Security of Multi-Sensor Fusion based Perception in Autonomous Driving

Under Physical-World Attacks

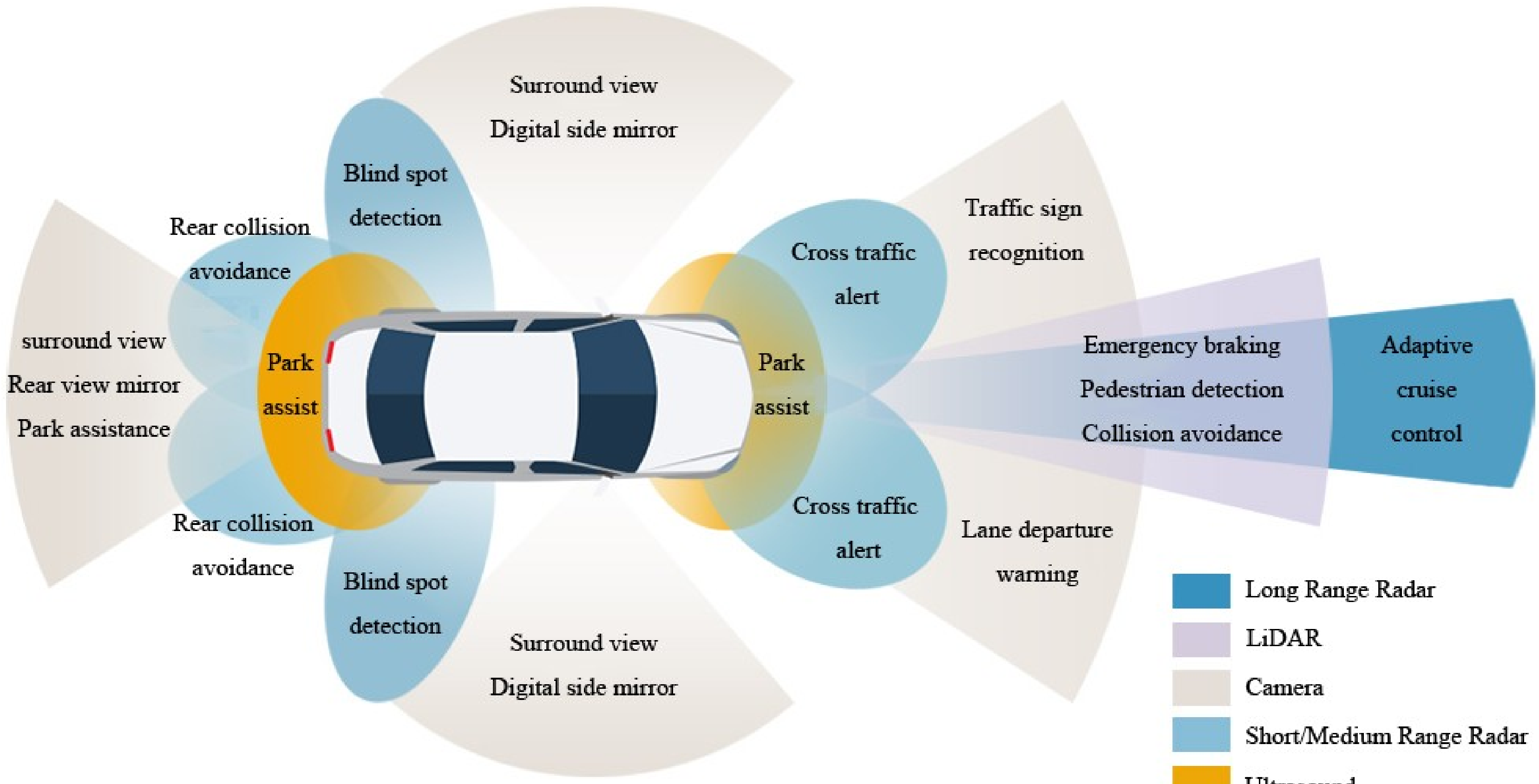
Chengzeng (Charles) You
Applications, Platforms and Systems Security Lab
Department of Computing

