π-Jack: Physical-World Adversarial Attack on Monocular Depth Estimation with Perspective Hijacking

Tianyue Zheng[1], Jingzhi Hu[2], Rui Tan[2], Yinqing Zhang[1], Ying He[2], and Jun Luo[2]

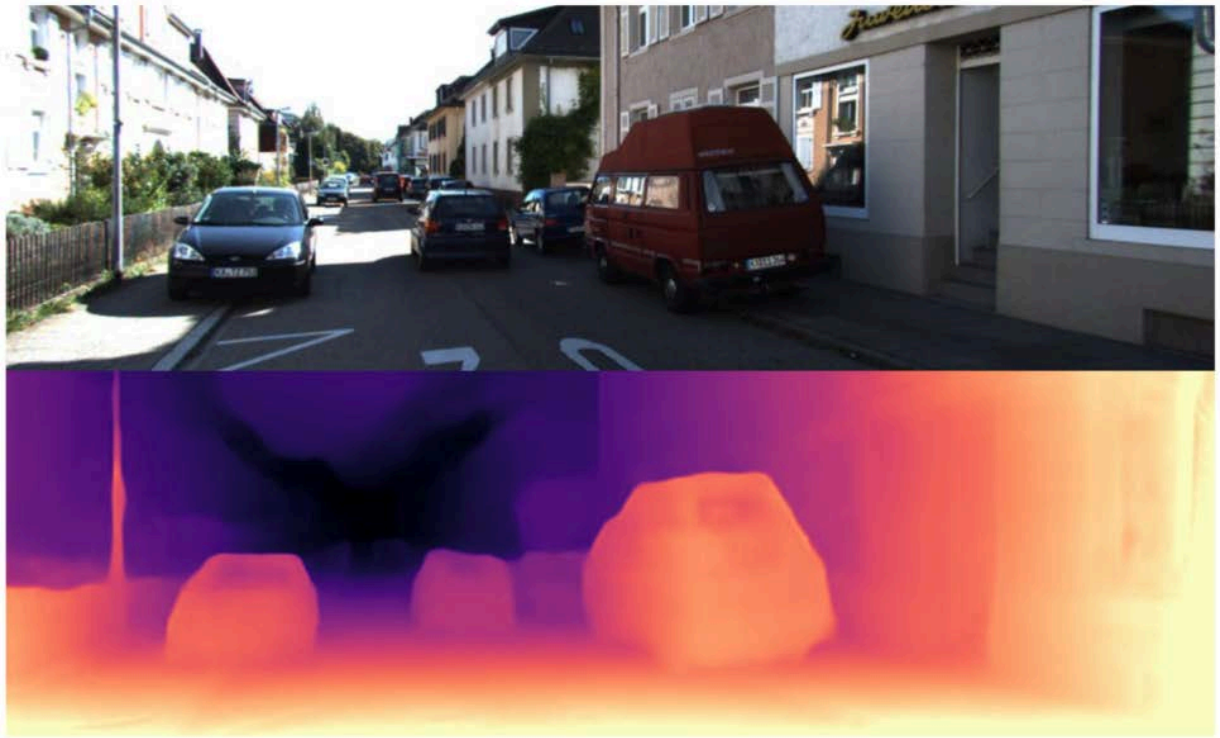
[1] School of Computer Science and Engineering, SUSTech, China

[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore

August, 2024

Motivation and Background

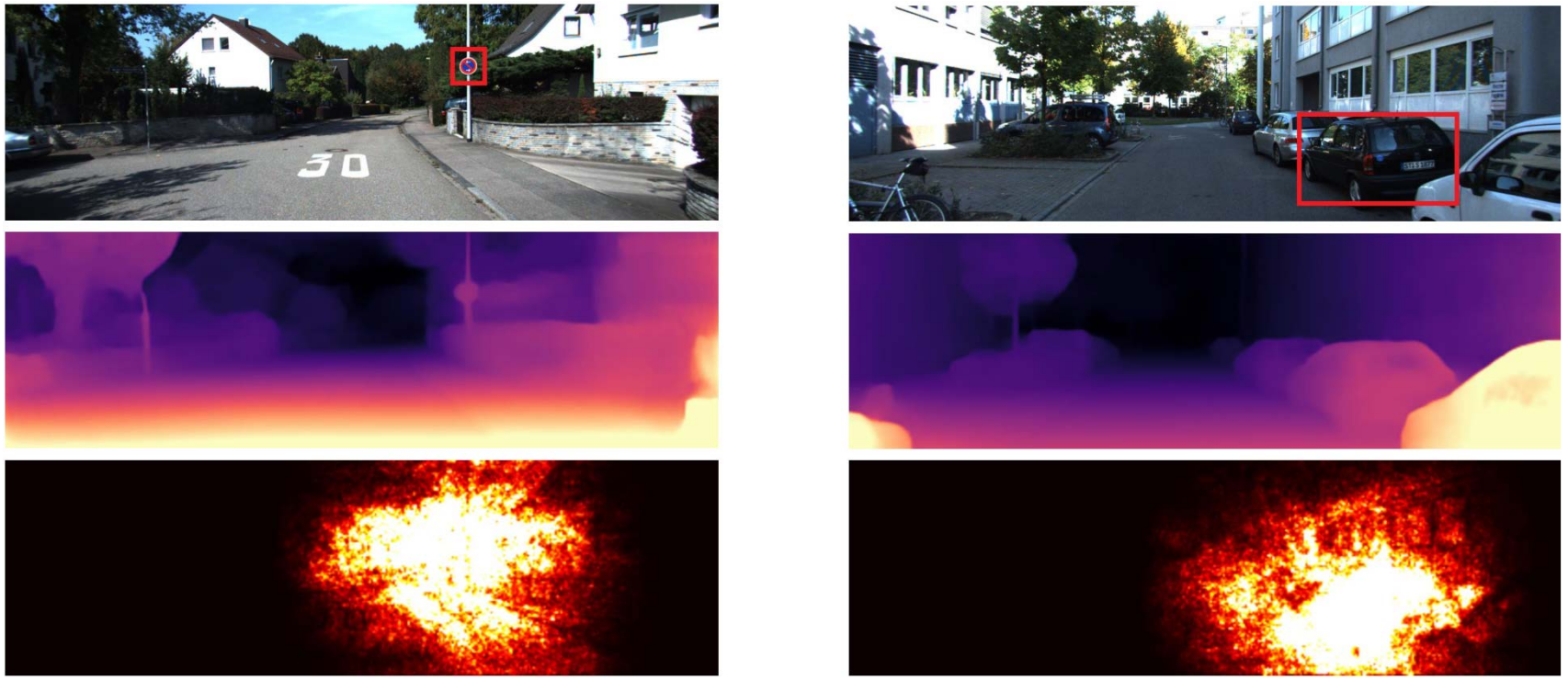
- Monocular depth estimation (MDE) estimates pixel-wise distances from a single RGB image.
- MDE is adopted by both academia and industry (e.g., Tesla, Waymo, and Toyota).
- Deficiencies in MDE could lead to AVs generating low-quality models of their surroundings.



