each row corresponds to the maps generated from the same sequence of actions in a single scene. The generated map is indicated in blue and the ground truth map is indicated in gray.

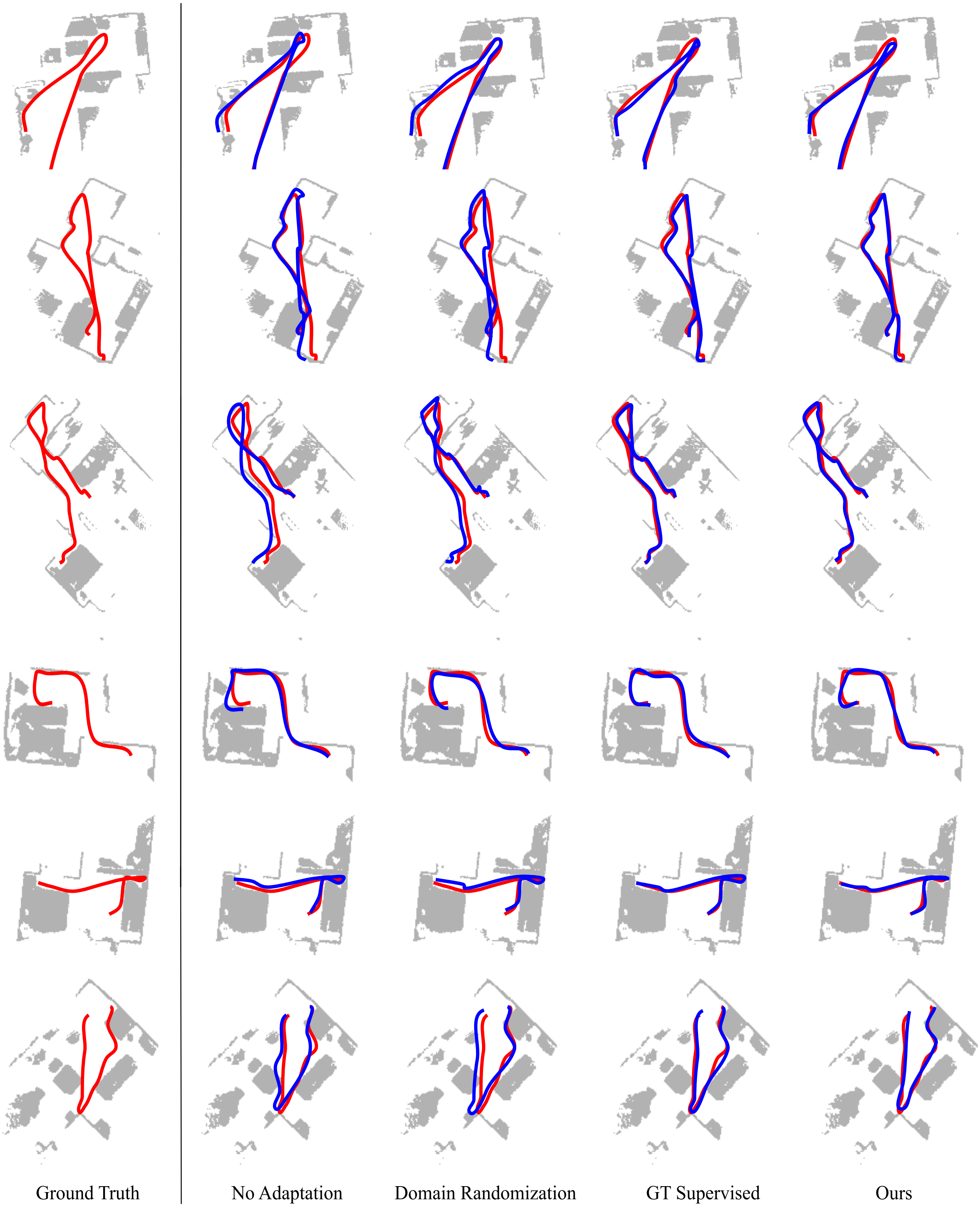

Figure 3. **Qualitative Result of Localization.** We show more qualitative result of the pose trajectories. Each row corresponds to the trajectories generated from the same sequence of actions in a single scene. The generated trajectory is indicated in blue and the ground-truth trajectory is indicated in red.