Language Embedded Radiance Fields for Zero-Shot Task-Oriented Grasping Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

A LERF-TOGO

A.1 Implementation Details

We implement LERF-TOGO on top of the Nerfacto method from Nerfstudio [1]. For faster convergence and smoother optimization, we modify several parameters from the original LERF paper. We use a smaller hashgrid with 16 levels and a maximum resolution of 256, and find that using larger MLPs for the density, color, and transient output heads of NeRF results in faster convergence and better ability to handle specularities and robot shadows. We introduce weight decay of 1e-7 to the LERF network which smooths training. In addition, we compress the DINO embeddings into 128 dimensions and supervise on these vectors rather than the original DINO outputs, but we do not normalize the resulting vectors like Tschernezki et al. [2]. While constructing the CLIP embedding pyramid for LERF, we use crops ranging from 5% to 35% of image height with 6 pyramid levels, biasing the pyramid to smaller crops as LERF-TOGO is primarily interested in object and part queries.

B Grasping

Point Cloud Extraction To extract a scene-wide point cloud for grasping, we use the method in Nerfstudio [1], which deprojects randomly sampled rays' depth from the train camera poses, then filters with outlier rejection. We then crop the point cloud to the workspace of the robot. For object centric point clouds, we deproject depth from views radially surrounding the object of interest.

Motion Planning A grasp is considered feasible if the robot can perform a collision-free trajectory with the following poses: the pre-grasp, grasp, and post-grasp configurations. The pre-grasp pose is positioned 5cm along the z-axis of the robot end effector, which allows the gripper to approach the target grasp pose with minimal additional motion. The post-grasp pose is located 10cm above the grasp pose, along the z-axis of the world frame. The UR5's IK solver calculates the set of viable joint configurations at these poses, and we calculate the trajectory as a linear interpolation between them. We additionally allow for a 180 degree rotation at the last wrist joint, as parallel-jaw grasps are rotationally symmetric. This facilitates the motion planning process, as the robot's camera mount is prone to colliding with the robot arm.

C Experiments

C.1 Setup Details

We use a UR5 arm with a Logitech BRIO webcam at 1600x896 resolution, with all camera settings frozen before each capture prevent color discrepancies among images. The camera mount points

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

Scene	Object Query ; Part Query
Kitchen	(black matte spoon, handle), (shiny black spoon, handle), (teapot, handle), (dish scrub brush, handle), (dust brush, handle)
Flowers	(daisy, plant stem), (rose, plant stem)
Mugs	(blue mug, handle), (pink teacup, handle), (turqouise mug, handle), (white mug, handle), (black mug, handle)
Tools	(Measuring tape, base), (screwdriver, handle), (wire cutters, handle) (soldering iron, handle), (hammer, handle)
Knives	(bread knife, handle), (steak knife, handle), (box cutter, handle)
Martinis	(red martini glass, stem), (grey martini glass, stem)
Fragile	(camera, strap), (pink sunglasses, earhooks) (blue sunglasses, earhooks), (lightbulb, screw)
Cords	(power strip, plug), (power strip, base), (ethernet dongle, usb) (ethernet dongle, ethernet)
Messy	(ice cream, cone), (green lollipop, stick), (blue lollipop, stick)
Pasta	(wine, cork), (wine, bottle neck), (saucepan, lid knob) (saucepan, handle), (corkscrew, handle)
Cleaning	(clorox, wet towel), (clorox, lid), (clorox, body) (tissue box, box), (tissue box, tissue)
Bottles	(meyer's cleaning spray, spray trigger), (meyer's cleaning spray, bottle neck) (meyer's cleaning spray, body) (purple cleaning spray, body) (purple cleaning spray, spray trigger), (purple cleaning spray, bottle neck)

Table 1: **Complete list of object and part queries**

orthogonally to the gripper axis, to maximize the reachable workspace of the camera while pointing towards the workspace center. During robot capture, pre-computation of DINO, CLIP, and ZoeDepth is parallelized across 3 NVIDIA 4090 GPUs to achieve real-time performance, and all subsequent operations are carried out on a single 4090. Capturing a scene takes 30 seconds, training the LERF to 2k iterations takes 78 seconds, and finally querying LERF-TOGO takes 10 seconds.

References

[1] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, et al. Nerfstudio: A modular framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*, 2023.

[2] V. Tschernezki, I. Laina, D. Larlus, and A. Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2022.