3DRefTransformer: Fine-Grained Object Identification in Real-World Scenes Using Natural Language

Ahmed Abdelreheem, Ujjwal Upadhyay, Ivan Skorokhodov,
Rawan Al Yahya, Jun Chen, Mohamed Elhoseiny
King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
{ahmed.abdelreheem, ujjwal.upadhyay, ivan.skorokhodov, rawan.yahya, jun.chen, mohamed.elhoseiny}@kaust.edu.sa

This document include the following:

- Hyper-parameters used in our model
- Additional experiments
- Qualitative examples (successful/failure)

1. Hyper-parameters in 3D RefTransformer

We summarize the hyper-parameters used in our proposed model 3DRefTransformer in Table. 1. We use the same hyper-parameters for all transformer encoders in our model (object, language, and multimodal transformer encoders).

Hyper-parameter	Value
Base Learning Rate	0.0005
Batch Size	16
Embedding dim d	128
Obj./Lang./MM transformer layers L	2
Obj./Lang./MM transformer attention heads	8
Obj./Lang./MM transformer FFN dim	512
Obj./Lang./MM transformer FFN dropout	0.1
Obj./Lang./MM transformer attention dropout	0.3
Optimizer	Adam

Table 1. Hyper-parameters of 3DRefTransformer

2. Additional experiments

Changing scale value in the scaled cosine distance. Following suggestions for choosing the scale value in [2] and [1], we used a scale value of 3.3. Moreover, we attempted different scale values η, as discussed in Section. 3.5 of the main paper. In Table. 2, we report the overall performance of this experiment.

3. Additional qualitative examples

We present some qualitative examples of our proposed model 3DRefTransformer in Figure. 1 and Figure. 2.

References

[1] Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross B. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. *CoRR*, abs/1512.04143, 2015.

[2] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed M. Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. *CoRR*, abs/1804.10660, 2018.

	Accuracy
$\eta = 1.0$	$30.2\% \pm 0.4\%$
$\eta = 2.0$	$35.2\% \pm 0.3\%$
$\eta = \mathbf{3.3}$	$\mathbf{39.0\% \pm 0.3\%}$
$\eta = 4.0$	$38.3\% \pm 0.5\%$
$\eta = 5.0$	$38.0\% \pm 0.4\%$

Table 2. This table shows the performance of our model 3DRefTransformer trained using different scale values. Using scale value of 3.3 which is suggested by [2] and [1], gave the best performance.

