DeepMTD

Moving Target Defense for Embedded Deep Visual Sensing against Adversarial Examples

针对嵌入视觉对抗样本的移动目标防御

Qun Song
宋群

Zhenyu Yan
鄢振宇

Rui Tan
谭睿

School of Computer Science & Engineering
Nanyang Technological University, Singapore
南洋理工大学, 计算机科学与工程学院

Deep Learning in Embedded Sensing

- Increasing applications
  – Automotive, healthcare, consumer electronics, etc
- Vulnerable to adversarial examples
  – Crafted inputs to mislead deep models, unnoticeable to human eyes
- Attacks in real world
  – Road sign classifiers, lane detectors

Stop Sign → Speed Limit Sign

[CVPR ’18]

Credit: Keen Security Lab *

Adversarial Examples

$$\min_{x^{adv}} \left\|x - x^{adv}\right\| \; s.t. \; \mathrm{NN}\left(x^{adv}\right) \neq \mathrm{NN}(x)$$

- Adversary's goal
  - **Targeted**: input misclassified to a specific class
  - **Non-targeted**: input misclassified to any class
- Adversary's knowledge
  - **Black box**: no/limited knowledge of model internals
  - **White box**: complete/lots of knowledge of model internals

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Related Work

Defenses

Model hardening

- **Adversarial training** [ICLR '14, ICLR '18]
  - Train on adversarial examples
  - Effective to considered adversarial examples only [NeurIPS '18]
- **Gradient masking** [S&P '16, ICLR '18]
  - Make gradients nonexistent or incorrect, randomized, or vanishing/exploding
  - Incomplete defense [ICLR '18]
  - Can be defeated by stronger attack [ICML '18]

Modified input

- **Data compression** [ICLR '18]
  Foveation [ICLR '16]
  Randomization [ICCV '17]
  - Result in loss of classification accuracy on clean examples [arXiv:1705.10686]
  - Does not affect adaptive attacker [ICML '18]

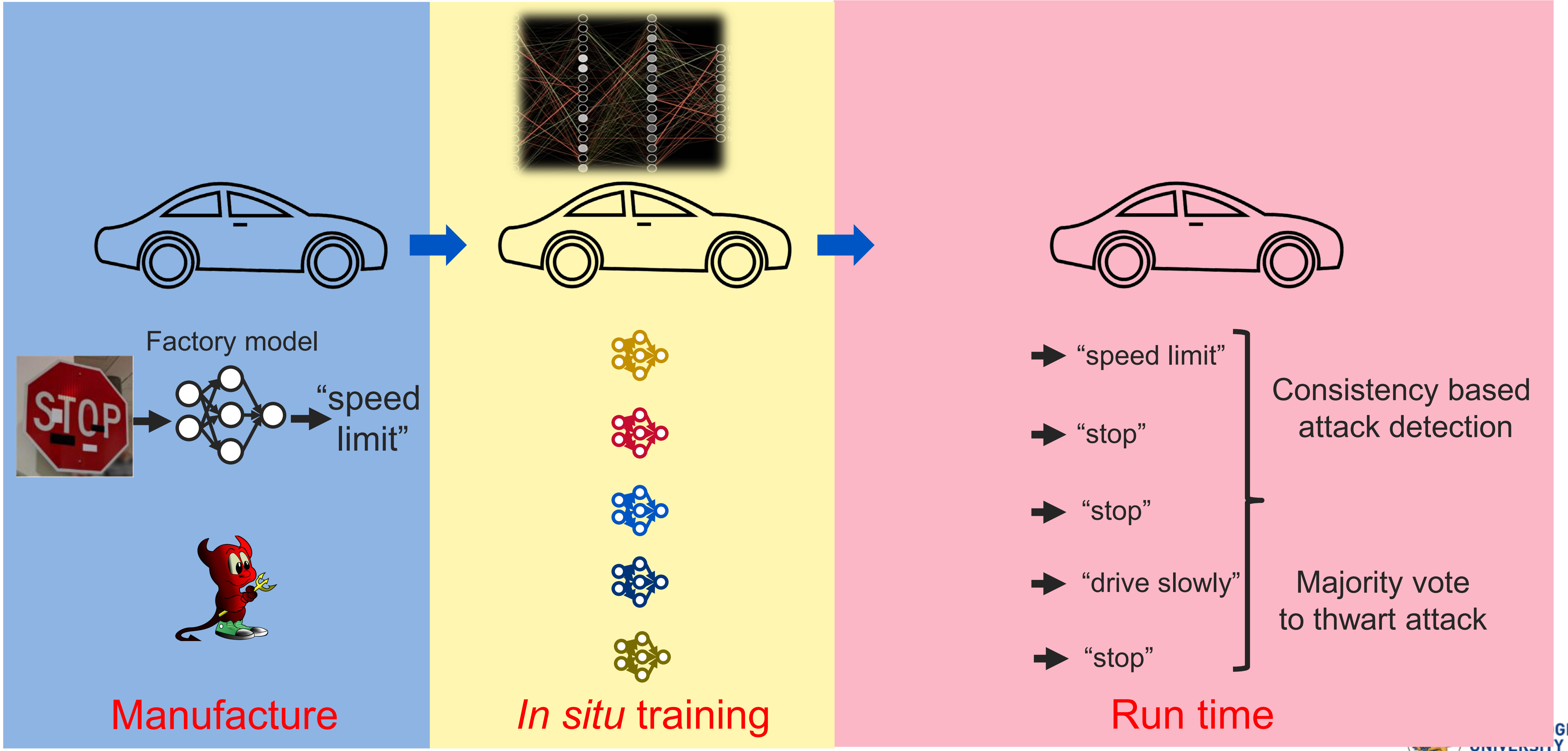
Static defense

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Moving Target Defense (MTD)