Content

Imperial College London

αpss RESEARCH LAB

1. Autonomous Driving Perception



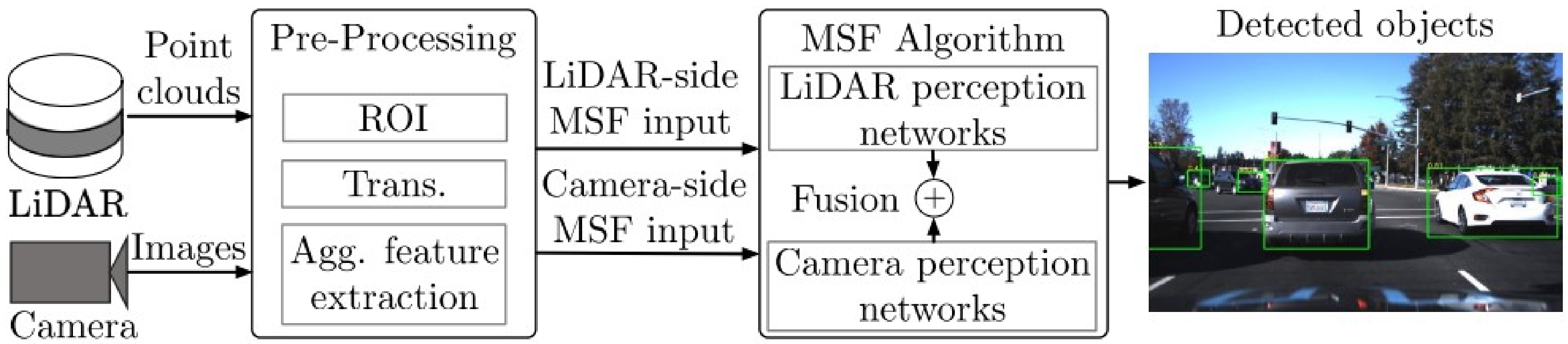
2. Physical-world Attack

- Realistic physical-world attacks: adding stickers, posters, or paintings to traffic signs [5]–[9], or shooting lasers to the LiDAR[10], [11].

Only focus on a single source of AD perception.

3. Multi-Sensor Fusion based Design

- Multi-Sensor Fusion (MSF) fuses the results from different perception sources to achieve overall higher accuracy and robustness [18]–[26].



First study on the security property of MSF-based perception in AD systems.

Imperial College London

αpss RESEARCH LAB

Content

1. Attack Goal and Threat Model

- **Attack goal**: fool the MSF-based AD perception in the victim AV to fail in detecting a front obstacle and thus crash into it.
- **Threat model**: white-box attack

2. Design Challenges

- **C1**. Lack of a single physical-world attack vector effective for both camera- and LiDAR-based AD perception.
- **C2**. Need to differentiably synthesize physically- consistent attack impacts onto both camera and LiDAR.

2. Design Challenges

- **C3**. Need to handle non-differentiable pre-processing steps in AD perception.

making the optimization difficult to be effective

Imperial College London

αpss RESEARCH LAB

Content

Imperial College London

1. Design Overview

- **Adversarial 3D object**: physically-realizable and stealthy attack vector for MSF-based AD perception.
- Causing road safety threats.

Imperial College London

αpss RESEARCH LAB

1. Design Overview

- Optimization-based adversarial 3D object generation.
- -Introduce shape manipulations to normal 3D mesh.
- -Synthesize the raw camera images and LiDAR point clouds.
- -Design the approximation function for the pre-processing step.

Imperial College London

αpss RESEARCH LAB

2. MSF-ADV Methodology Overview

$$\min_{S^a} \quad \mathbb{E}_{t\sim T}[\mathcal{L}_a(t(S^a);\mathcal{R}^l,\mathcal{R}^c,\mathcal{P},\mathcal{M}) + \lambda \cdot \mathcal{L}_r(S^a,\ S)] \quad (1)$$

$$\text{where} \quad \mathrm{PC}^a = \mathcal{R}^l(t(S^a),\ \mathrm{PC}) \quad (2)$$

$$\mathrm{IMG}^a = \mathcal{R}^c(t(\ S^a),\ \mathrm{IMG},\mathrm{C}) \quad (3)$$

$$F^a = \mathcal{P}(\mathrm{PC}^a,\ \mathrm{IMG}^a) \quad (4)$$

$$\mathcal{L}_a(t(S^a);\mathcal{R}^l,\mathcal{R}^c,\mathcal{P},\mathcal{M}) = \mathcal{O}(\mathcal{M}(F^a)) \quad (5)$$

$$\text{subject to} \quad \Delta(S^a,\ S) \le \epsilon \quad (6)$$

S: original benign object   Sa: adversarial one.

M(·): MSF algorithm

La: adversarial loss   Lr(·): realizability loss

Et: Expectation over Transformation (EoT)

λ: balancing hyper-parameter

Imperial College London

αpss RESEARCH LAB

2. MSF-ADV Methodology Overview

$$
\begin{aligned}
\min_{S^a} \quad & \mathbb{E}_{t\sim T}[\mathcal{L}_a(t(S^a);\mathcal{R}^l,\mathcal{R}^c,\mathcal{P},\mathcal{M}) + \lambda \cdot \mathcal{L}_r(S^a,\ S)] && (1)\\
\text{where} \quad & \mathrm{PC}^a = \mathcal{R}^l(t(S^a),\ \mathrm{PC}) && (2)\\
& \mathrm{IMG}^a = \mathcal{R}^c(t(\ S^a),\ \mathrm{IMG},\mathrm{C}) && (3)\\
& F^a = \mathcal{P}(\mathrm{PC}^a,\ \mathrm{IMG}^a) && (4)\\
& \mathcal{L}_a(t(S^a);\mathcal{R}^l,\mathcal{R}^c,\mathcal{P},\mathcal{M}) = \mathcal{O}(\mathcal{M}(F^a)) && (5)\\
\text{subject to} \quad & \Delta(S^a,\ S) \leq \epsilon && (6)
\end{aligned}
$$

S: original benign object Sa: adversarial one.
M(·): MSF algorithm
La: adversarial loss Lr(·): realizability loss
Et: Expectation over Transformation (EoT)
λ: balancing hyper-parameter

