Segment-Anything Models Achieve Zero-shot Robustness in Autonomous Driving

Jun Yan[1] Pengyu Wang[1] Danni Wang[1] Weiquan Huang[1] Daniel Watzenig [2] Huilin Yin[1]

[1]*College of Electronic and Information Engineering, Tongji University, China*

2 Graz University of Technology

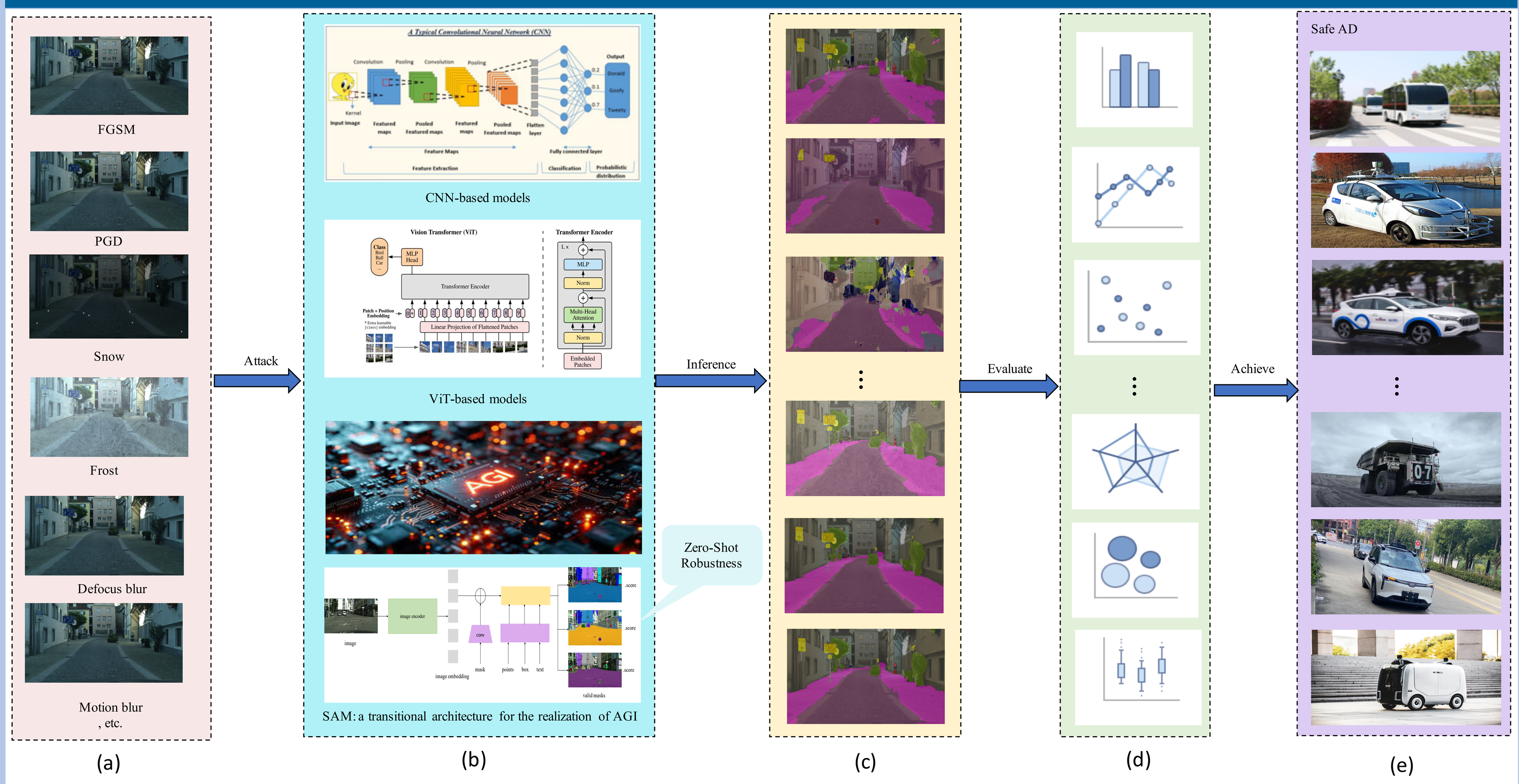
Introduction

Semantic segmentation models are vulnerable to adversarial attacks. Previous study[1] has explored the robustness of the classical CNN-based semantic segmentation models. The emergence of new visual foundation models like SAM[2] calls for the new research paradigm in validation of autonomous driving.

