Supplementary Material for KG-Nav: Bridging the Gap between Visual Servoing and Image-Goal Navigation via Keypoint Graph-Based Reinforcement Learning

Author Names Omitted for Anonymous Review. Paper-ID 395

I. Robot Experiments

We conduct a zero-shot transfer to the real world as shown in Figure 1. We deploy a mobile robot with a perspective camera and a panoramic camera. Depicted in the figure are two paths that are taken day and night, respectively. Our method generalizes well in unseen environments due to the use of the keypoint graph constructed with local interest points. We report on the experimental details and results in the following subsections.

A. Experiment Settings

To implement KG-Nav for a real-world setting, we deploy a Clearpath Jackal robot. Our robot is equipped with an Intel RealSense D435 and a Ricoh Theta V camera for the perspective camera setting and the panoramic camera setting, respectively. For the perspective camera setting, we use the RGB image captured by the D435 which has a size of 640×360 and a field of view (FoV) of 69 °×42 °. For the panoramic camera setting, we use the equirectangular image which is captured by the Theta V and resized to 512×256. We can adapt the policy of our KG-Nav agents trained with the different image sizes since we use the normalized coordinates for computing the flow features of the keypoint graph. Therefore, we directly apply our KG-Nav and KG-Nav-Equirect agent trained with the Gibson dataset [1] and test the zero-shot transfer performance of the KG-Nav agents.

For test sites, we choose three indoor places and two outdoor places on the campus which are entrances to an elevator, a stationery store, and a bank inside a building and two entrances to a building from outside. With initial distances of 1.5∼10 m from the target viewpoints, we test three episodes per test site, thereby 15 episodes in total. The number of keypoints we extract from the observation is the same as in simulation experiments and the stopping condition ρ_{th} is 0.5 in the real-world experiments. Exceeding the episode length of 100 or colliding with obstacles also terminates an episode. We measure the approximate distance at the agent's final location to the target viewpoint with Perspective-n-Point (PnP) RANSAC methods.

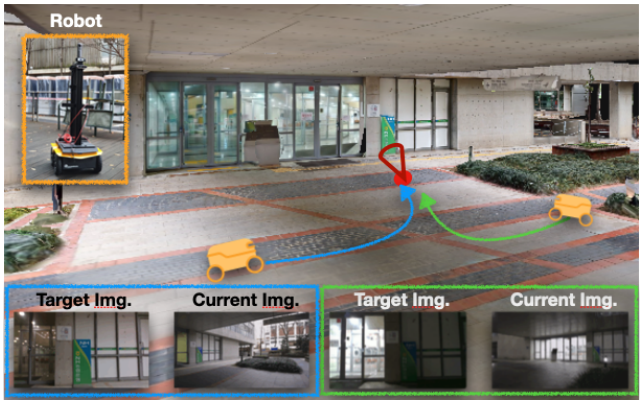

Fig. 1. Sim2Real Experiments. Two paths and the target and currently observed image of each path are depicted.

B. Quantitative Results

In the perspective camera setting, the KG-Nav agent succeeded with a success rate of 73.3%, and an average final distance 1.65 m. In the perspective camera setting, the KG-Nav-Equirect agent succeeded with a success rate of 46.7%, and an average final distance 0.80 m. OVRL [2] could not succeed in any episode since their offline representation learning is overfitted to the training environment. On the other hand, KG-Nav is shown to generalize well in unseen environments thanks to the use of the keypoint graph constructed with local interest points. In addition, its performance is mostly affected by the FoV of the observations.

C. Qualitative Results

We provide qualitative results in the various settings in the attached video.

References

[1] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson Env: Real-World Perception for Embodied Agents," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] K. Yadav, R. Ramrakhya, A. Majumdar, V.-P. Berges, S. Kuhar, D. Batra, A. Baevski, and O. Maksymets, "Offline Visual Representation Learning for Embodied Navigation," *arXiv:2204.13226*, 2022.