- Static defenses grant the advantage of time to attackers
- MTD revokes the advantage

Preview: MTD against Adversarial Examples

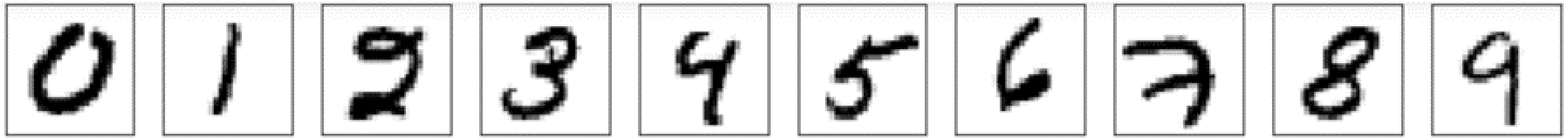


Outline

- Background & Motivation
- **Approach Design & Evaluation**
- Implementation
- Conclusion

Used Datasets

- **MNIST**: 10 handwritten digits



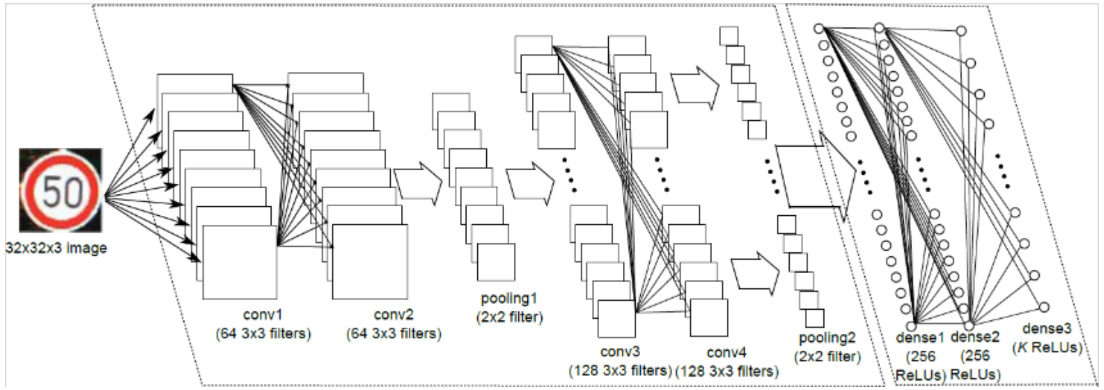
- **CIFAR10**: 10 classes of objects

- **German Traffic Sign Recognition Benchmark (GTSRB)**: 43 classes

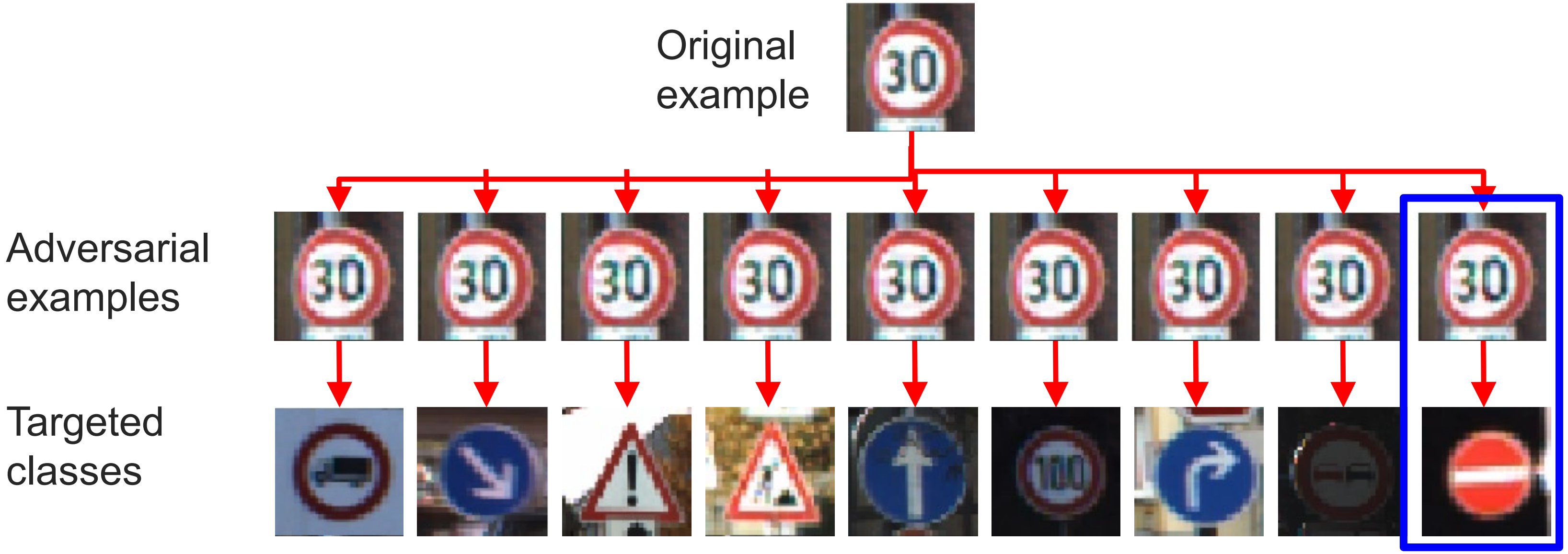
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Deep Model & Adversarial Examples

