JUNE 18-22, 2023
CVPR VANCOUVER, CANADA

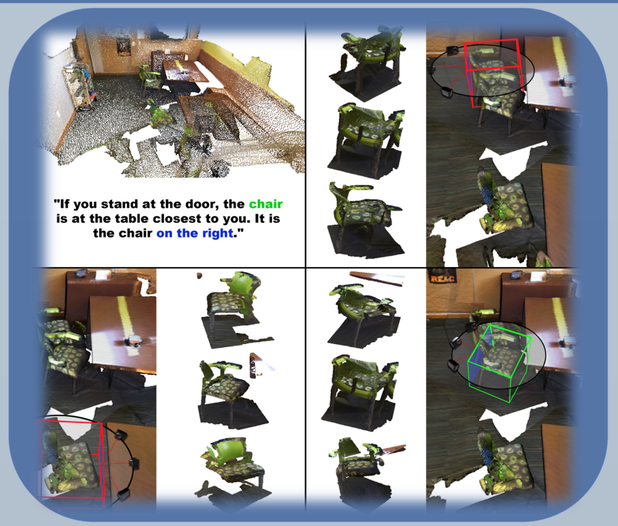
Look Around and Refer: 2D Synthetic Semantics Knowledge Distillation for 3D Visual Grounding

Eslam Mohamed Bakr, Yasmeen AlSaedy, Mohamed Elhoseiny

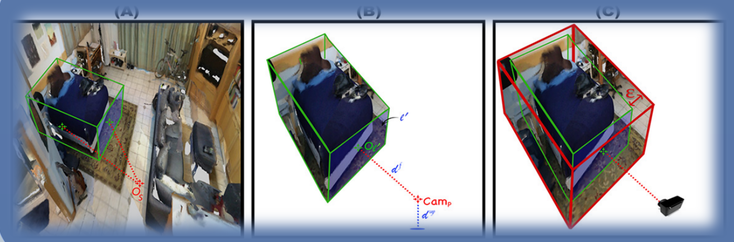
جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Introduction

The 3D visual grounding task has been explored with visual and language streams comprehending referential language to identify target objects in 3D scenes. However, most existing methods devote the visual stream to capturing the 3D visual clues using off-the-shelf point clouds encoders. The main question we address in this paper is “can we consolidate the 3D visual stream by 2D clues synthesized from point clouds and efficiently utilize them in training and testing?”. The main idea is to assist the 3D encoder by incorporating rich 2D object representations without requiring extra 2D inputs. To this end, we leverage 2D clues, synthetically generated from 3D point clouds, and empirically show their aptitude to boost the quality of the learned visual representations.

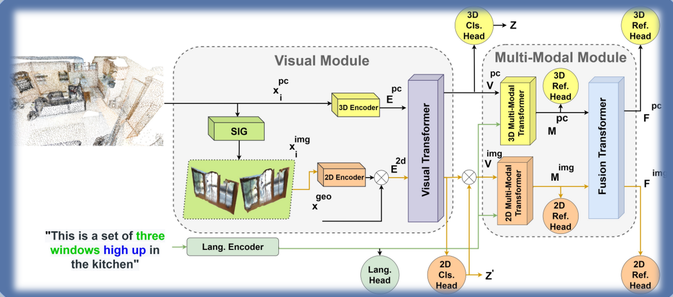


Look Around: 2D Synthetic Images Generator



Simplified overview of our 2D Synthetic Images Generator (SIG) module. First, we determine the prominent face of each object w.r.t the scene center. Then, the camera is located at a distance d f from that face and at a distance d up from the room’s floor. Finally, we randomly extend the region of interest by ϵ.

Refer: Visiolinguistic Transformer Architecture



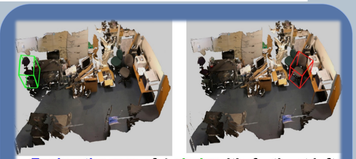
Results

LAR achieves state-of-the-art results across three different 3D grounding benchmarks, which do not rely on extra training data, by a large margin, i.e., 5.0%, 1.9% and 2.3% on Nr3D, Sr3D and ScanRefer, respectively. Hence, despite the unfair comparison with SAT, which uses extra 2D images, LAR surpasses SAT by 1.6%, 2.7%, and 2.3% on Nr3D, Sr3D, and ScanRefer, respectively.

Method	Additional Input	Nr3D				
		Overall(σ)	Easy	Hard	View-dep.	View-indep.
ReferIt3D [2]	-	35.6%	43.6%	27.9%	32.5%	37.1%
Text-Guided-GNNs [17]	-	37.3%	44.2%	30.6%	35.8%	38.0%
InstanceRefer [58]	S_{pc}	38.8%	46.0%	31.8%	34.5%	41.9%
3DRefTransformer [1]	-	39.0%	46.4%	32.0%	34.7%	41.2%
3DVG-Transformer [59]	-	40.8%	48.5%	34.8%	34.8%	43.7%
FFL-3DOG [12]	-	41.7%	48.2%	35.0%	37.1%	44.7%
TransRefer3D [13]	-	42.1%	48.5%	36.0%	36.5%	44.9%
LanguageRefer [37]	-	43.9%	51.0%	36.6%	41.7%	45.0%
Non-SAT [52]	-	37.7%	44.5%	31.2%	34.1%	39.5%
SAT [52]	$2D_{Img}$	49.2%	56.3	42.4	46.9	50.4
SAT † [52]	$2D_{Img}$	47.3%	55.8%	41.4%	46.9%	50.4%
LAR (**Ours**)	-	**48.9%**±0.2	**58.4%**	**42.3%**	**47.4%**	**52.1%**

Method	Additional Input	Sr3D				
		Overall(σ)	Easy	Hard	View-dep.	View-indep.
ReferIt3D [2]	-	40.8%	44.7%	31.5%	39.2%	40.8%
Text-Guided-GNNs [17]	-	45.0%	48.5%	36.9%	45.8%	45.0%
InstanceRefer [58]	S_{pc}	48.0%	51.1%	40.5%	45.4%	48.1%
3DRefTransformer [1]	-	47.0%	50.7%	38.3%	44.3%	47.1%
3DVG-Transformer [59]	-	51.4%	54.2%	44.9%	44.6%	51.7%
FFL-3DOG [12]	-	-	-	-	-	-
TransRefer3D [13]	-	57.4%	60.5%	50.2%	49.9%	57.7%
LanguageRefer [37]	-	56.0%	58.9%	49.3%	49.2%	56.3%
Non-SAT [52]	-	43.9%	-	-	-	-
SAT [52]	$2D_{Img}$	57.9%	61.2	50.0	49.2	58.3
SAT † [52]	$2D_{Img}$	56.6%	60.6%	49.7%	48.7%	57.4%
LAR (**Ours**)	-	**59.35%**±0.1	**63.0%**	**51.2%**	**50.0%**	**59.1%**


The towel hanging closest to the mirror


Facing the row of 4 chairs, it's furthest left