SUSTech

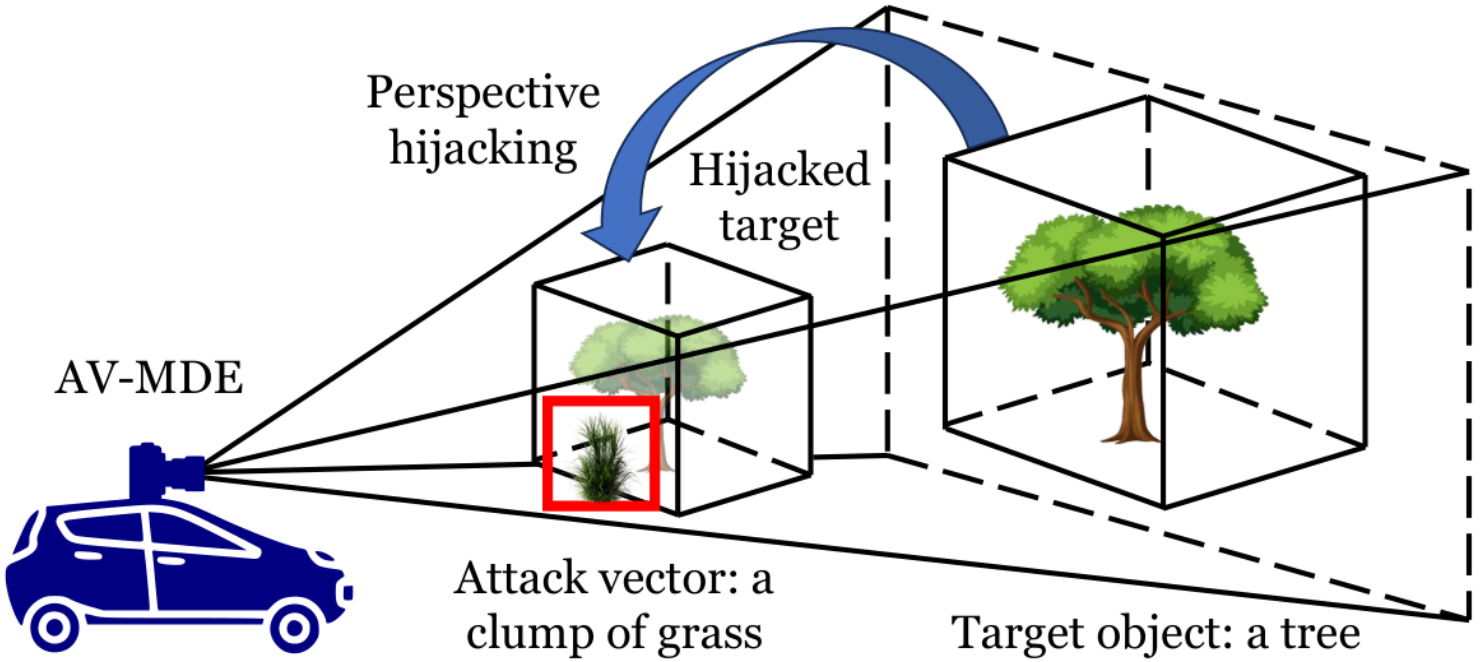
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Motivation and Background



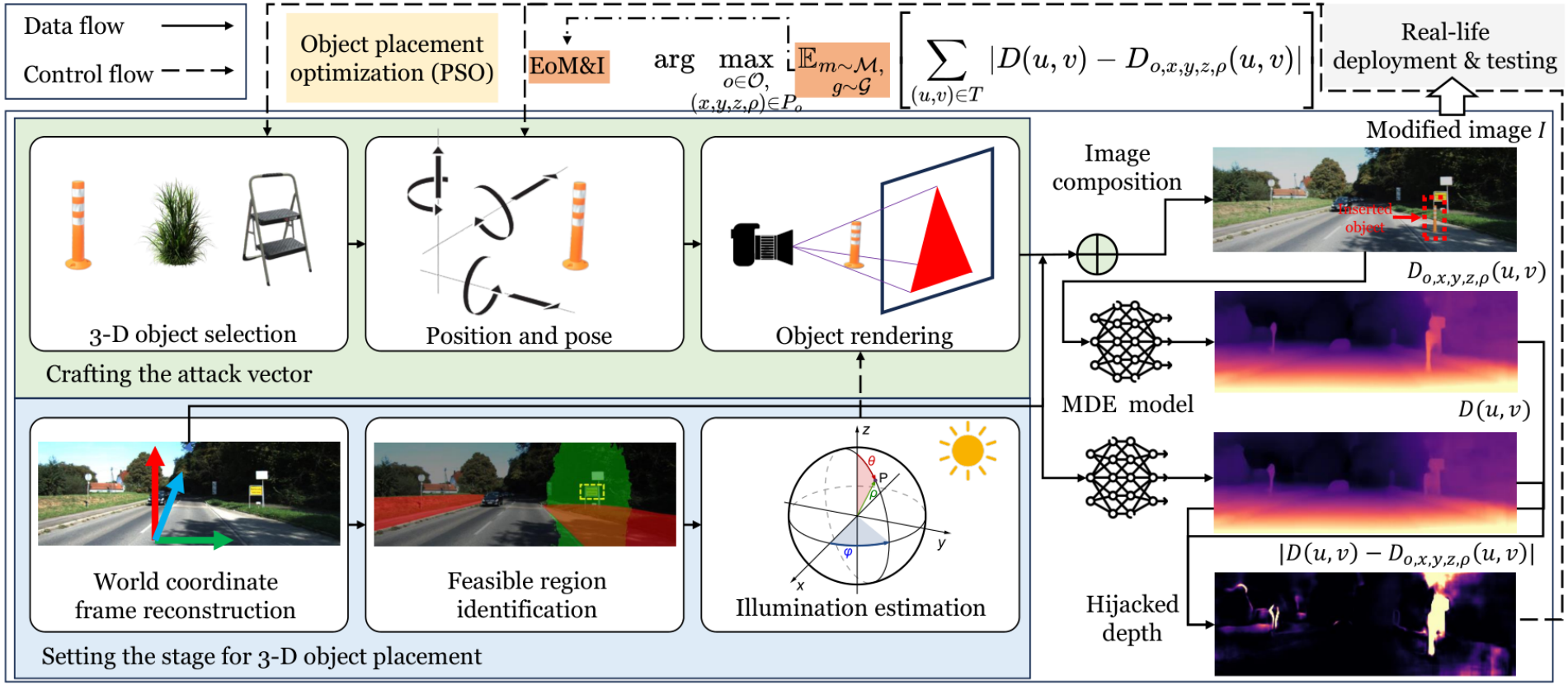
We then use the saliency map to understand what the depth estimation model focuses on when performing MDE, it turns out that MDE rely on perspective cues for depth inference.