2. MSF-ADV Methodology Overview

$$\min_{S^a} \quad \mathbb{E}_{t\sim T}[\mathcal{L}_a(t(S^a); \mathcal{R}^l, \mathcal{R}^c, \mathcal{P}, \mathcal{M}) + \lambda \cdot \mathcal{L}_r(S^a,\ S)] \quad (1)$$

$$\text{where} \quad \mathrm{PC}^a = \mathcal{R}^l(t(S^a),\ \mathrm{PC}) \quad (2)$$

$$\mathrm{IMG}^a = \mathcal{R}^c(t(\ S^a),\ \mathrm{IMG}, \mathrm{C}) \quad (3)$$

$$\boxed{F^a = \mathcal{P}(\mathrm{PC}^a,\ \mathrm{IMG}^a) \quad (4)}$$

$$\mathcal{L}_a(t(S^a); \mathcal{R}^l, \mathcal{R}^c, \mathcal{P}, \mathcal{M}) = \mathcal{O}(\mathcal{M}(F^a)) \quad (5)$$

$$\text{subject to} \quad \Delta(S^a,\ S) \le \epsilon \quad (6)$$

S: original benign object Sa: adversarial one.
M(·): MSF algorithm
La: adversarial loss Lr(·): realizability loss
Et: Expectation over Transformation (EoT)
λ: balancing hyper-parameter

2. MSF-ADV Methodology Overview

$$\min_{S^a} \quad \mathbb{E}_{t\sim T}[\mathcal{L}_a(t(S^a);\mathcal{R}^l,\mathcal{R}^c,\mathcal{P},\mathcal{M}) + \lambda \cdot \mathcal{L}_r(S^a,\ S)] \quad (1)$$

$$\text{where} \quad \mathrm{PC}^a = \mathcal{R}^l(t(S^a),\ \mathrm{PC}) \quad (2)$$

$$\mathrm{IMG}^a = \mathcal{R}^c(t(\ S^a),\ \mathrm{IMG},\mathrm{C}) \quad (3)$$

$$F^a = \mathcal{P}(\mathrm{PC}^a,\ \mathrm{IMG}^a) \quad (4)$$

$$\boxed{\mathcal{L}_a(t(S^a);\mathcal{R}^l,\mathcal{R}^c,\mathcal{P},\mathcal{M}) = \mathcal{O}(\mathcal{M}(F^a)) \quad (5)}$$

$$\text{subject to} \quad \Delta(S^a,\ S) \le \epsilon \quad (6)$$

S: original benign object Sa: adversarial one.
M(·): MSF algorithm
La: adversarial loss Lr(·): realizability loss
Et: Expectation over Transformation (EoT)
λ: balancing hyper-parameter

Imperial College London

RESEARCH LAB

2. MSF-ADV Methodology Overview

$$\min_{S^a} \quad \mathbb{E}_{t\sim T}[\mathcal{L}_a(t(S^a);\mathcal{R}^l,\mathcal{R}^c,\mathcal{P},\mathcal{M}) + \lambda\cdot\mathcal{L}_r(S^a,\ S)] \quad (1)$$

$$\text{where} \quad \mathrm{PC}^a = \mathcal{R}^l(t(S^a),\ \mathrm{PC}) \quad (2)$$

$$\mathrm{IMG}^a = \mathcal{R}^c(t(\ S^a),\ \mathrm{IMG},\mathrm{C}) \quad (3)$$

$$F^a = \mathcal{P}(\mathrm{PC}^a,\ \mathrm{IMG}^a) \quad (4)$$

$$\mathcal{L}_a(t(S^a);\mathcal{R}^l,\mathcal{R}^c,\mathcal{P},\mathcal{M}) = \mathcal{O}(\mathcal{M}(F^a)) \quad (5)$$

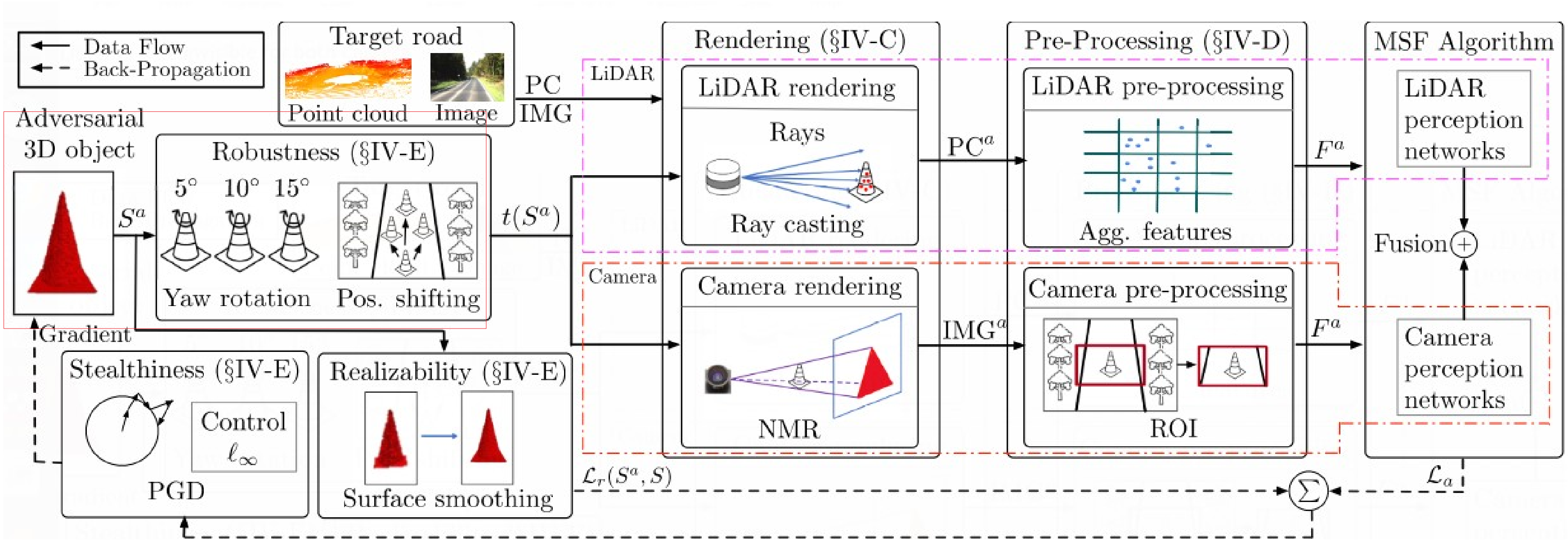
$$\boxed{\text{subject to} \quad \Delta(S^a,\ S) \le \epsilon \quad (6)}$$