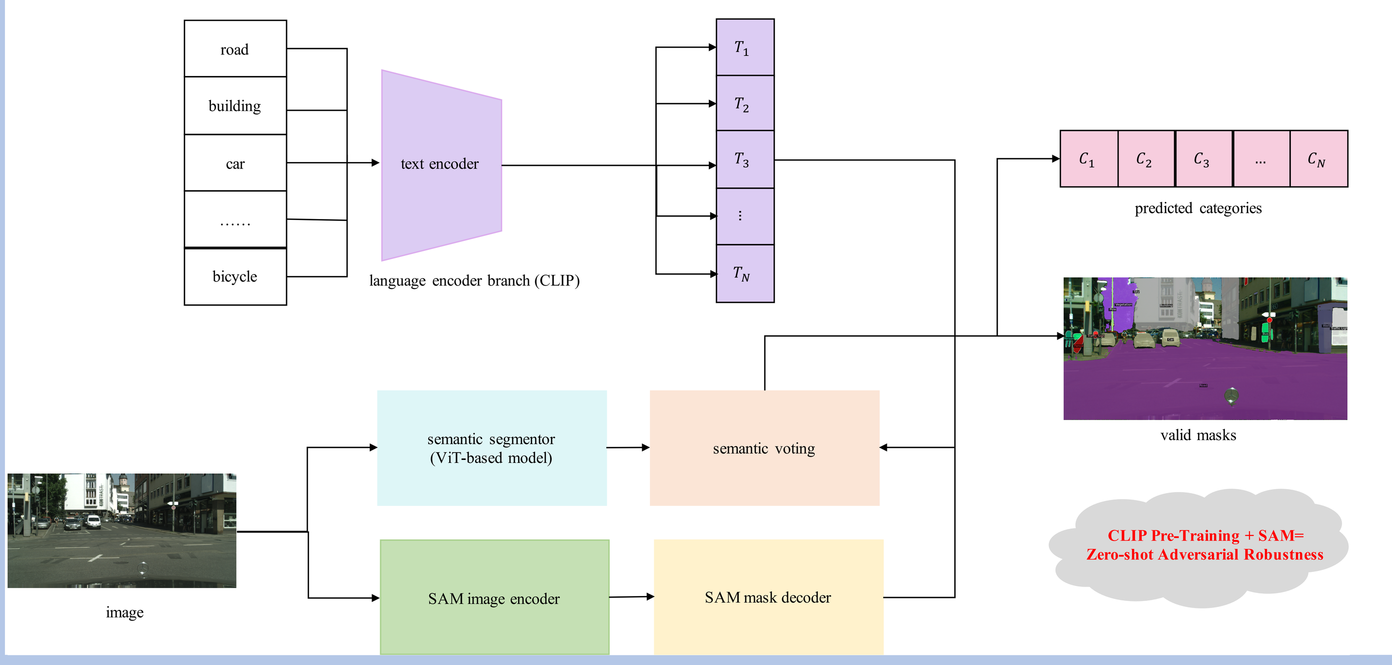
The main contributions of **this work** are summarized below:

- (Methodology-wise) In the semantic segmentation task, this study shows a SAM pipeline with the assistance of text encoder achieves a robust in-context learning ability under the adversarial attacks.
- (Empirical-study-wise) We evaluate the robustness of CNN models, ViT models, and SAM models under the white-box attacks and black-box attacks on the dataset of Cityscapes.

Method Overview



Proposed Framework



SAM
Adding CLIP=
Robust SSAM
(Segment SAM)

Black-box Attacks

TABLE I
ROBUSTNESS OF SEMANTIC SEGMENTATION MODELS UNDER THE BLACK-BOX CORRUPTIONS (SEVERITY=1-3)