Overview of π-Jack



High-level idea of π-Jack: by strategically placing an attack vector (a clump of grass), MDE for a target object (a tree) can be hijacked.

Workflow of π–Jack's Attack Strategy



- 3-D object selection
- Setting the stage for 3-D object placement
- Object placement and rendering
- Robust design and analysis

3-D Object Selection

- the 3-D object should possess structures similar to the target
- the 3-D object should exhibit a texture akin to the target
- the 3-D object should have an extended shape
- the object should be ordinary and inconspicuous

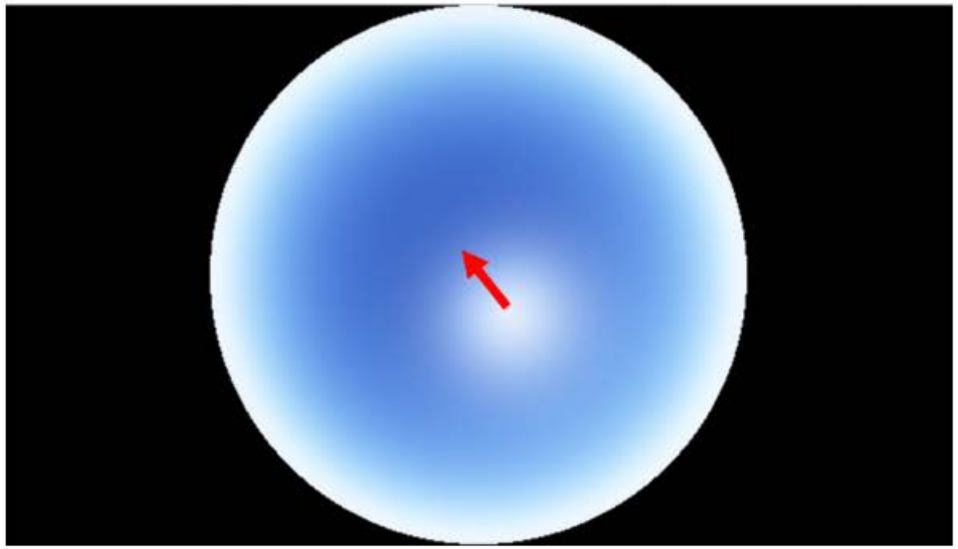
	Barrier pole	Flag	Grass clump	Ladder	Safety sign	Garbage bin	Traffic sign	Roadblock	Hydrant post
Structural similarity	Tree trunk, window	Tree trunk, window	Tree, bush	Scaffolding, fire escape	Vehicle, tree	Vehicle, building	Lamp post signs	Vehicle	Tree trunk, lamp post
Texture	Metallic	Glossy	Leafy	Wooden	Plastic	Glossy	Glossy	Coarse	Metallic
Extensibility	Good	Good	Fair	Good	Poor	Fair	Good	Fair	Fair
Typical height×width	$0.20\,m^2$	$0.92\,m^2$	$0.43\,m^2$	$1.20\,m^2$	$0.56\,m^2$	$1.4\,m^2$	$0.52\,m^2$	$0.73\,m^2$	$0.142\,m^2$
Stealthiness	Good	Fair	Good	Fair	Fair	Good	Fair	Fair	Fair

Properties of the selected 3-D objects.

Setting the stage for 3-D object placement

$$\mathbf{l}_s^*, \omega^*, \tau^* = \arg\min_{\mathbf{l}_s,\omega,\tau} \sum_q \left(Q(q) - \omega g_{\mathrm{RGB}}\left(\mathbf{l}_q, \tau, \mathbf{l}_s\right)\right)^2$$

$$L_z^* = \arg\min_{L_z} \sum_q \left(Q(q) - \omega g_{\mathrm{RGB}}\left(\mathbf{l}_q, L_z\right)\right)^2$$