S: original benign object Sa: adversarial one.
M(·): MSF algorithm
La: adversarial loss Lr(·): realizability loss
Et: Expectation over Transformation (EoT)
λ: balancing hyper-parameter

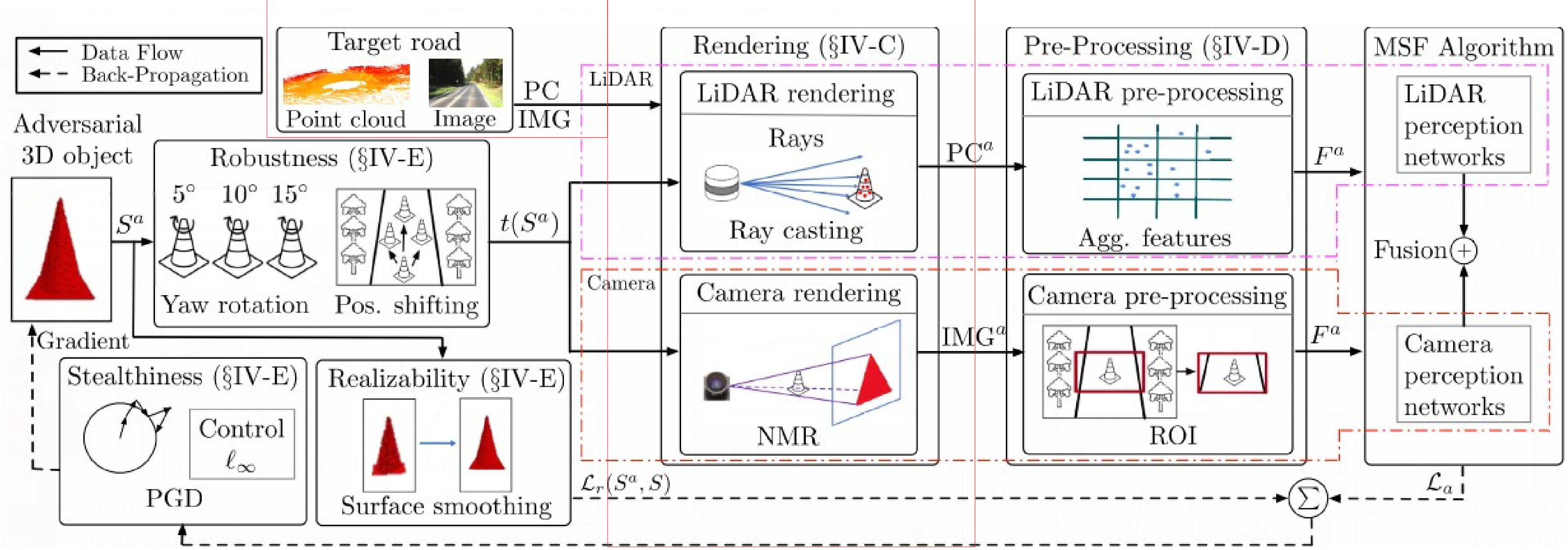
Imperial College London

αpss RESEARCH LAB