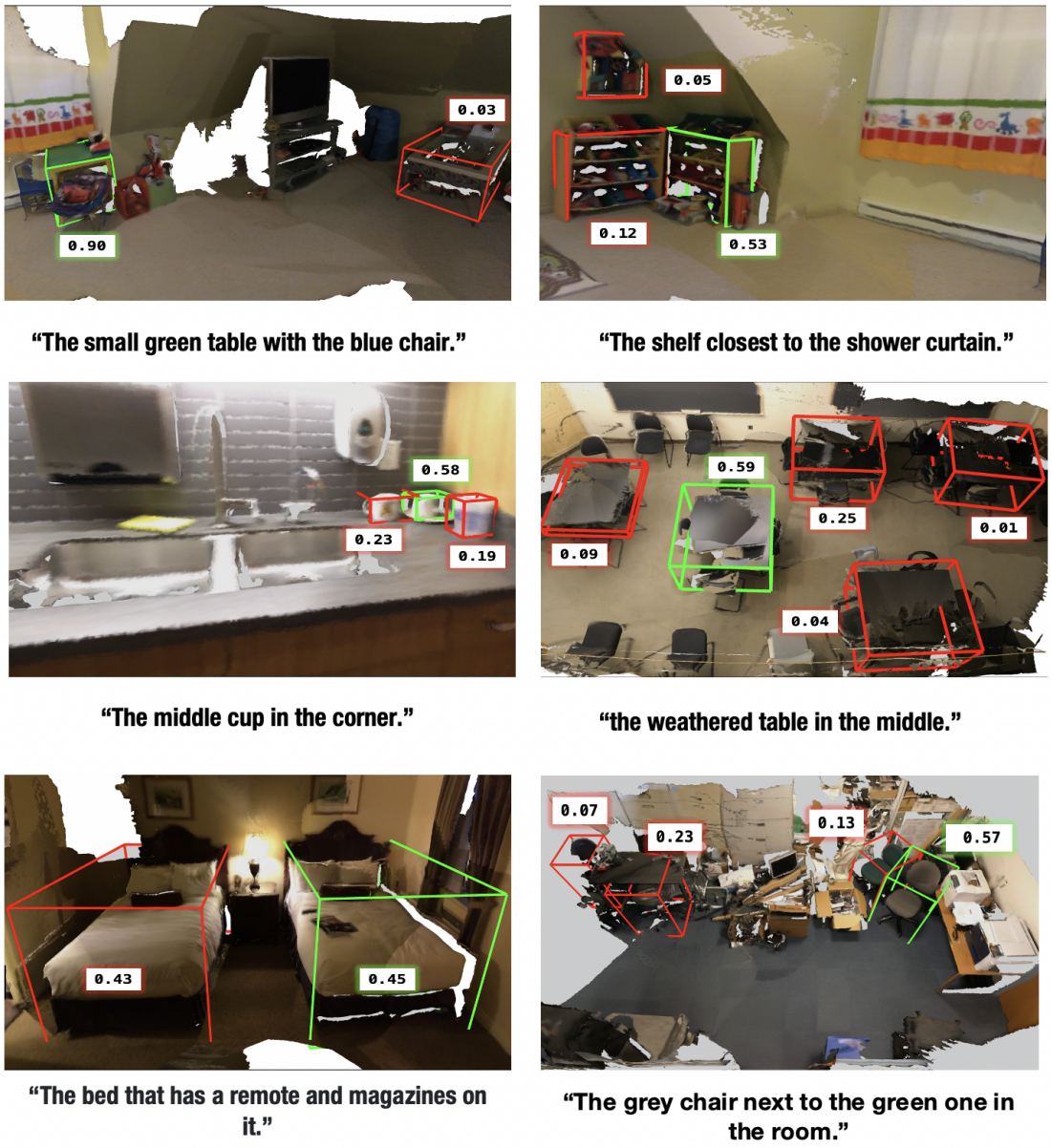
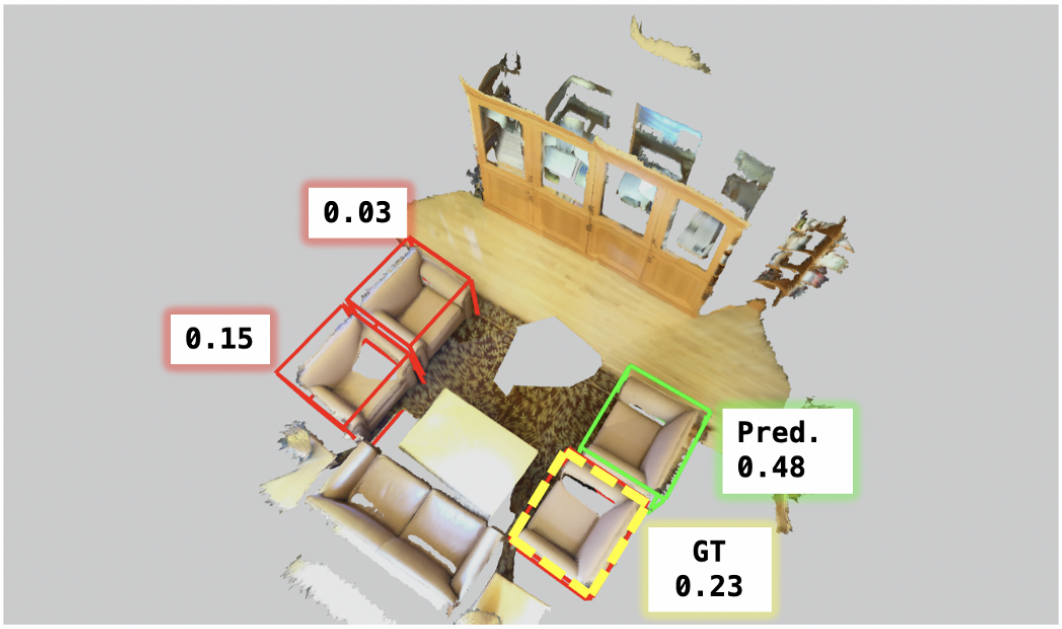
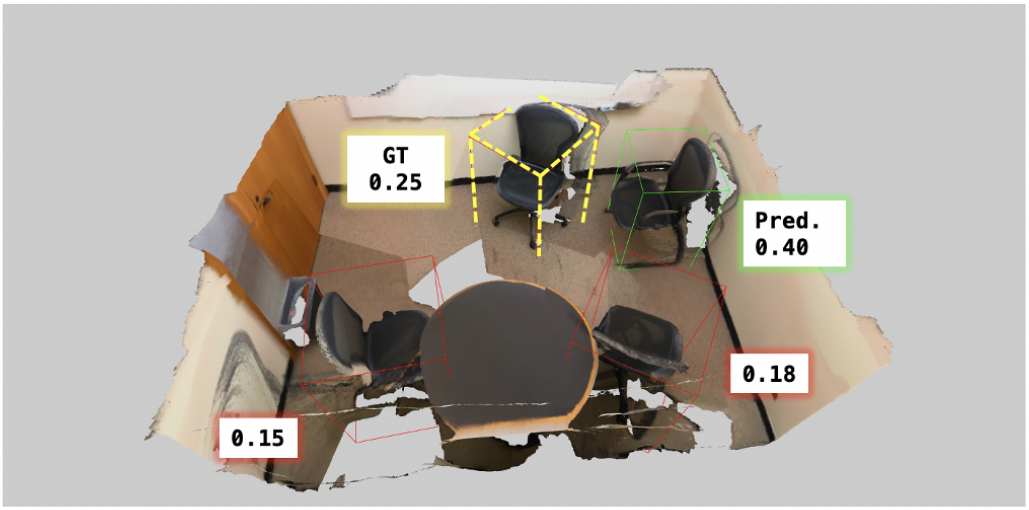


Figure 1. Successful examples using our most successful model.



"Walking towards the sofas, it's the left sofa furthest away from wooden cabinets."



"The chair closest to the whiteboard."

Figure 2. Failure examples using our best performing model.