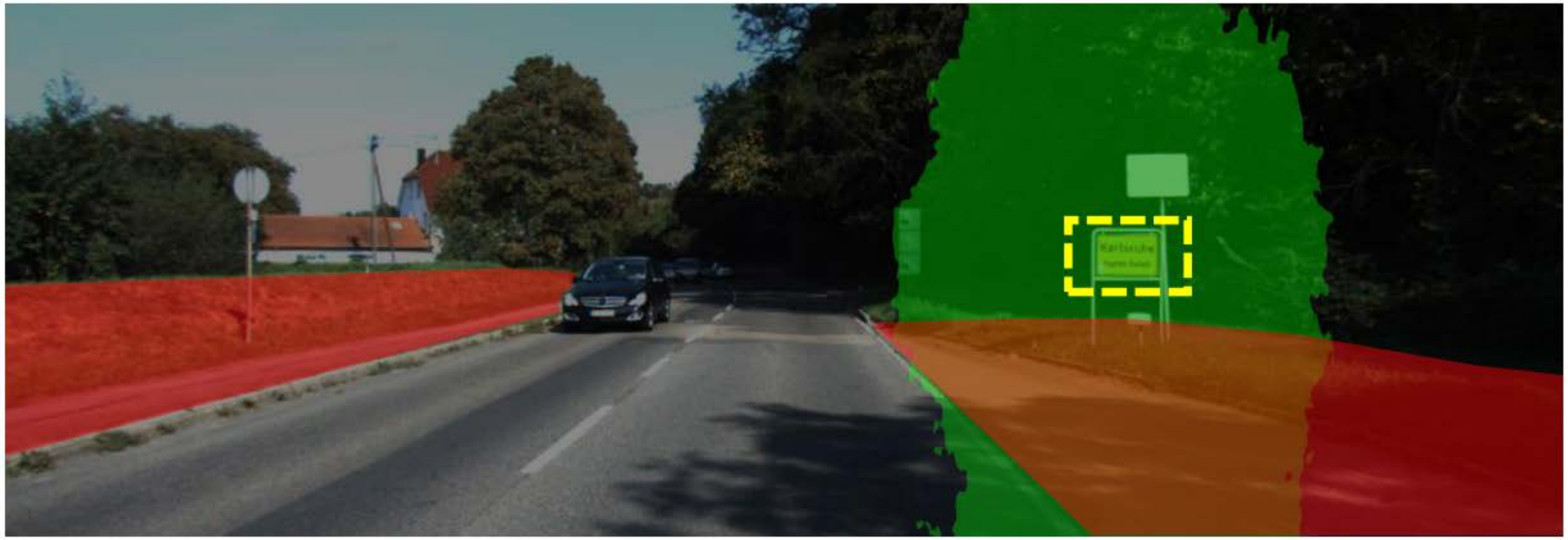
Illumination estimation

Setting the stage for 3-D object placement

$$S' = \sum_{(u_t, v_t) \in T} \left| \theta(I)(u_t, v_t) - \theta(\chi(I \odot B))(u_t, v_t) \right|$$

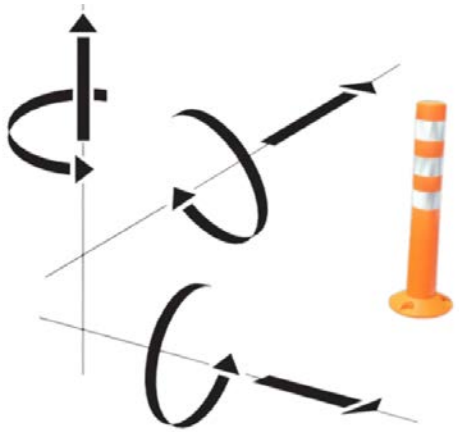
$$\mathcal{F}_{\text{sal}}(u, v) = \mathbb{I}\left[(G * S')(u, v) > \mathcal{T} \right]$$

$$\mathcal{F}(u, v) = \mathcal{F}_{\text{sal}}(u, v) \cap \mathcal{F}_{\text{val}}(u, v)$$

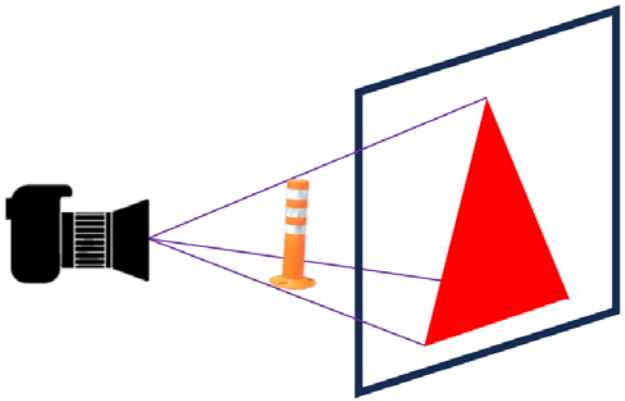


Feasible region identification

Object Placement and Rendering



Placement



Rendering

$$o^*, x^*, y^*, z^*, \rho^* = \arg\max_{o \in O, (x,y,z,\rho) \in P_o} \sum_{(u,v) \in T} \left| D(u,v) - D_{o,x,y,z,\rho}(u,v) \right|$$

