LOBSTAR: Language Model-based Obstruction Detection for Augmented Reality

Yanming Xiu, Tim Scargill, Maria Gorlatova

Intelligent Interactive Internet of Things (I^3T) Lab, Duke University

Introduction

- In augmented reality (**AR**), improperly placed content can obstruct real-world information[1]
- This produces the obstruction attacks

(a) (b) (c)

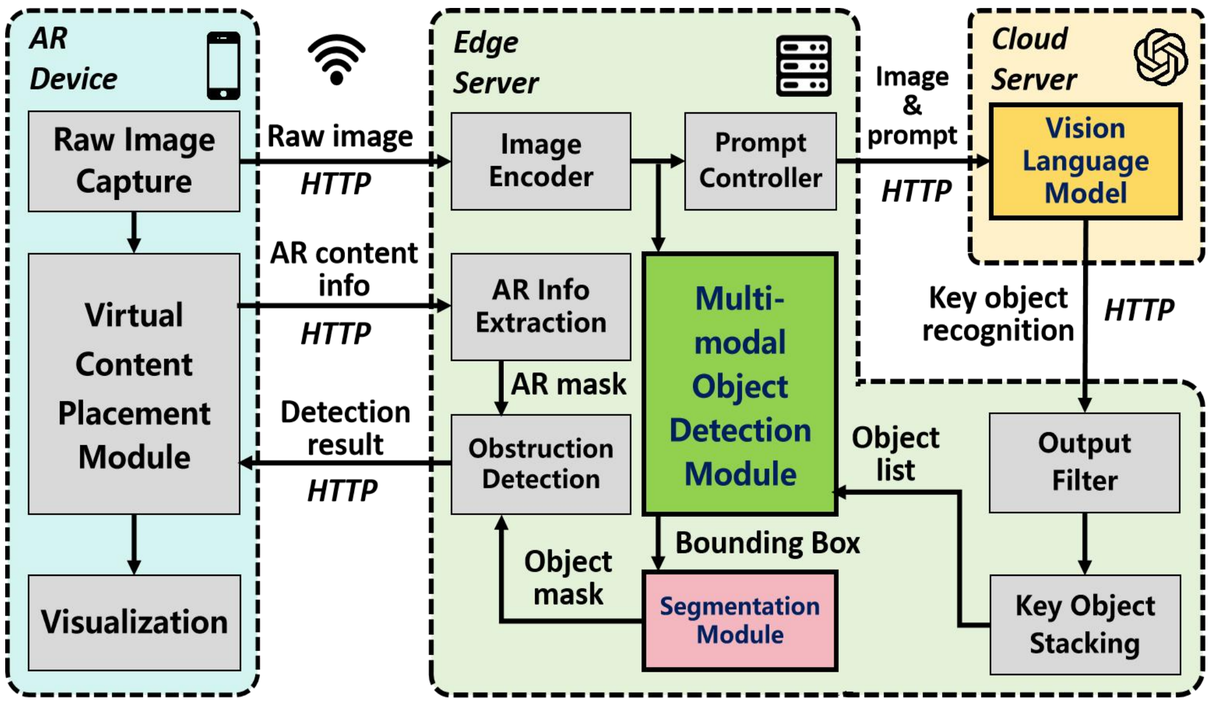
Example of an obstruction attack in AR: (a) real-world view, (b) AR view with a stop sign obstructed by a virtual arrow and (c) make the arrow translucent to mitigate obstruction

Motivation

- Traditional image quality assessment methods often lack a comprehensive understanding of the environmental information[3], cannot accurately detect key object obstruction in the image
- Vision language models (VLMs) have pushed the recent adavncement in scene understanding
- VLMs has a more holistic understanding of complex scenes and provides a potential improvement over classic methods in detecting obstruction attacks

LOBSTAR Design

LOBSTAR is deployed across three devices: (1) an AR device, (2) an edge server, and (3) a cloud server.



System architecture of the LOBSTAR system

- Vision language model: OpenAI GPT-4o
- Object detection model: GroundingDINO
- Segmentation model: Segment Anything VIT-B

System Evaluation

- Created two datasets: virtual scene and a real-world, each has raw and augmented image pairs, labeled with key object and obstruction status
- Virtual: 15 classes of key object, 280 image pairs; Real: 17 classes of key object, 144 image pairs

(a) (b) (c) (d) (e) (f) (g) (h)

Dataset samples: (a-d) real environment: original image, AR view, object mask, virtual content mask; (e-h) Virtual scene

- Result on dataset: > 96% detection accuracy

Dataset	Experiment	Obstruction Detection Accuracy	Segmentation Mean IoU	Detection Mean Confidence
Real	**LOBSTAR**	**96.53%**	**84.65%**	**75.35%**
	Blind	91.67%	76.78%	71.43%
	Prior	94.44%	87.14%	69.83%
	Saliency	56.25%	-	-
Virtual	**LOBSTAR**	**96.42%**	**85.99%**	**77.76%**
	Blind	92.14%	78.44%	74.57%
	Prior	97.14%	87.56%	73.73%
	Saliency	62.00%	-	-

Blind: very limited instruction in text prompt for VLM; Prior: directly tell object detection model the key object, with no VLM involved; Saliency: a saliency map-based baseline.

- Real-time evaluation: AR device: Google Pixel 7 Pro; edge server: ubuntu with NVIDIA GeForce RTX 3090; cloud server access: OpenAI API
- 580ms latency with a 5GHz WiFi connection, 950ms latency in a remote connection scenario

Conlusion and Future Work

- Proposed LOBSTAR, the first system using a VLM to detect obstruction attacks in AR
- Achieved high deetctionaccuracy and low latency
- Will apply LOBSTAR to head-mounted devices

Related Publications / Citations

[1] K. Cheng, F. Roesner, et al. Exploring user reactions and mental models towards perceptual manipulation attacks in mixed reality. Proceedings of USENIX Security, 2023.
[2] S. Davari, F. Lu, and D. A. Bowman. Occlusion management techniques for everyday glanceable AR interfaces. In IEEE VR Abstracts and Workshops, 2020.
[3] H. Duan, X. Min, Y. Zhu, G. Zhai, X. Yang, and P. Le Callet. Confusing image quality assessment: Toward better augmented reality experience. IEEE Transactions on Image Processing, 31:7206–7221, 2022.

Acknowledgements

DoD COE-ACS, NSF, IBM, Meta, DARPA, YOUNG FACULTY AWARD, ATHENA

Duke UNIVERSITY