- Training and validation accuracy of **99.93%** and **96.64%** on GTSRB

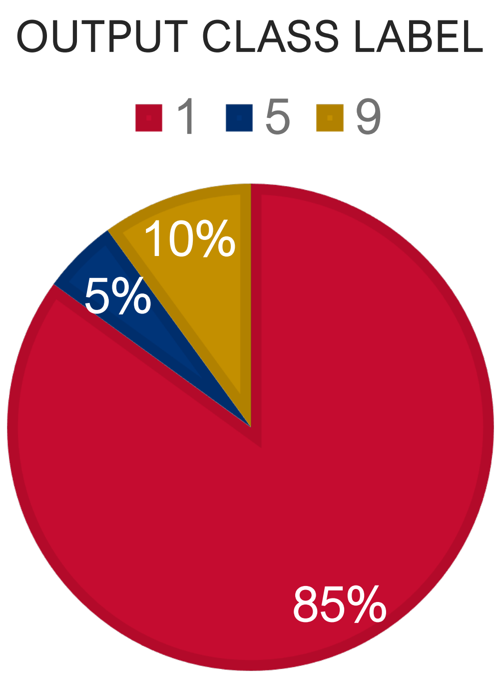


- White-box adversarial attack: C&W attack [S&P '17]

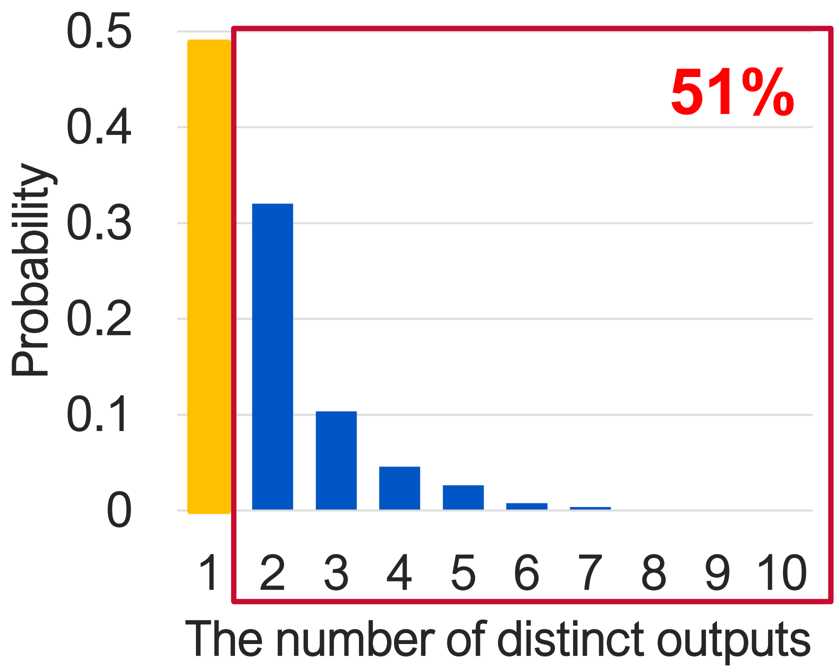


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Challenge 1: Transferability of Adversarial Examples



Distribution of new models' outputs (true label = 1 and target label = 0)