Robust design

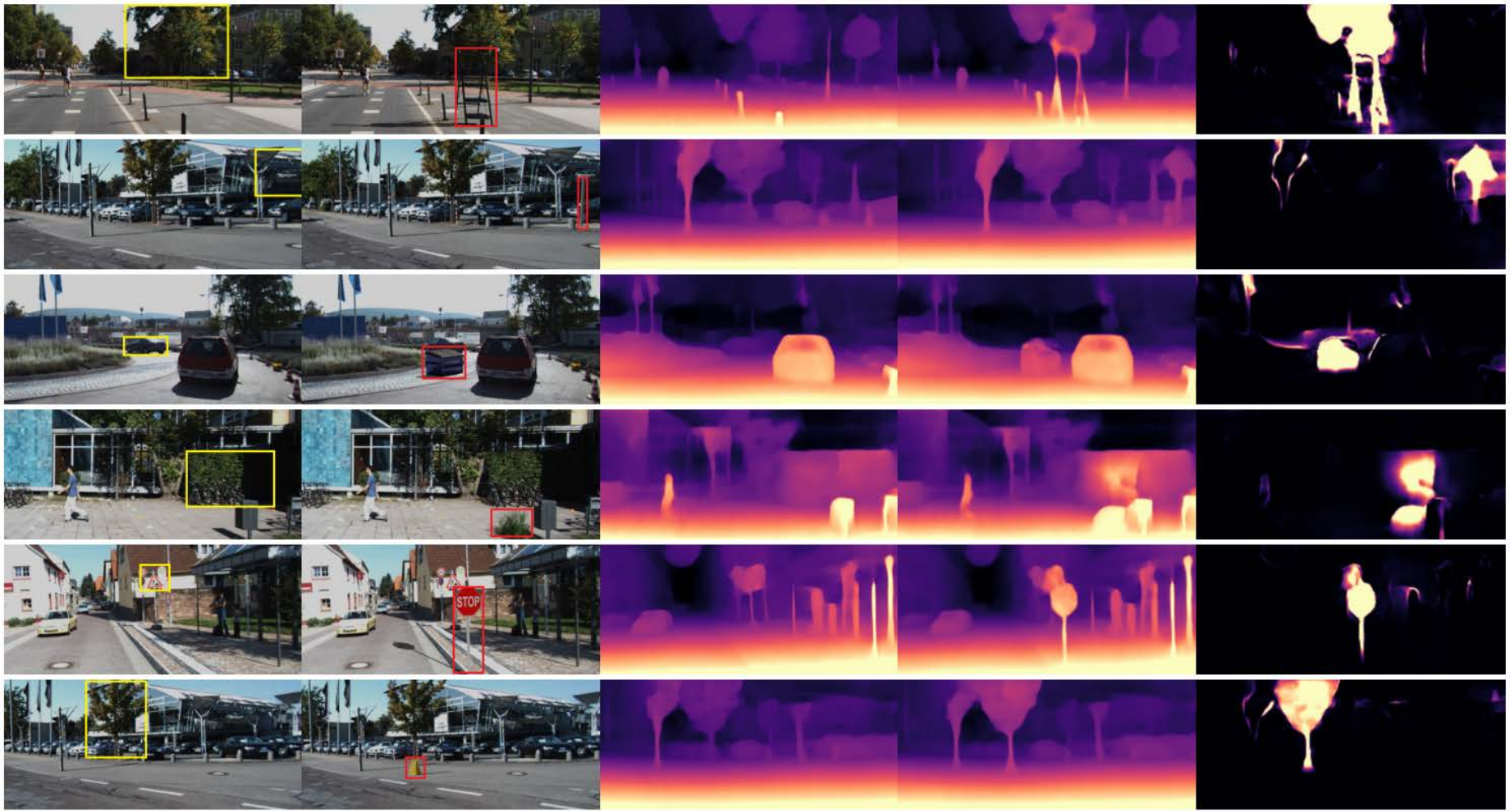
EoM&I

$$\arg\max_{\substack{o \in O, \\ (x,y,z,\rho) \in P_o}} \mathbb{E}_{\substack{m \sim \mathcal{M}, \\ g \sim \mathcal{G}}} \left[\sum_{(u,v) \in T} \left| D(u,v) - D_{o,x,y,z,\rho}(u,v) \right| \right]$$

SUSTech

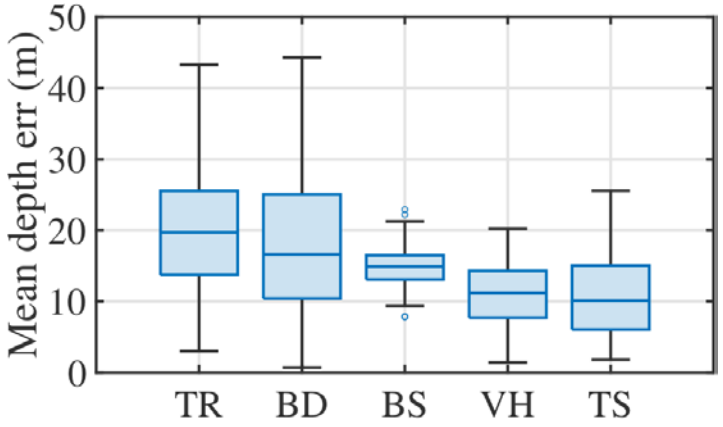
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Evaluation

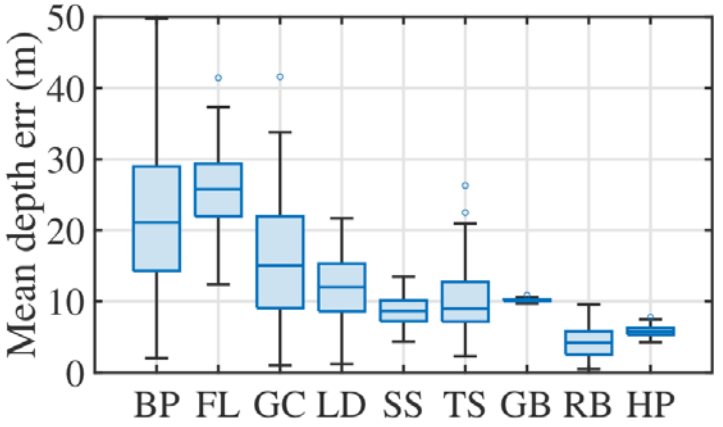


Example estimated depth maps before and after π-Jack attack

Evaluation

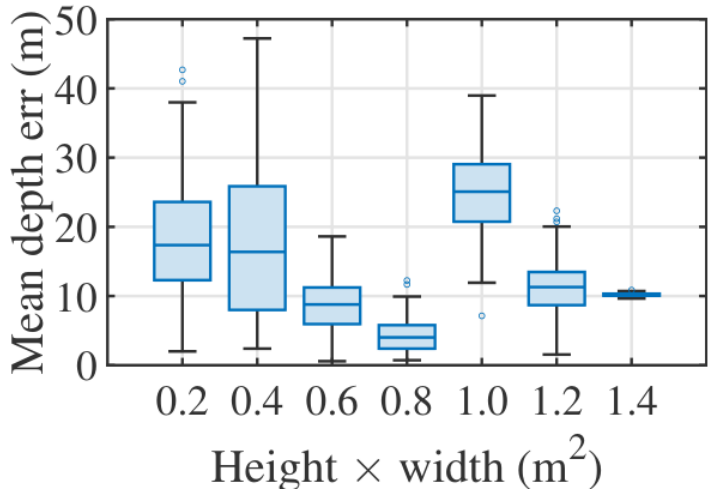


Target object

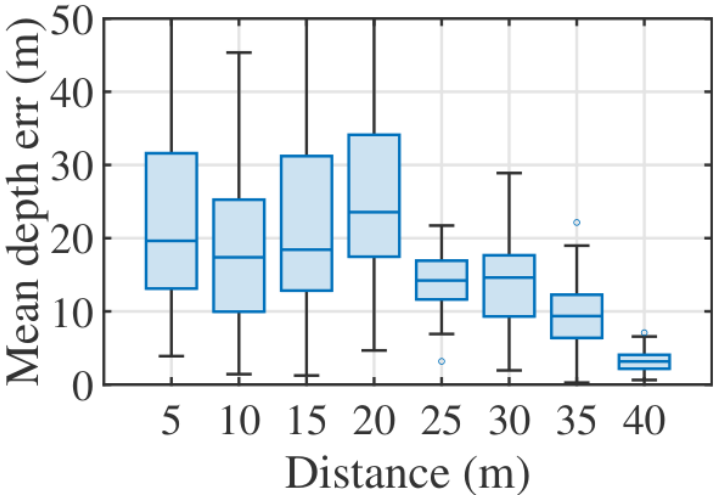


Attack vector

Overall π-Jack performance.