2. MSF-ADV Methodology Overview

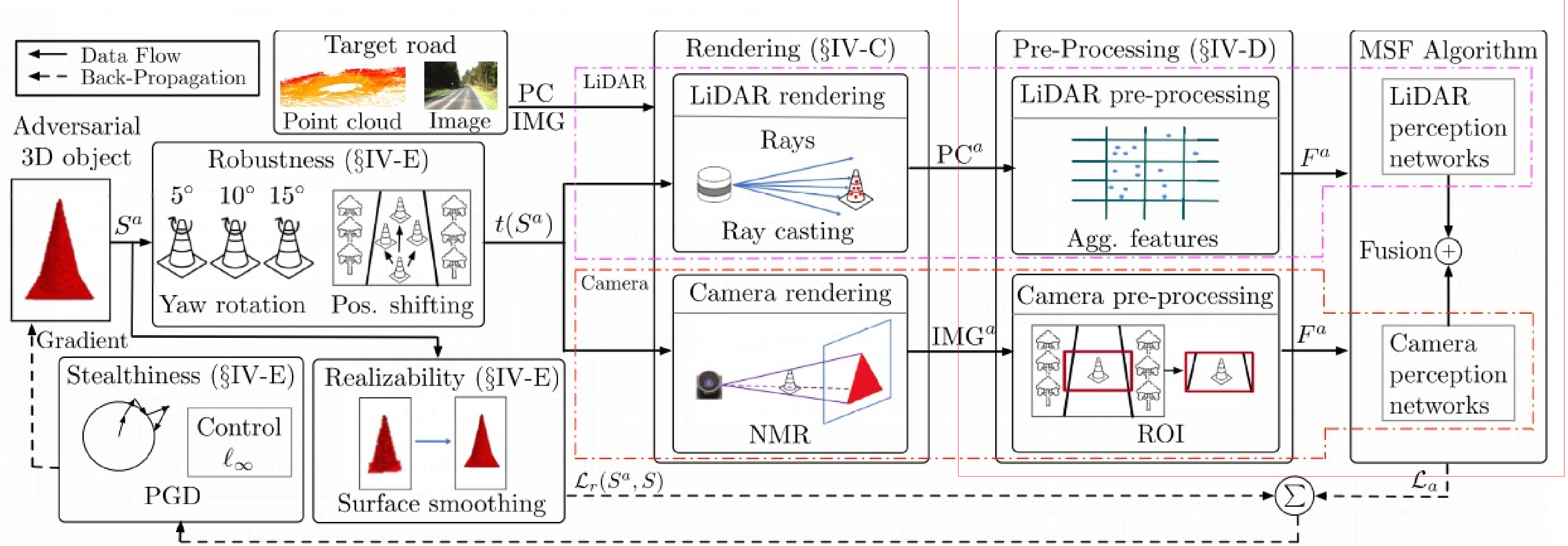


2. MSF-ADV Methodology Overview



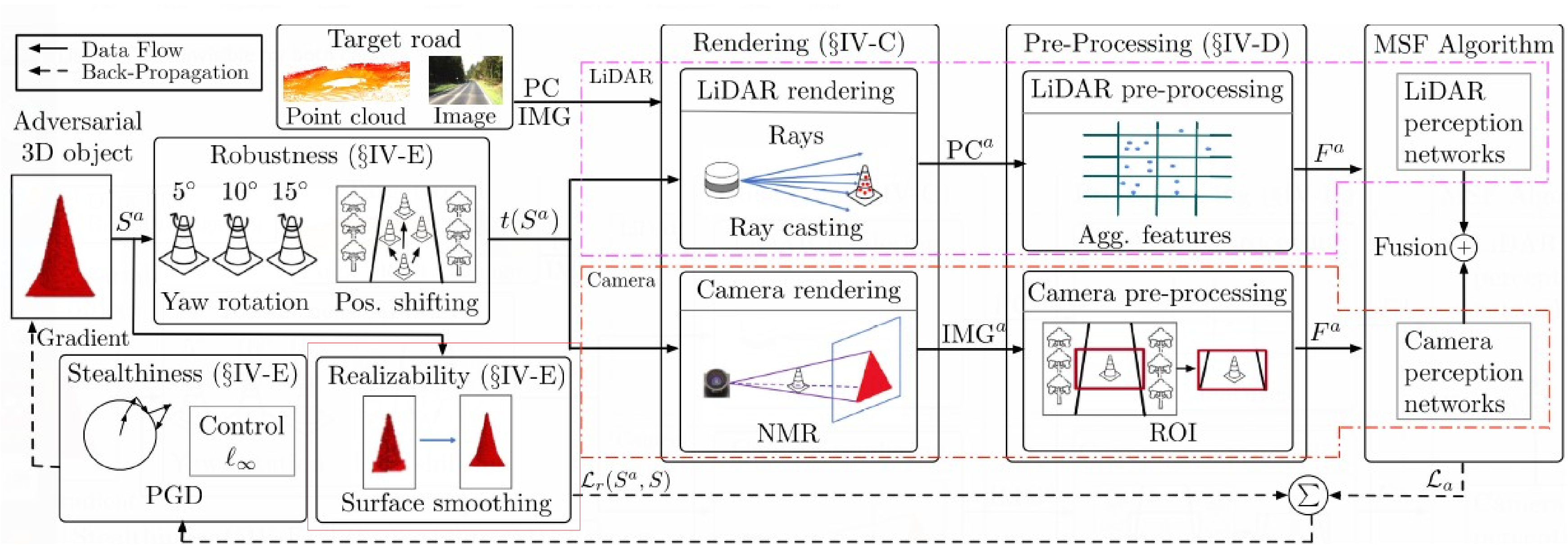
Imperial College London

2. MSF-ADV Methodology Overview

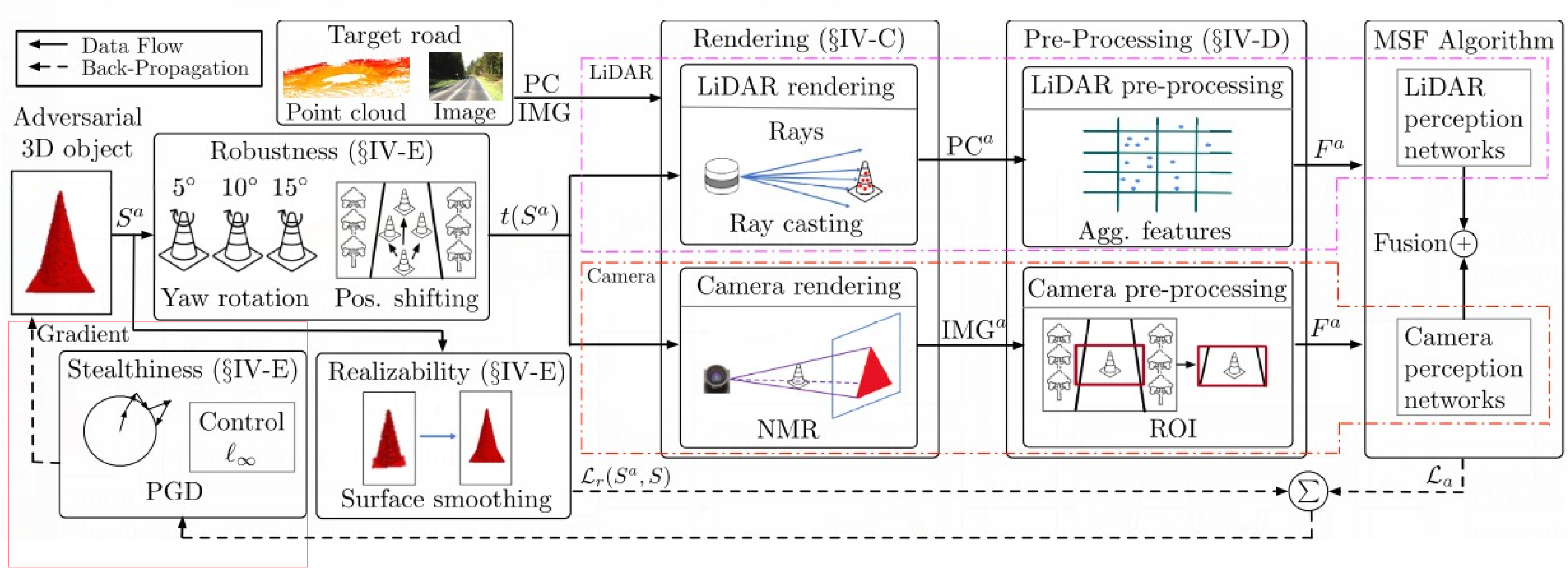


Imperial College London