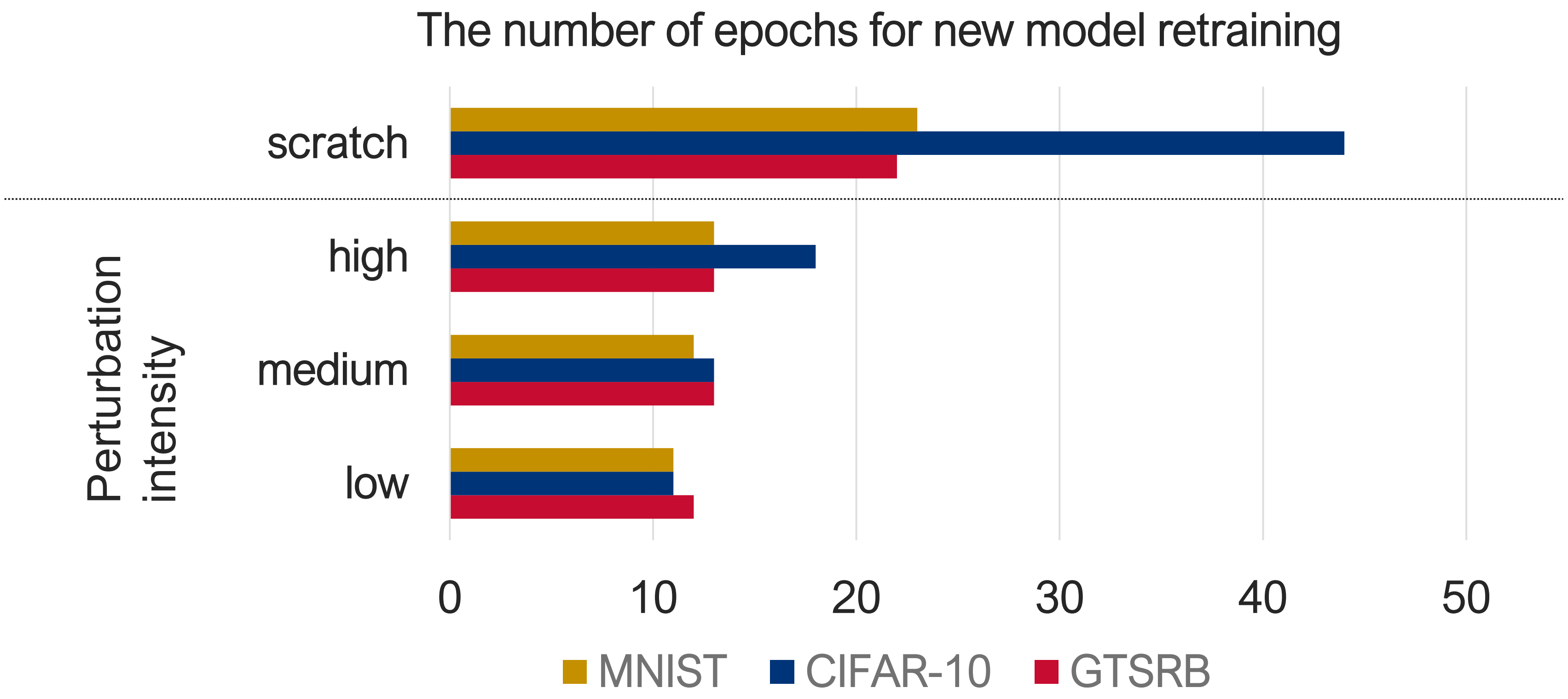
Distribution of the number of distinct outputs

- Attack misleads new models with some probability
- A single new model may not thwart the attack

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Challenge 2: Overhead of *In Situ* Retraining

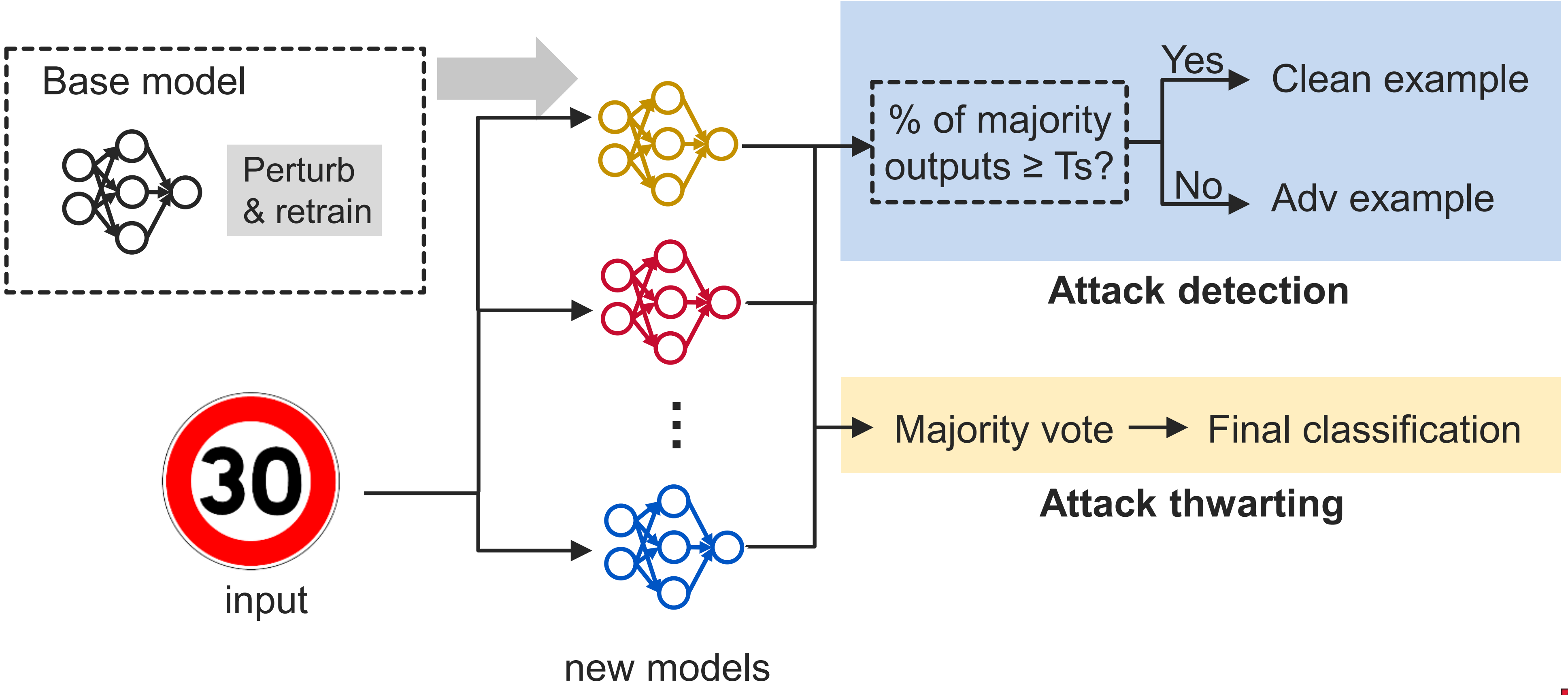
- Retraining new models incurs computation overhead
- Add perturbations to base model and retrain



