VL-Fields: Towards Language-Grounded Neural Implicit Spatial Representations

Nikolaos Tsagkas, Oisin Mac Aodha, Chris Xiaoxuan Lu
{n.tsagkas, oisin.macaodha, xiaoxuan.lu}@ed.ac.uk

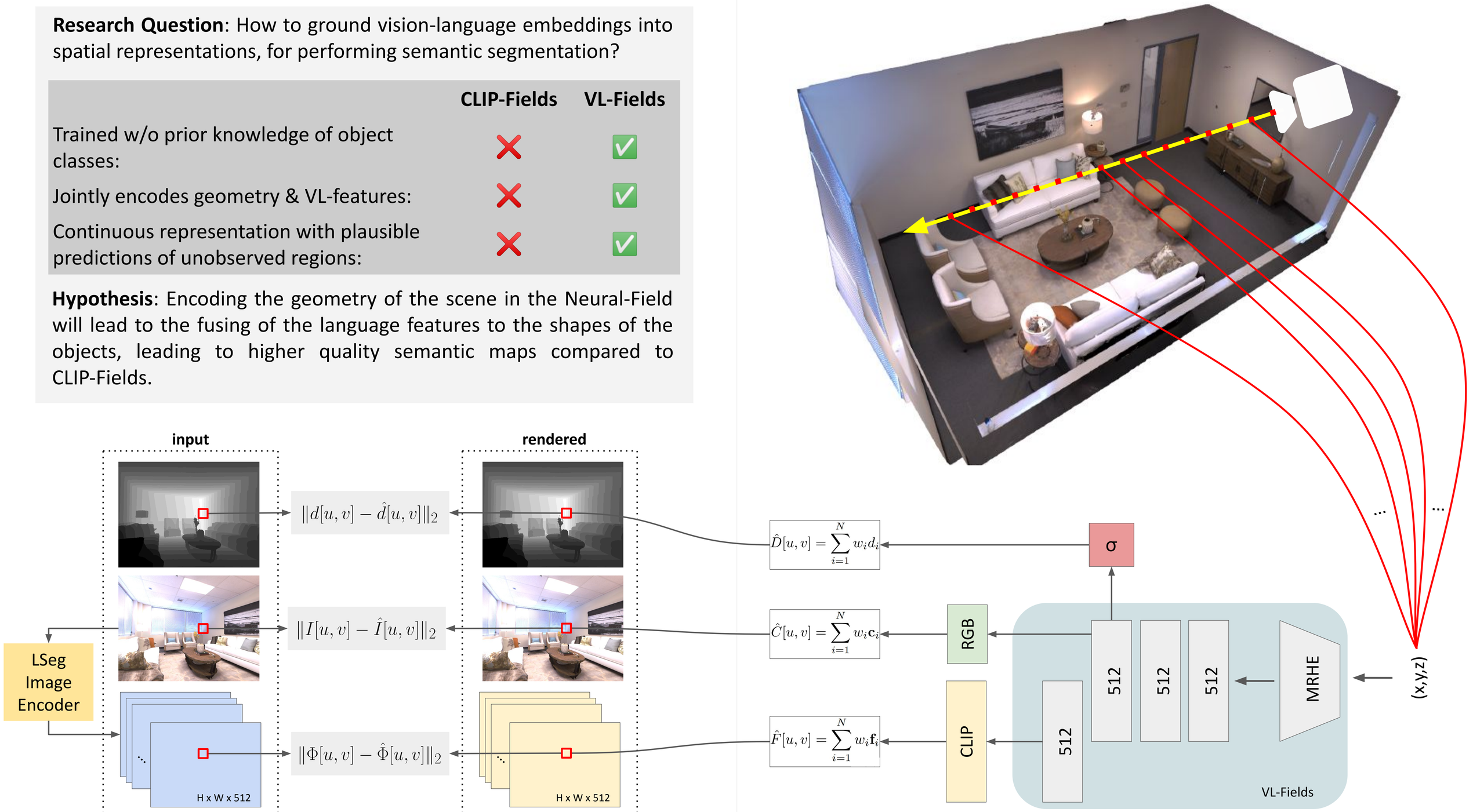
What is VL-Fields?

Research Question: How to ground vision-language embeddings into spatial representations, for performing semantic segmentation?