2. MSF-ADV Methodology Overview



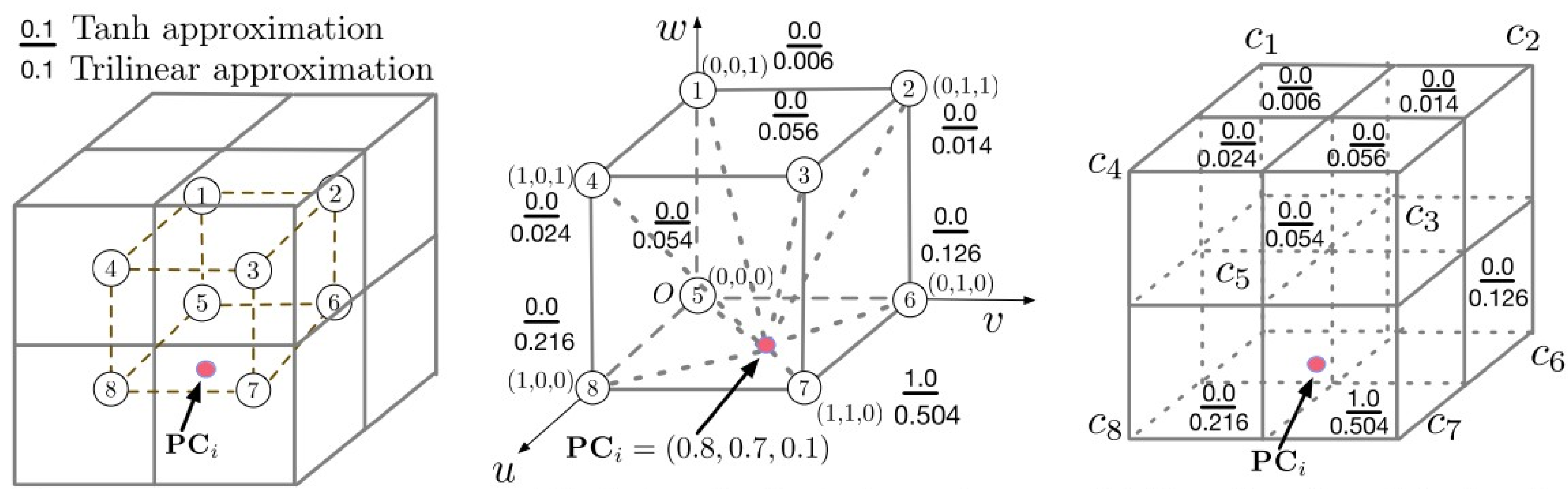
2. MSF-ADV Methodology Overview



3. Differentiable Rendering

- Define Sa in the LiDAR coordinate system.
- Use a calibration matrix C to transform Sa from the LiDAR coordinate system to the camera coordinate system.