		Blur					Weather				Noise				Digital					
Architecture	**clean**	Gaussian	**defocus**	**motion**	glass	zoom	**snow**	**frost**	**fog**	**spatter**	speckle	Gaussian	shot	impulse	brightness	contrast	JPEG	saturate	pixelate	elastic
DP-ResNet50 [4]	76.6	66.0	62.2	65.3	52.9	24.3	19.6	33.7	72.5	45.3	32.8	10.7	14.0	9.5	74.4	70.6	34.0	74.6	63.3	74.3
DP-ResNet101 [4]	77.3	67.6	64.9	66.4	54.2	26.8	24.9	37.9	74.0	52.3	41.2	15.6	20.2	14.3	75.1	73.3	40.4	75.5	66.6	74.6
DP-MobileNetV2 [4]	72.9	59.6	55.8	60.0	44.6	21.8	19.6	28.1	66.9	47.8	26.0	9.6	10.6	12.1	69.4	64.3	22.1	69.8	54.7	71.1
DP-Xception65 [4]	78.4	71.6	68.6	69.6	62.5	25.0	19.9	26.5	65.8	60.8	28.4	7.0	8.0	5.1	66.6	71.9	34.1	67.1	71.1	76.2
FCN-ResNet50 [1]	67.9	55.9	52.5	56.5	43.4	22.6	17.3	31.1	48.6	45.2	10.1	1.8	2.2	1.5	54.6	60.4	15.4	52.0	33.6	64.8
FCN-ResNet101 [1]	69.0	57.3	54.1	56.6	43.5	21.9	17.3	27.9	53.4	47.1	17.0	4.8	5.1	6.0	58.0	59.6	22.9	52.8	39.8	65.9
FCN32s-VGG16 [1]	55.1	42.5	38.9	41.6	28.7	17.9	12.9	18.8	34.8	36.8	14.0	7.7	8.8	2.7	40.0	41.7	15.5	38.2	22.1	52.7
FCN16s-VGG16 [1]	58.4	42.9	38.4	42.9	29.8	18.0	12.6	17.9	33.5	37.7	8.6	2.4	3.3	2.3	40.9	44.0	14.3	39.3	21.9	56.1
FCN8s-VGG16 [1]	60.3	44.3	40.1	43.4	32.2	18.0	12.2	16.9	34.0	39.4	8.6	3.2	3.8	2.3	42.8	43.6	15.4	41.9	22.4	57.5
PSPNet-ResNet50 [3]	69.3	57.1	54.1	57.4	35.4	22.8	10.3	17.8	45.0	42.9	9.5	2.8	3.1	4.6	56.8	56.1	13.9	49.0	25.9	65.6
PSPNet-ResNet101 [3]	70.7	58.8	55.4	59.0	43.8	24.6	15.4	32.7	54.9	44.1	12.9	3.4	4.5	3.5	60.3	60.9	27.1	53.6	43.6	66.9
SegNet-VGG16 [2]	62.7	52.4	49.0	54.8	51.8	23.4	18.5	24.9	42.4	49.5	40.2	18.4	22.8	13.0	56.2	52.9	37.7	52.9	62.0	61.5
OCRNet-ResNet100 [7]	80.2	68.9	68.9	69.0	56.3	24.1	21.4	43.4	76.3	56.6	31.7	6.4	9.6	12.9	79.1	73.4	28.5	78.6	69.3	78.4
OCRNet-HRNet-W48 [7]	80.5	70.7	70.3	71.1	63.0	21.5	18.7	44.2	75.2	63.9	42.4	16.2	17.8	18.0	79.5	76.5	36.1	78.0	76.6	78.9
OCRNet-HRNet-W18s [7]	73.6	59.0	62.3	62.9	52.8	20.1	18.5	33.7	61.0	54.6	29.5	10.9	11.8	8.8	69.8	69.3	30.4	70.0	69.5	72.4
ISANet (ResNet50) [6]	78.4	64.7	63.2	64.6	51.0	19.7	11.8	33.5	69.5	50.2	30.0	9.0	11.8	11.7	76.8	69.0	23.0	76.3	60.8	76.1
ISANet (ResNet101) [6]	79.6	67.4	67.6	67.6	56.2	20.0	19.8	37.1	75.1	55.1	33.9	10.6	13.8	14.2	78.2	72.8	28.9	77.7	65.3	76.9
STDC (Pre-training) [8]	75.0	64.5	65.1	64.9	54.7	21.8	11.3	32.2	68.9	51.1	29.0	26.4	10.9	6.4	73.2	67.9	31.1	73.1	61.0	73.2
STDC (No pre-training) [8]	71.8	59.6	63.6	62.5	52.5	24.6	11.4	27.4	59.1	48.9	35.0	12.4	15.1	12.7	68.9	62.1	49.6	69.0	70.9	71.1
SegFormer-b3 [5]	81.9	74.5	74.2	74.2	68.1	31.6	43.8	55.1	79.2	70.4	68.5	51.8	57.2	50.5	81.4	80.7	60.6	81.1	74.1	80.5
SegFormer-b0 [5]	76.5	64.6	67.5	67.7	57.2	27.5	27.5	40.9	71.2	56.3	51.9	26.5	31.1	27.6	74.9	73.7	47.5	74.3	68.1	74.3
OneFormer-SwinTransfomer [30]	83.0	79.8	78.0	77.4	73.9	35.2	65.9	56.8	81.6	78.8	77.8	67.5	72.7	73.6	82.5	82.0	71.8	82.5	77.4	81.7
OneFormer-ConvXNet [30]	83.0	79.0	78.0	77.4	73.9	35.2	57.3	56.6	82.3	76.1	79.1	71.0	74.4	72.3	82.7	82.3	71.0	82.4	78.6	81.0
SAM-SegFormer	73.0	65.7	63.1	64.4	63.1	24.5	38.9	44.1	67.5	60.0	63.5	52.9	56.7	50.5	71.7	69.6	60.9	72.8	69.3	71.6
SAM-OneFormer	80.0	75.5	73.7	72.3	70.8	29.5	58.3	51.7	76.9	72.4	74.0	63.9	68.7	67.6	78.9	77.2	68.9	79.5	72.7	77.8
MobileSAM-SegFormer	68.9	61.7	59.0	60.0	58.9	20.9	32.1	37.4	60.7	53.7	57.5	47.7	51.4	44.9	67.5	61.9	56.3	68.5	65.5	67.6
MobileSAM-OneFormer	75.3	70.5	68.9	67.5	66.5	25.7	44.2	44.6	69.7	64.0	67.0	56.5	61.1	58.5	73.9	68.5	63.3	74.6	68.5	73.1