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

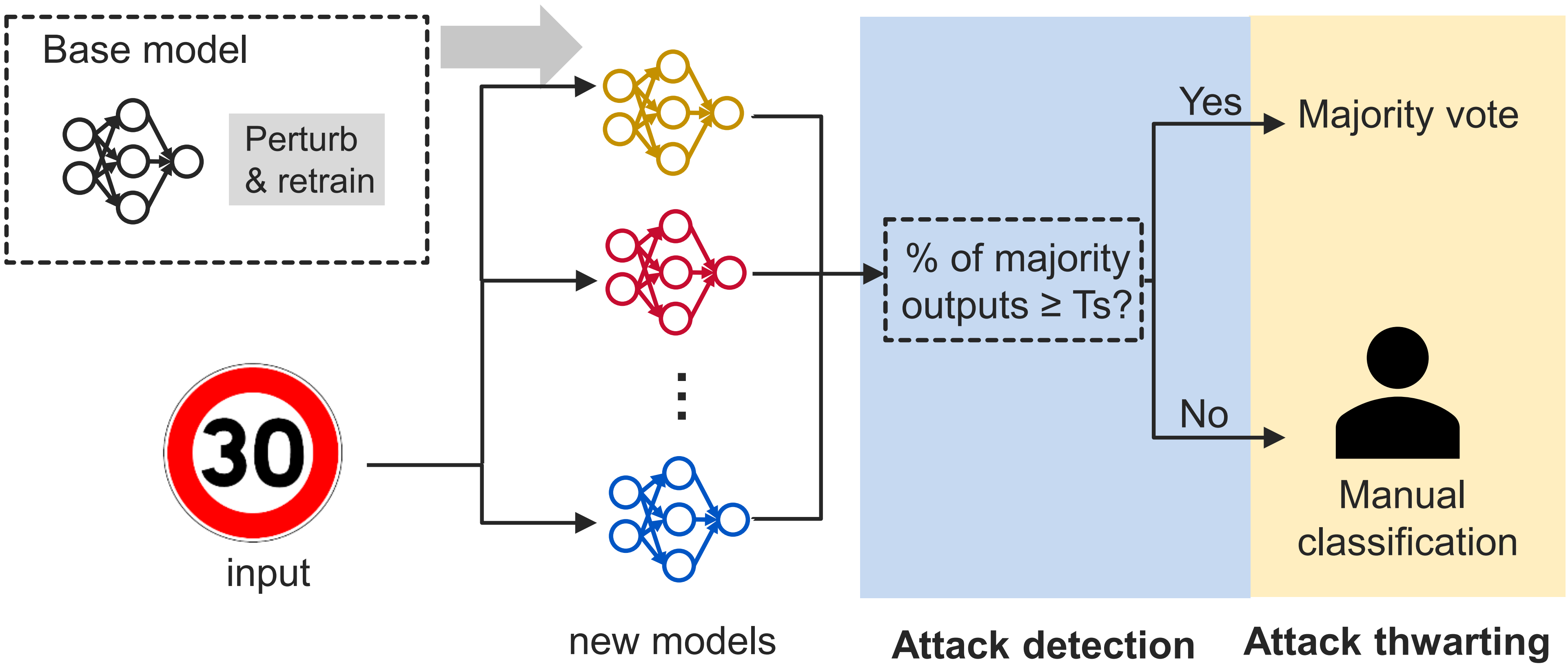
DeepMTD Work Flow

- ***Autonomous***

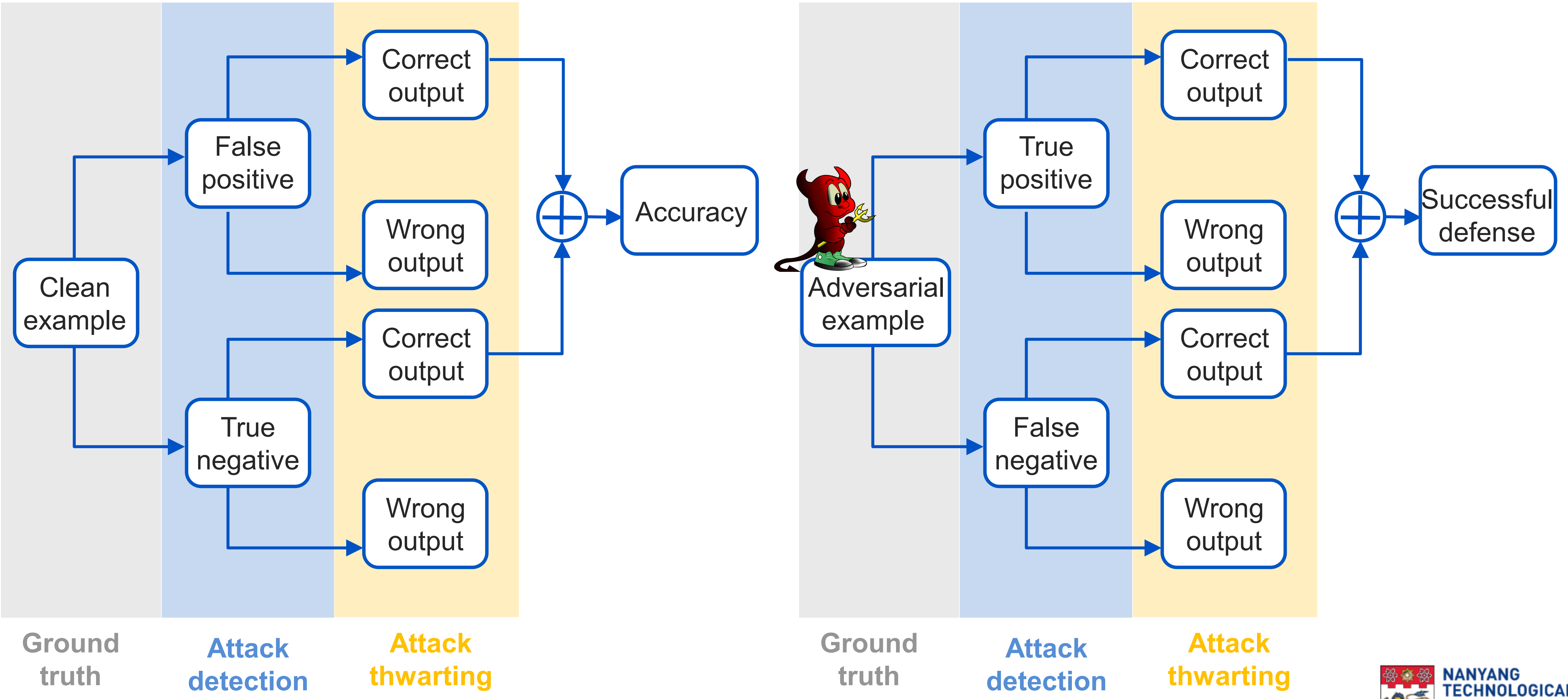


DeepMTD Work Flow

- ***Human-in-the-loop***

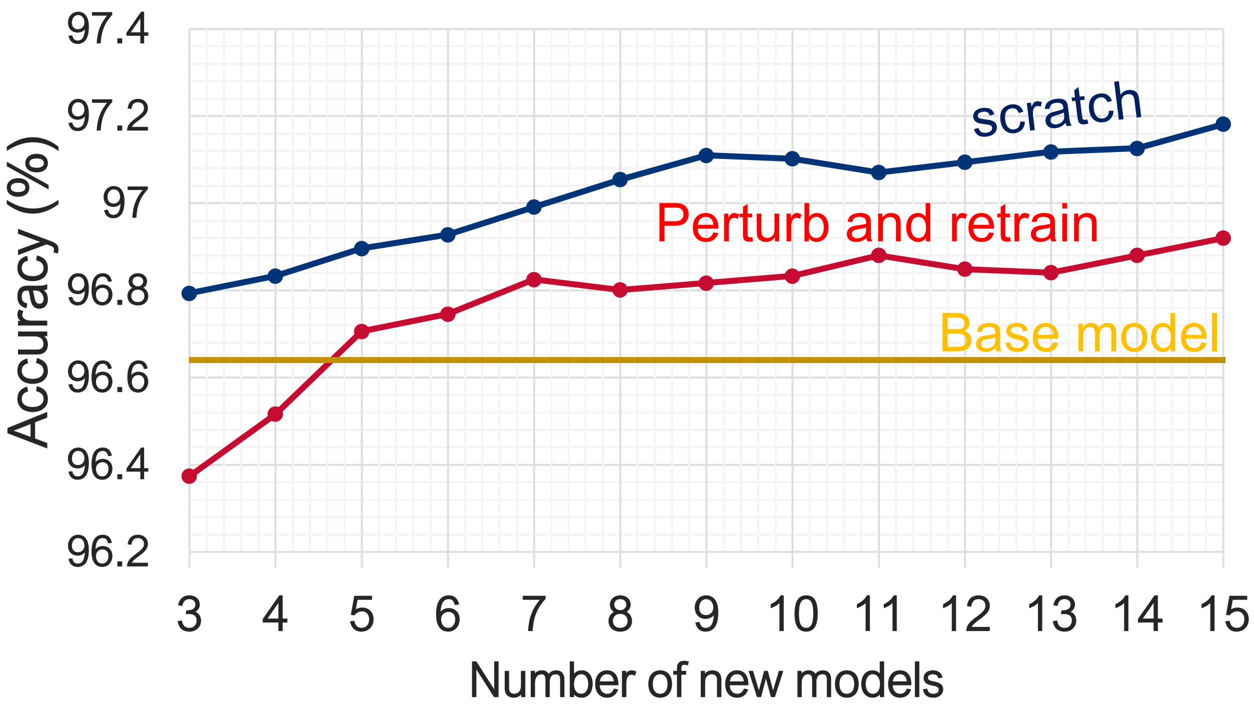


Evaluation Metric