4. Pre-Processing Step Approximation



(a) 8 cells & formed cube (b) Soft point-inclusion calc. (c) Result assigned to 8 cells

Imperial College London

αpss RESEARCH LAB

4. Pre-Processing Step Approximation

$$\begin{aligned}\mathrm{softPI}(\mathbf{PC}_i, c_m) =& (1 - \frac{d(u_m, u_i)}{L}) \cdot (1 - \frac{d(v_m, v_i)}{W}) \\ & \cdot (1 - \frac{d(w_m, w_i)}{H})\end{aligned} \tag{7}$$

$$d(u_1, u_2) = \frac{L}{2} + \frac{L}{2} \cdot \tanh(\mu \cdot (\mid u_1 - u_2 \mid - \frac{L}{2})) \tag{8}$$

4. Pre-Processing Step Approximation

With an accurate SoftPI(·), we can then differentiably approximate all the cell-level aggregated features:

- **Count and density**. The count feature calculates the number of points in a cell. The density feature calculates the density of points in a cell.
- **Occupancy**. The occupancy feature calculates whether a cell has points or not.
- **Height and intensity**. The max/min/mean height features calculate the maximum, minimum, and the average height of the points inside a cell.

5. Objective Function Design

Adversarial Loss La: minimize the confidence value of the regions of Sa.

Realizability Loss Lr(·): (1) improve the printability of Sa at 3D printers(2) prevent the generation of Sa that is underneath the road surface.

Improving the stealthiness of Sa: (1)the realizability loss above can improve its surface smoothness. (2)control how small Sa looks compared to the benign one S.

Improving the robustness of Sa: implement Transformation T via random yaw-dimension rotations and ground-plane position shifting of Sa.

Content

1. Setup

MSF algorithm selection. On the LiDAR side, Apollo v5.5 and v2.5. On the camera side, the latest version of Apollo and the pre-trained YOLO v3.

3D object type selection. (1) a traffic cone of size 0.5 m × 0.5 m × 1.0 m, for A5-L + A5-C and A2-L + A5-C, (2) a bench of size 0.6 m × 0.5 m × 1.5m, for A5-L + Y3 and A2-L + Y3, and (3) a toy car of size 0.6 m × 0.7 m × 1.6 m for all 4 MSF combinations.

Attack scenario selection. For each object type, we select 100 real-world driving scenarios from the KITTI dataset.

Object placement. 7 meters (m) in front of the victim.

Imperial College London

αpss RESEARCH LAB

2. Attack Effectiveness

MSF Comb.		A5-L⊕A5-C		A5-L⊕Y3		A2-L⊕A5-C		A2-L⊕Y3	
Object Type		Traffic cone	Toy car	Bench	Toy car	Traffic cone	Toy car	Bench	Toy car
Success Rate		100%	91%	100%	93%	100%	96%	100%	97%
Dist. (cm)	$\Delta\ell_1$	5.92	5.95	5.93	5.97	5.93	5.63	5.90	5.61
	$\Delta\ell_2$	3.28	3.46	3.39	3.37	3.43	3.34	3.30	3.25
	$\Delta\ell_\infty$	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
LPIPS		0.06	0.02	0.20	0.04	0.07	0.17	0.20	0.06

Imperial College London

αpss RESEARCH LAB

2. Attack Effectiveness

MSF Comb.		A5-L⊕A5-C		A5-L⊕Y3		A2-L⊕A5-C		A2-L⊕Y3	
Object Type		Traffic cone	Toy car	Bench	Toy car	Traffic cone	Toy car	Bench	Toy car
Success Rate		100%	91%	100%	93%	100%	96%	100%	97%
Dist. (cm)	$\Delta\ell_1$	5.92	5.95	5.93	5.97	5.93	5.63	5.90	5.61
	$\Delta\ell_2$	3.28	3.46	3.39	3.37	3.43	3.34	3.30	3.25
	$\Delta\ell_\infty$	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
LPIPS		0.06	0.02	0.20	0.04	0.07	0.17	0.20	0.06