TABLE II
ROBUSTNESS OF SEMANTIC SEGMENTATION MODELS UNDER THE WORST BLACK-BOX CORRUPTIONS (SEVERITY=5).

		Blur					Weather				Noise				Digital					
Architecture	**clean**	Gaussian	**defocus**	**motion**	glass	zoom	**snow**	**frost**	**fog**	**spatter**	speckle	Gaussian	shot	impulse	brightness	contrast	JPEG	saturate	pixelate	elastic
SAM-SegFormer	73.0	38.2	42.4	47.3	41.5	17.4	23.6	25.7	57.2	47.5	47.1	20.1	27.7	22.0	68.7	48.0	38.8	64.6	59.4	67.9
SAM-OneFormer	80.0	58.2	59.9	55.6	53.6	19.4	36.0	34.2	70.4	63.3	62.0	26.0	30.4	32.2	76.6	62.4	48.1	71.7	13.4	73.1
MobileSAM-SegFormer	68.9	35.6	39.8	42.7	38.0	14.1	16.0	19.7	54.1	24.2	40.8	15.7	22.0	15.6	68.7	38.8	35.5	59.7	56.4	63.9
MobileSAM-OneFormer	75.3	52.9	55.4	50.8	50.3	16.0	21.0	25.0	66.3	31.4	67.0	20.4	25.7	25.4	76.6	49.7	43.8	65.9	13.2	69.1

The robustness of SAMs under the black-box corruptions is considerable, which is beneficial for the SOTIF in autonomous driving.

White-box Attacks

TABLE III
ROBUSTNESS (MIOU) OF DIFFERENT SEMANTIC SEGMENTATION MODELS UNDER THE FGSM ATTACKS ($\varepsilon = 8.0/255.0, 16.0/255.0$).

	$\varepsilon = 8.0/255.0$	$\varepsilon = 16.0/255.0$
DP-ResNet50 [4]	36.9	16.6
DP-ResNet101 [4]	44.1	19.3
DP-MobileNetV2 [4]	30.6	10.8
DP-Xception65 [4]	17.1	4.9
FCN32s-ResNet50 [1]	15.4	3.3
FCN32s-ResNet101 [1]	26.6	10.0
FCN32s-VGG16 [1]	14.3	7.2
FCN16s-VGG16 [1]	10.3	6.8
FCN8s-VGG16 [1]	12.2	6.5
PSPNet-ResNet50 [3]	12.3	6.0
PSPNet-ResNet101 [3]	18.3	3.9
SegNet-VGG16 [2]	22.9	10.8
STDC (Pre-training) [8]	9.8	2.0
STDC (No pre-training) [8]	11.7	4.8
SegFormer_mit-b5 [5]	56.6	49.2
SegFormer_mit-b3 [5]	49.8	37.2
SegFormer_mit-b0 [5]	35.3	22.0
OCRNet-ResNet100 [7]	11.6	1.7
OCRNet-HRNet-W48 [7]	30.3	3.5
OCRNet-HRNet-W18 [7]	11.6	2.4
ISANet (ResNet50) [6]	13.9	3.2
ISANet (ResNet101) [6]	29.7	8.1
SAM-SegFormer	51.6	44.8
SAM-OneFormer	59.1	57.6
MobileSAM-SegFormer	49.4	43.0
MobileSAM-OneFormer	56.3	52.9