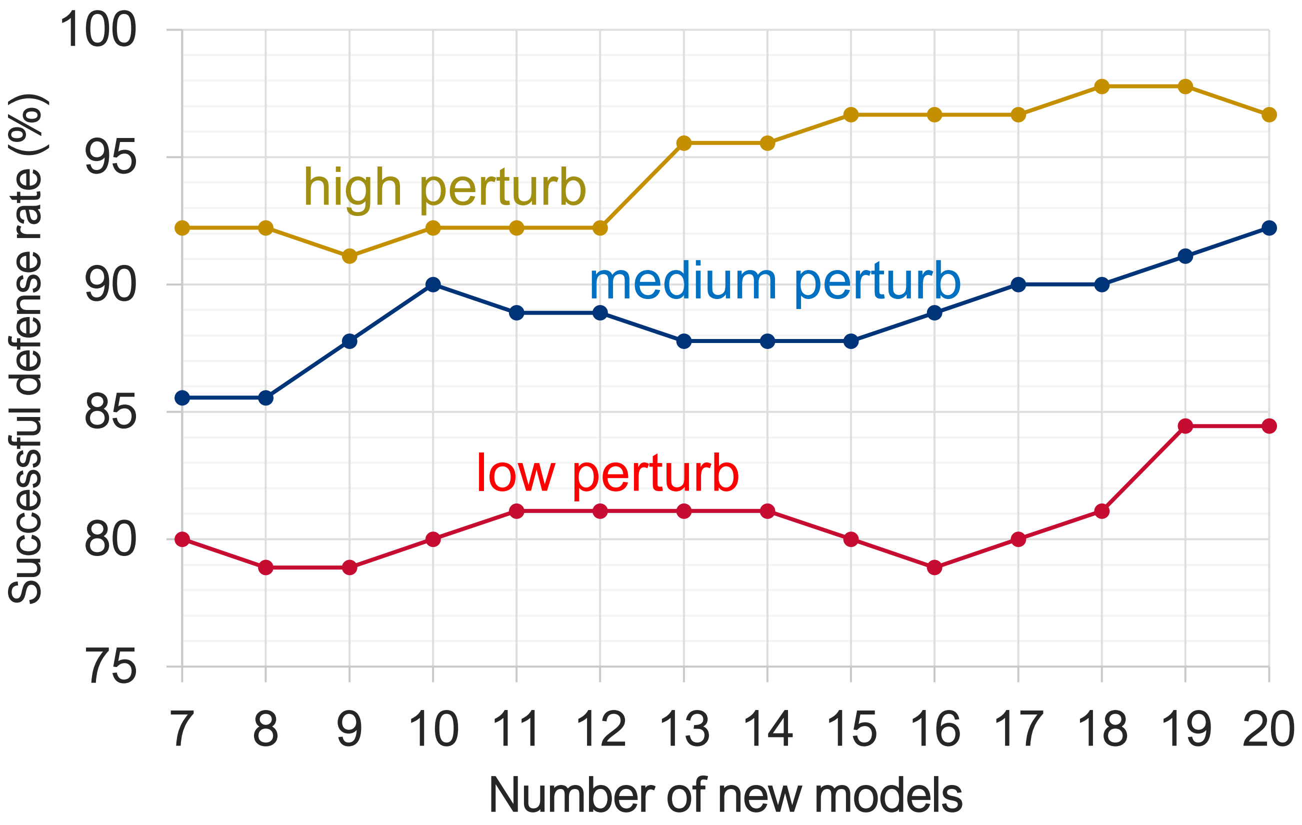
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Accuracy When No Attack (Auto)



- Trade-off btw accuracy & compute overhead
- Improved accuracy on clean examples

Successful Defense Rate (Auto)



- Trade-off btw compute overhead & security

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Human in the Loop

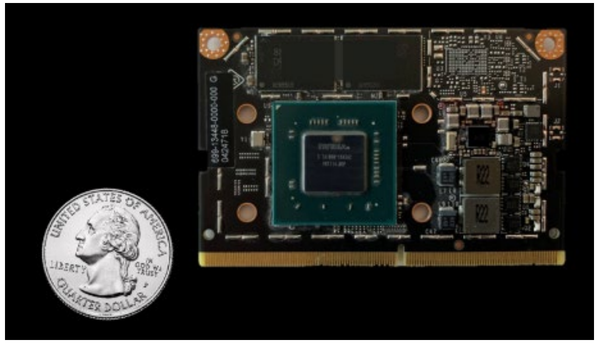
- True positives
  - Human is not affected by adversarial examples
  - Security improved
- False positives
  - Unnecessary overhead to human