Impact of attack vector size

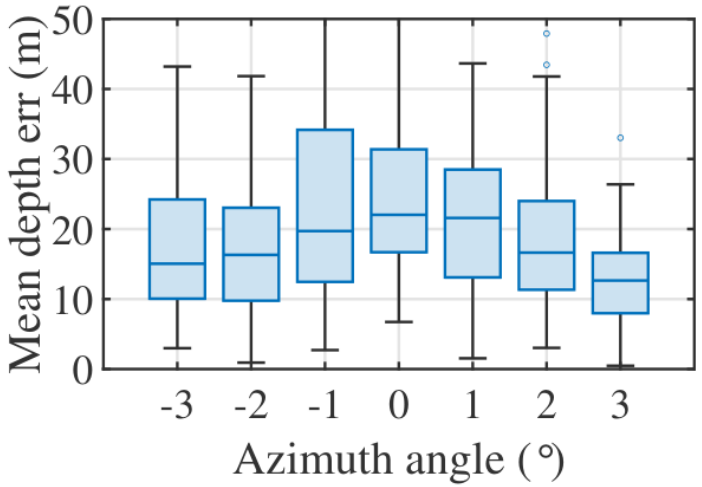


Impact of distance

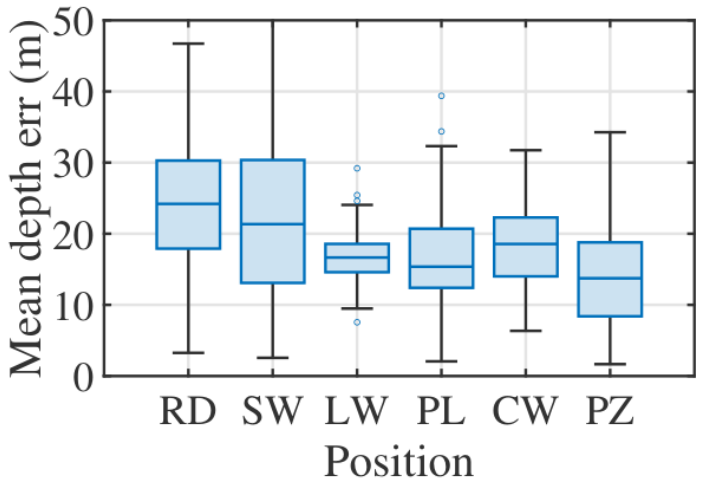
SUSTech

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

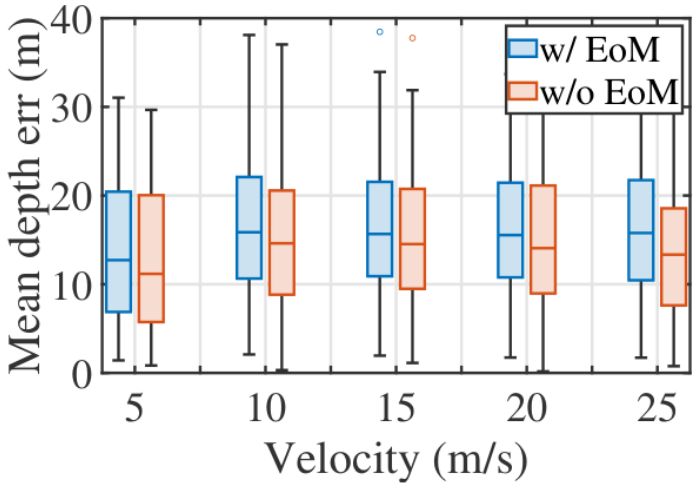
Evaluation



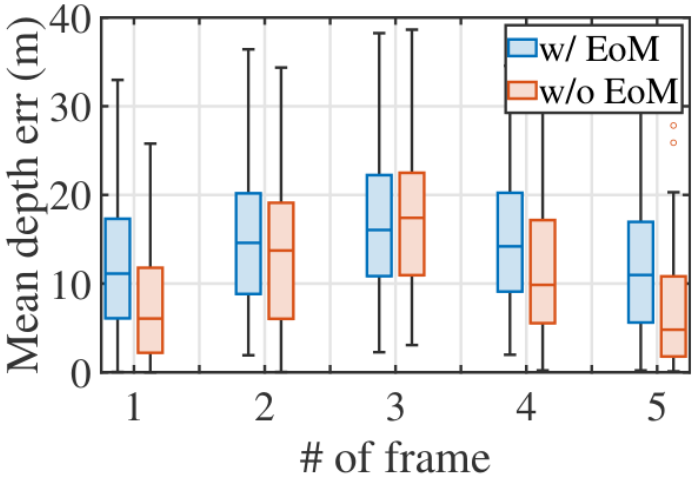
Impact of angle



Impact of position



Impact of velocity

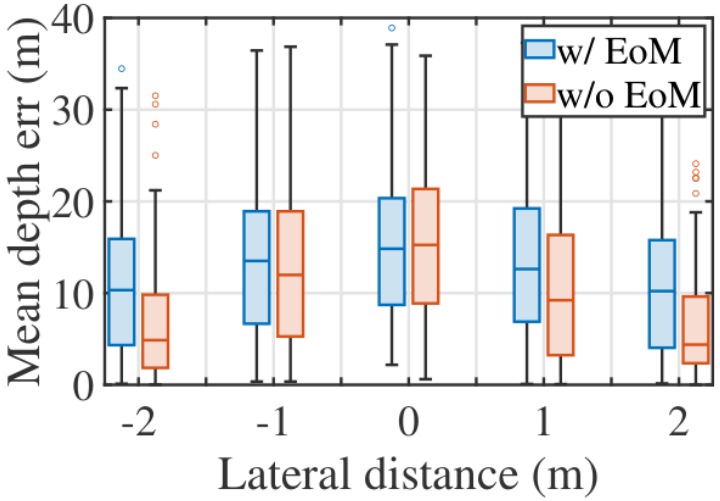


Impact of number of frames

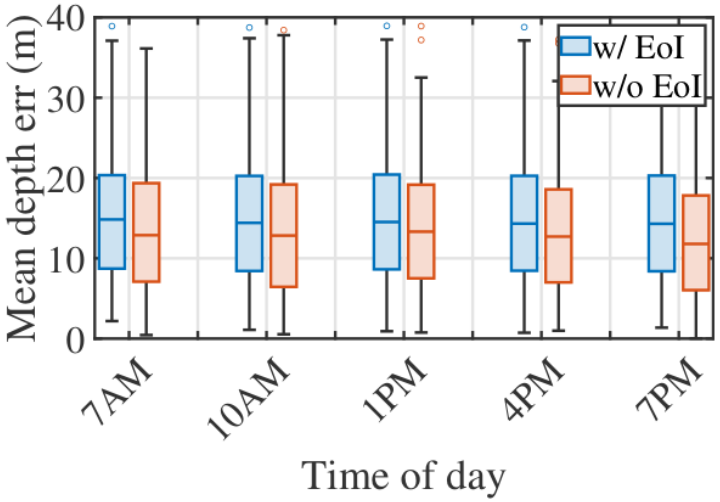
SUSTech

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

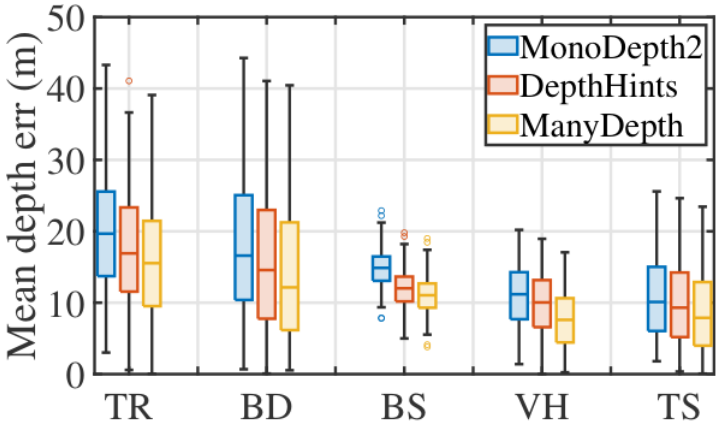