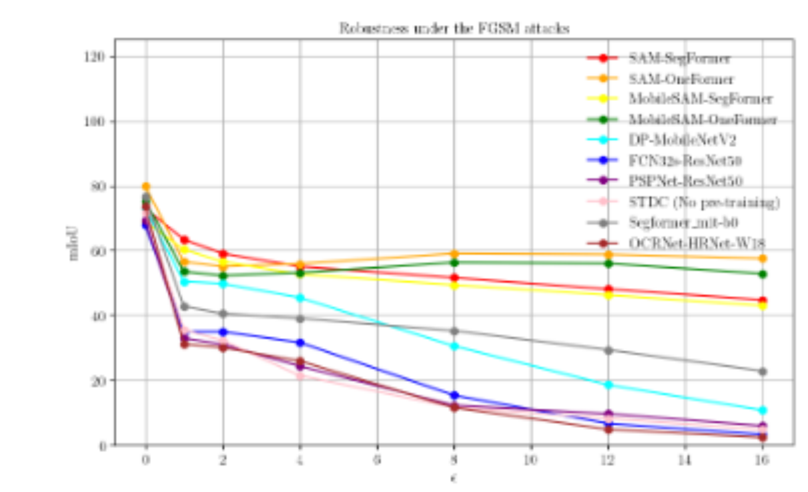


Fig. 3. Robustness study under the FGSM attacks

TABLE IV
ROBUSTNESS (MIOU) OF DIFFERENT SEMANTIC SEGMENTATION MODELS UNDER THE PGD-10 ATTACKS ($\varepsilon = 8.0/255.0, 16.0/255.0$).

	$\varepsilon = 8.0/255.0$	$\varepsilon = 16.0/255.0$
PSPNet [3]	28.8	26.0
DeepLabV3 [49]	29.5	26.5
SAM-SegFormer	21.6	21.3
SAM-OneFormer	53.5	52.1
MobileSAM-SegFormer	19.8	20.6
MobileSAM-OneFormer	49.6	52.1

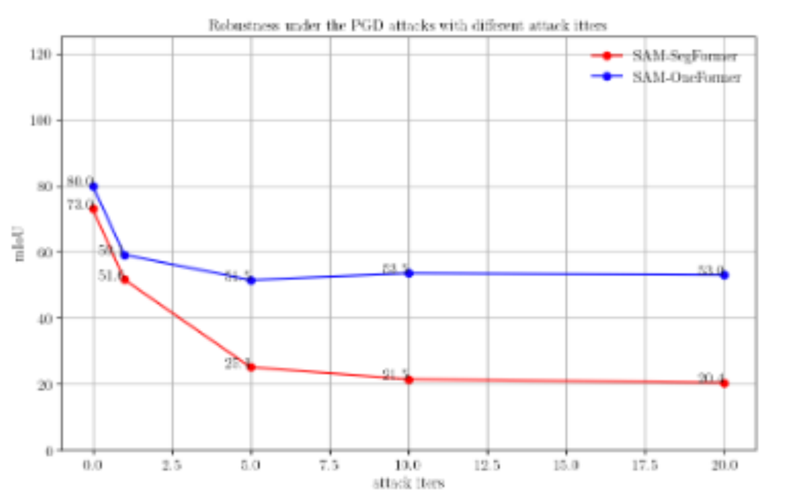


Fig. 4. Robustness study under the PGD attacks

TABLE V
COMPARISON BETWEEN SAM AND MOBILESAM

	Params (M)	Inference Speed (ms)
SAM [18]	632	452
MobileSAM [34]	5.78	8

The white-box can be attribute to the security issue. SAMs are robust under these gradient-based attacks. However, the defense is built upon the randomization that the model trained on the generalize data is transferred to the Cityscapes dataset.

Conclusion

- Overparameterization + broad training data = zero-shot robustness
- OneFormer backbone with the task unification is more robust
- MobileSAM is more suitable for the real scenarios, which requires further adversarial finetuning
- More testing scenarios like SegPGD and CosPGD is essential
- This framework can be scaled to aerial vehicles or unmanned boats

Acknowledgment

This work was supported by the Shanghai International Science and Technology Cooperation Project No.22510712000 and the Special Funds of the Tongji University for "Sino-German Cooperation 2.0 Strategy" No. ZD2023001. The authors would like to thank TÜV SUD for the kind and generous support. We thank sincerely to Ms. He Zhang and Mr. Ronghui Liu for their help.

Reference

[1] Yin, Huilin, et al. "On adversarial robustness of semantic segmentation models for automated driving." *IV*, 2022.
[2] Kirillov, Alexander, et al. "Segment anything." *ICCV*. 2023.

IEEE IAVVC2024
IEEE International Automated Vehicle Validation Conference
October 21-23, 2024 | Pittsburgh, PA