ASCOMP-2024
Fall 2024

NeRF with semantics for reconstruction of 3D scene and object segmentation by text prompts

Julia Kudryavtseva[1]

Problem Statement

Develop a system capable of matching the points in a 3D scene with their respective visual-semantic meaning. Given a text prompt the system is capable of segmenting objects in 3D space and extracting their coordinates in the real world.

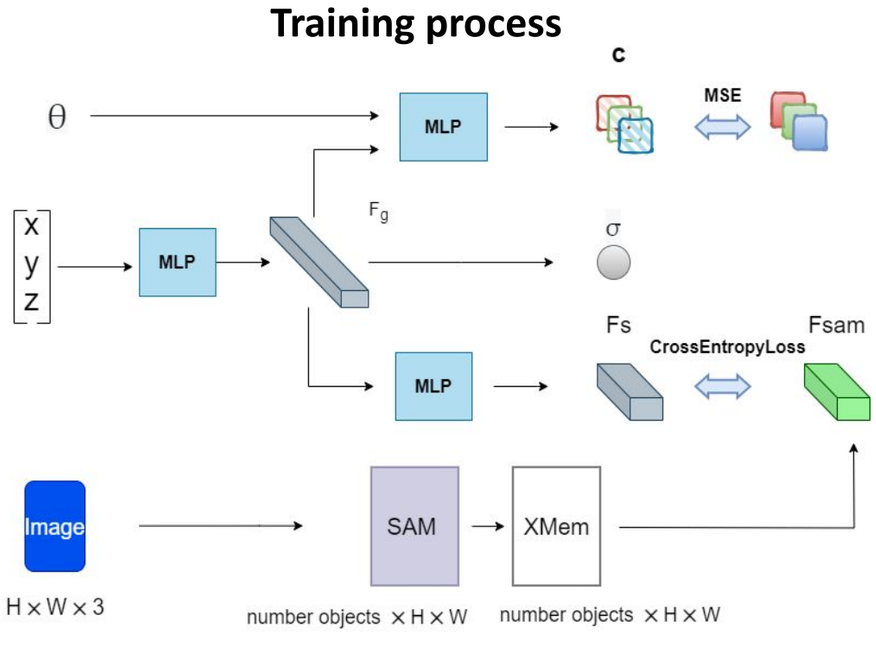
Motivation

To bring an object at user's request, a mobile robot has to find a queried item and determine its coordinates in a 3D space. The robot has to understand the geometry of an object to interact with it effectively.

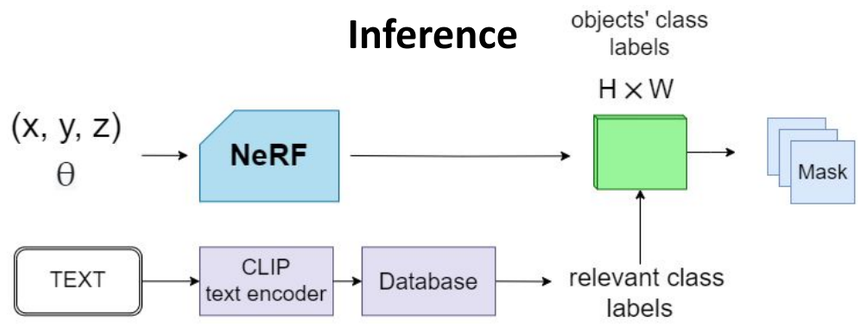
Research structure:

- **Preprocessing:** SAM model generates high-quality masks selected by an NMS-like algorithm. The VOS model (XMem) makes them consistent: receives object mask as input and tracks it on the video. To capture all occurrences on the video, XMem is sequentially restarted for each frame. The mask detected earlier is taken.
- **Image-language embeddings database:** For every class label it's consistent masks are multiplied by the corresponding image frames to eliminate the background. Then they are encoded using the CLIP model, which are averaged over the dataset to capture an object from multi-views.
- **Training:** Hash-NeRF predict labels of semantic class and RGB color for every pixel taking as input 3d coordinates and view direction.

Training process



Inference



Model	Use GPU for inference	Time inference	Memory
NeRF + SAM 2D	Yes	-	-
LERF	Yes	-	-
CLIP-Fields	No	*O*(number of points × CLIP dim)	*O*(number of points × CLIP dim)
OpenScene	No	*O*(number of points × CLIP dim)	*O*(number of points × CLIP dim)
Hash-NeRF (ours)	**No**	***O*(number of classes × CLIP dim)**	***O*(number of points + CLIP dim)***

Tab 1. **IoU metrics** of segmentation masks generated with text prompts on test set of LERF dataset.

text prompt	NeRF+SAM2D	LERF	Hash-NeRF
nerf gun	0,524	0,626	**0,862**
typewriter	0,439	0,640	**0,807**
white cabinet	**0,800**	0,265	0,389
yellow bulldozer	0,423	0,819	**0,878**
scene: `dozer_nerfgun_waldo`	0,546	0,588	**0,734**
apple	**0,930**	0,595	0,879
bear	0,778	0,480	**0.911**
mug	**0,871**	0,531	0,700
plate	**0,985**	0,688	0,934
scene: `teatime`	**0,891**	0,573	0.856
knives	0,599	0,238	**0,698**
refrigerator	0,683	0,170	**0,746**
sink	**0,910**	0,103	0,779
`mIoU scene: waldo_kitchen`	0,731	0,170	**0,741**

Tab 2. **Accuracy of object** localization by text prompts on test set of LERF dataset scenes. The object is localized successfully if its IoU metrics > 0,5 .

text prompt	NeRF+SAM2D	LERF	Hash-NeRF
nerf gun	0,541	0,514	**0,919**
typewriter	0,486	0,730	**0,865**
white cabinet	**0,838**	0,189	0,324
yellow bulldozer	0,459	0,892	**0,973**
scene: `dozer_nerfgun_waldo`	0,581	0,588	**0,770**
apple	**0,944**	0,611	**0,944**
bear	0,833	0,500	**0,944**
mug	**0,944**	0,889	**0,944**
plate	**0.999**	**0.999**	**0.999**
scene: `teatime`	0,930	0,750	**0,958**
knives	0,526	0,158	**0,789**
refrigerator	0,684	0,1	**0,789**
sink	**0,947**	0,105	0,737
`mAcc scene: waldo_kitchen`	0,719	0.121	**0,772**

Results:

- outperforms baselines on both segmentation quality and consistency on all scenes in average.
- addresses both general and specific language concepts with significant quality improvements.
- could overcome ambiguous queries with user guidance.

Fig. 1: Segmented point cloud produced by Hash-NeRF for "teatime" scene LeRF dataset.

Conclusion

Hash-NeRF could localize semantic information into robot memory effectively and interact with a user by text queries. It characterizes with: high-quality of segmentation masks, open-vocabulary, object, localization on occluded areas, superiority in time inference and memory-saving.

1 Skolkovo Institute of Science and Technology, Skolkovo, Moscow Region, Russia
Group contacts: julia.kudr.alex@gmail.com