Evaluation



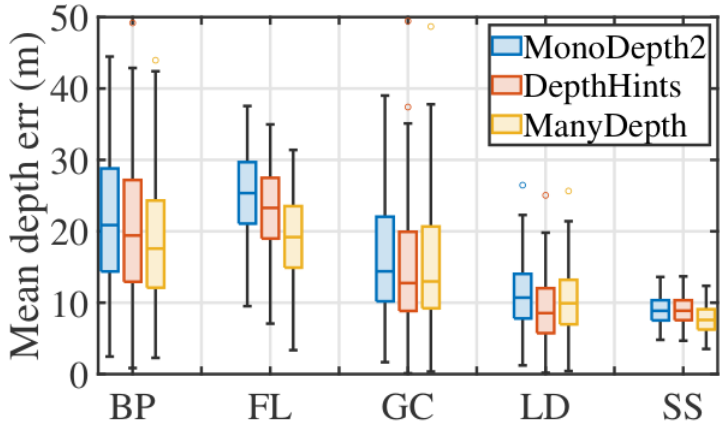
Impact of lateral distance



Impact of time of day



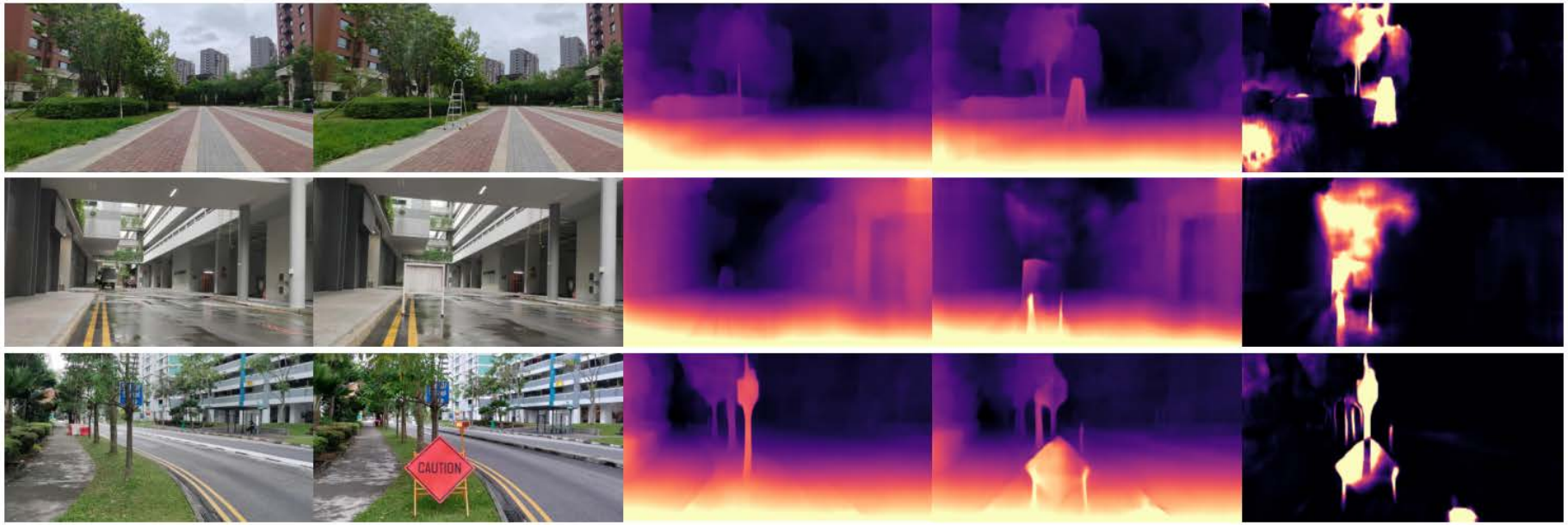
Target object



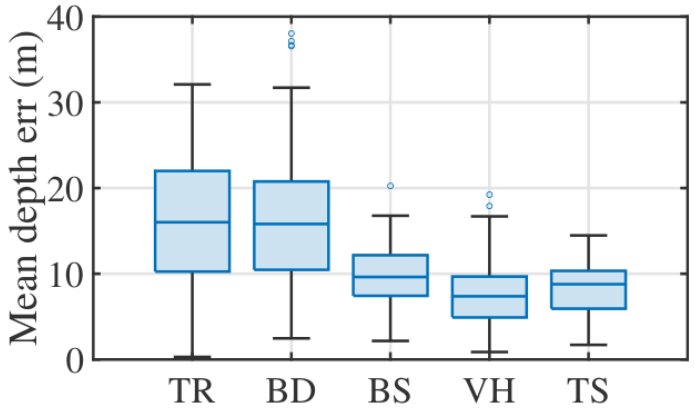
Attack vector

Generalizability of π-Jack to different models

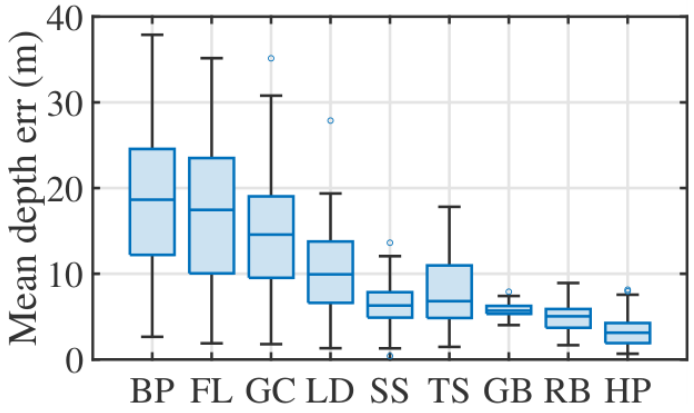
Evaluation



Evaluation with real-world scenes and attack vectors



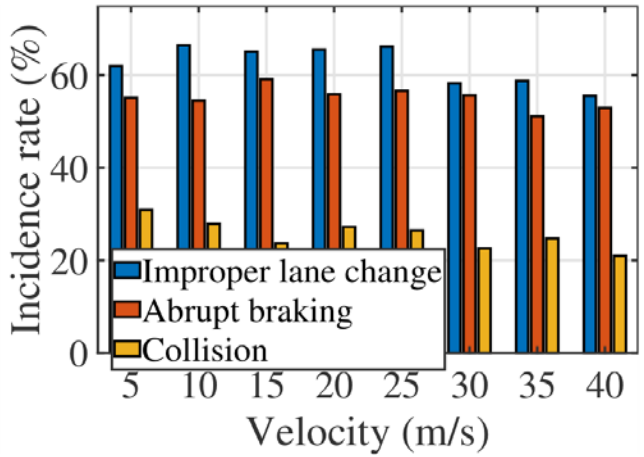
Target object



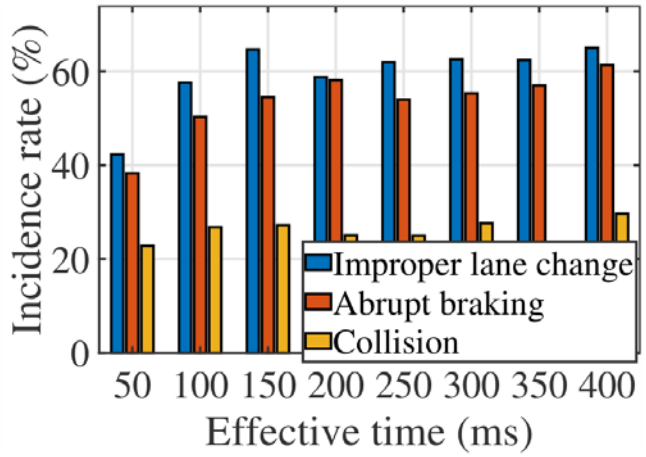
Attack vector

Summary of real-world evaluations

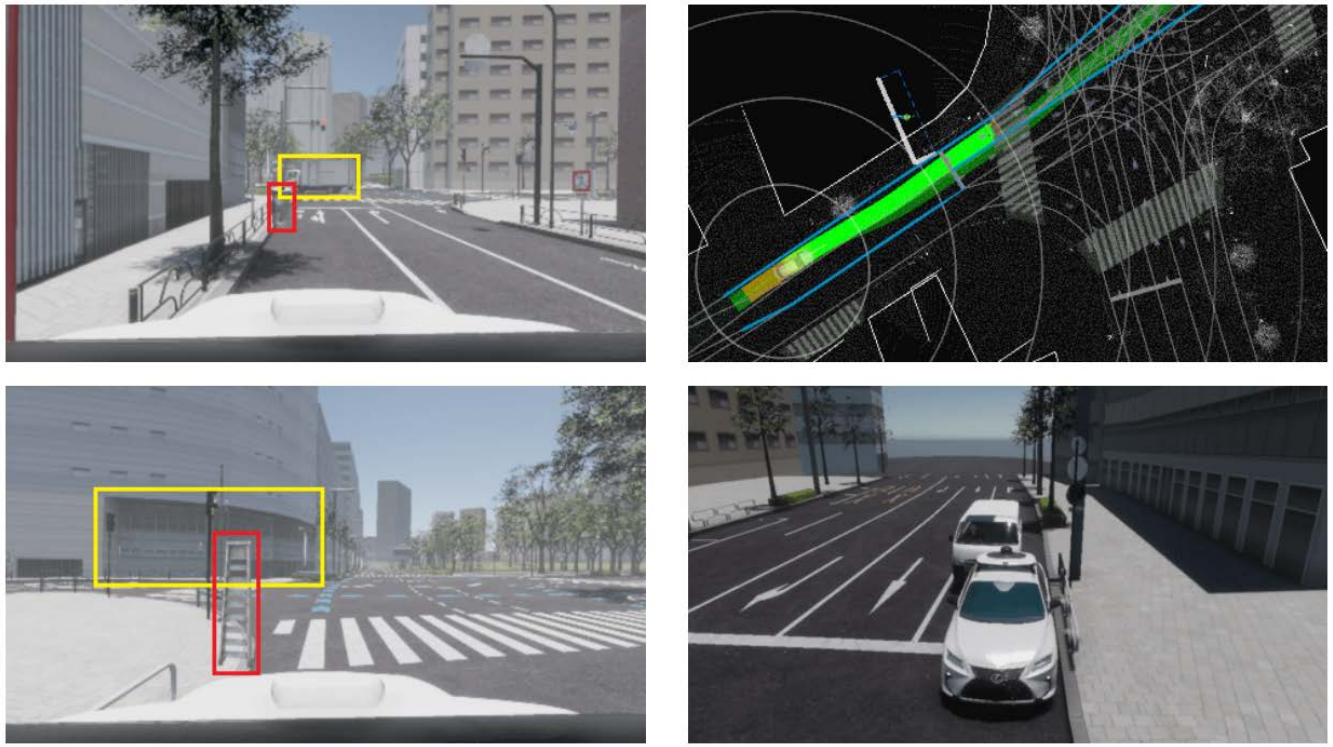
Evaluation



Impact of velocity



Impact of time



Example scenes from AWSIM and Autoware.

Conclusion and Future Work

- The first physical adversarial attack on AV-MDE systems utilizing perspective hijacking
- Exploiting ordinary 3-D objects as attack vectors, π-Jack offers superior effectiveness, robustness, accessibility, and inconspicuity.
- Experiments validate the high attack success rate and large depth difference achieved by π-Jack, demonstrating its successful application in both composited and real-world AV scenes.

Thank you!