Imperial College London

αpss RESEARCH LAB

2. Attack Effectiveness

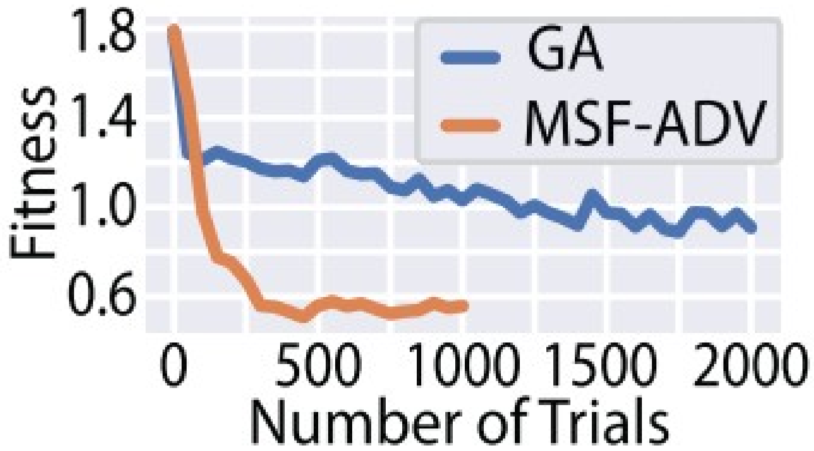
MSF Comb.		A5-L⊕A5-C		A5-L⊕Y3		A2-L⊕A5-C		A2-L⊕Y3	
Object Type		Traffic cone	Toy car	Bench	Toy car	Traffic cone	Toy car	Bench	Toy car
Success Rate		100%	91%	100%	93%	100%	96%	100%	97%
Dist. (cm)	$\Delta\ell_1$	5.92	5.95	5.93	5.97	5.93	5.63	5.90	5.61
	$\Delta\ell_2$	3.28	3.46	3.39	3.37	3.43	3.34	3.30	3.25
	$\Delta\ell_\infty$	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
LPIPS		0.06	0.02	0.20	0.04	0.07	0.17	0.20	0.06

Imperial College London

3. Comparison with Baseline Attack Methods

Attack Method	Success Rate	$\Delta\ell_p$ Dist. (cm)		
		$\Delta\ell_1$	$\Delta\ell_2$	$\Delta\ell_\infty$
GN	8%	21.8	3.35	10.3
GA	9%	2.85	1.84	2.00
Ours	100%	5.92	3.28	2.00

3. Comparison with Baseline Attack Methods



4. Attack Robustness

	Y = (-0.1 m, 0.1 m)		
	X = (5 m, 15 m)	(15 m, 25 m)	(25 m, 35 m)
w/o EoT	80.3%	79.2%	79.9%
w/ EoT	96.3%	95.5%	96.6%

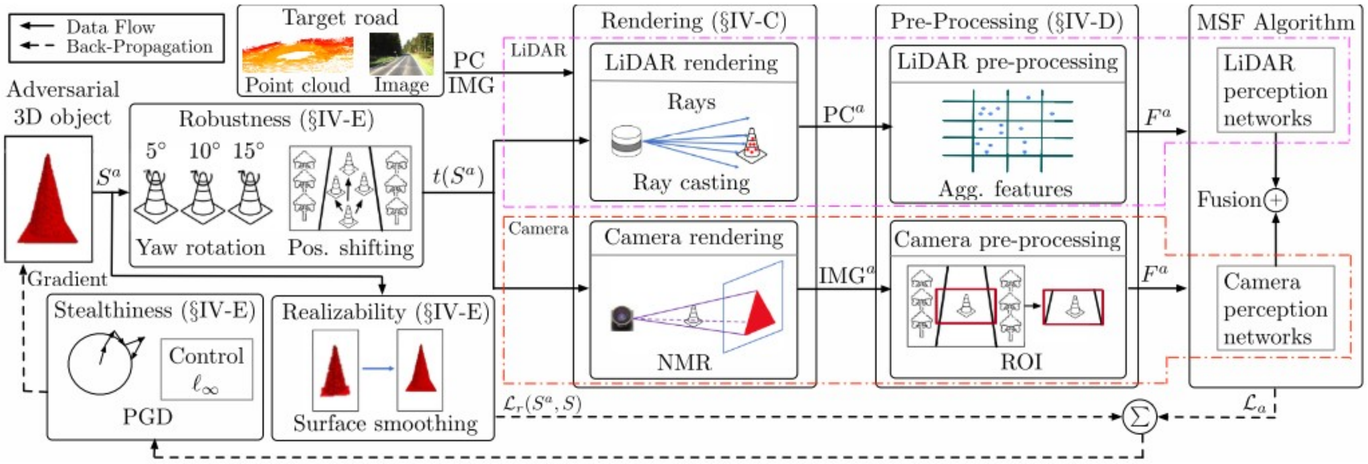
Table V. Average success rate on A5-L⊕A5-C with traffic cone in different victim approaching distance ranges.

Content

1. Limitations

Did not perform an end-to-end attack evaluation on a real AV in the physical world due to the cost and safety considerations.
there also exists another type of fusion design: DNN-based fusion [18]– [24]. Thus, it is still unclear how effective MSF-ADV can be for DNN-based MSF algorithms

Thank you



Data Flow
Back-Propagation
Target road
Point cloud
Image
PC
IMG
LiDAR
Adversarial 3D object
Robustness (§IV-E)
5° 10° 15°
Yaw rotation
Pos. shifting
S^a
$t(S^a)$
Rendering (§IV-C)
LiDAR rendering
Rays
Ray casting
Camera
Camera rendering
NMR
PC^a
IMG^a
Pre-Processing (§IV-D)
LiDAR pre-processing
Agg. features
Camera pre-processing
ROI
F^a
F^a
MSF Algorithm
LiDAR perception networks
Fusion
Camera perception networks
Gradient
Stealthiness (§IV-E)
Control ℓ_∞
PGD
Realizability (§IV-E)
Surface smoothing
$\mathcal{L}_r(S^a, S)$
$\mathcal{L}_a$