	CLIP-Fields	VL-Fields
Trained w/o prior knowledge of object classes:	❌	✅
Jointly encodes geometry & VL-features:	❌	✅
Continuous representation with plausible predictions of unobserved regions:	❌	✅

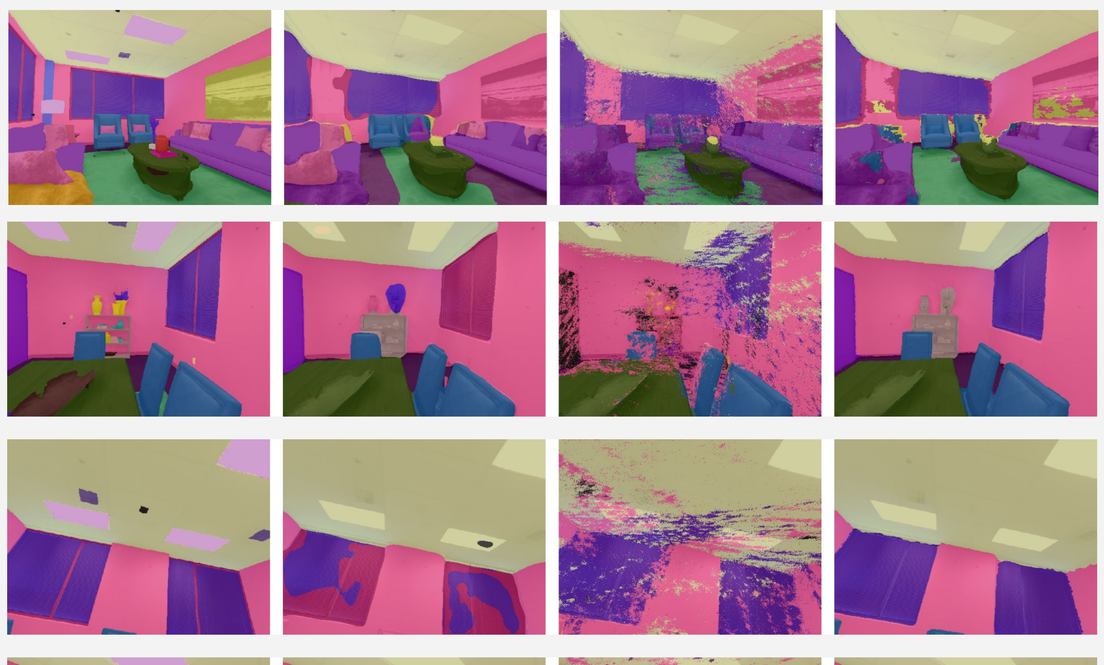
Hypothesis: Encoding the geometry of the scene in the Neural-Field will lead to the fusing of the language features to the shapes of the objects, leading to higher quality semantic maps compared to CLIP-Fields.

Training Pipeline



Qualitative & Quantitative Evaluation

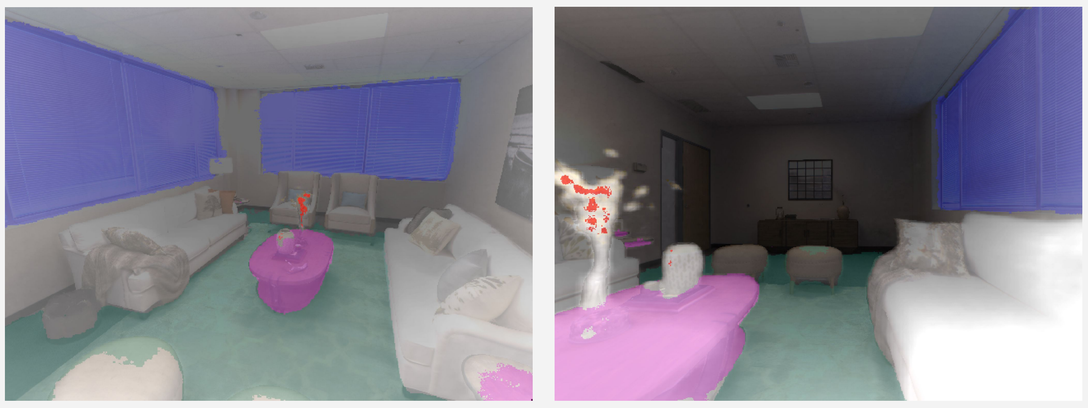
Semantic Segmentation



Ground Truth | LSeg | CLIP-Fields | VL-Fields (ours)

	mIoU				
	room_0	room_1	room_2	office_0	office_1
LSeg	0.603	0.643	0.771	0.755	0.759
CLIP-Fields	0.544	0.640	0.748	0.718	0.678
VL-Fields	**0.629**	**0.657**	**0.821**	**0.768**	**0.761**

Open-Vocabulary Queries



Open-vocabulary language-based queries in 3D space: “vacuum the rug”, “clean the table”, “pick up the plant”, “dust the blinds”. The colors indicate the areas in the encoded 3D space that correspond to each command.

Limitations

Smaller objects are fused semantically with larger object

LSeg loses CLIP’s ability to identify long-tail objects