Trade-off btw security improvement
& overhead to human

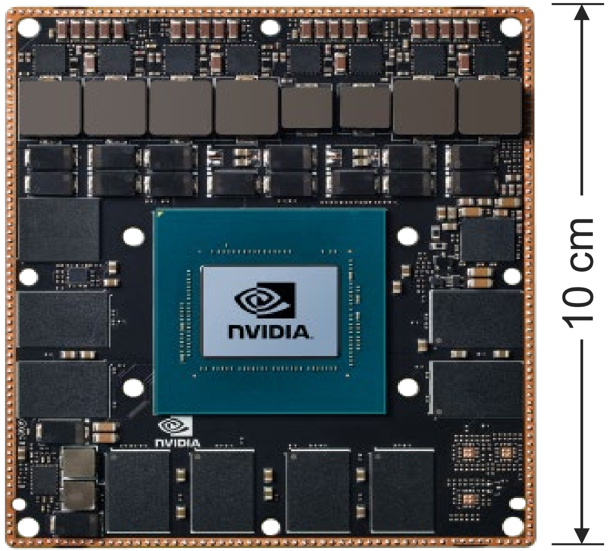
Outline

- Background & Motivation
- Approach Design & Evaluation
- **Implementation**
- Conclusion

Implementation



NVIDIA Jetson Nano
4-core CPU, 128 tensor cores, 4GB mem



NVIDIA Jetson AGX
8-core CPU, 512 tensor cores, 16GB mem

- **Parallel vs. serial DeepMTD**
  - Parallel DeepMTD brings ~20% improvement in inference time

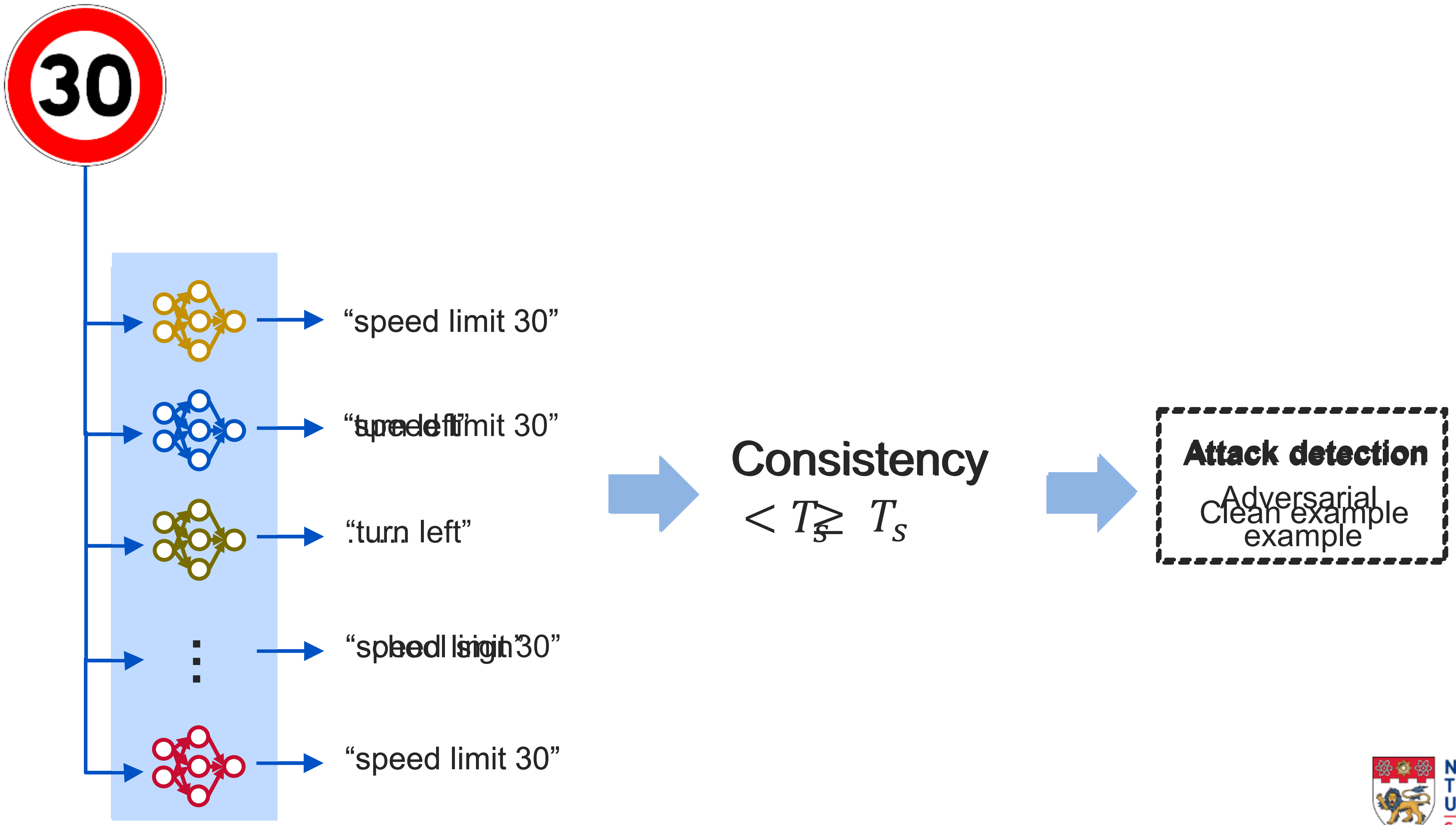
K Keras Parallel mode

0% 0% 0%

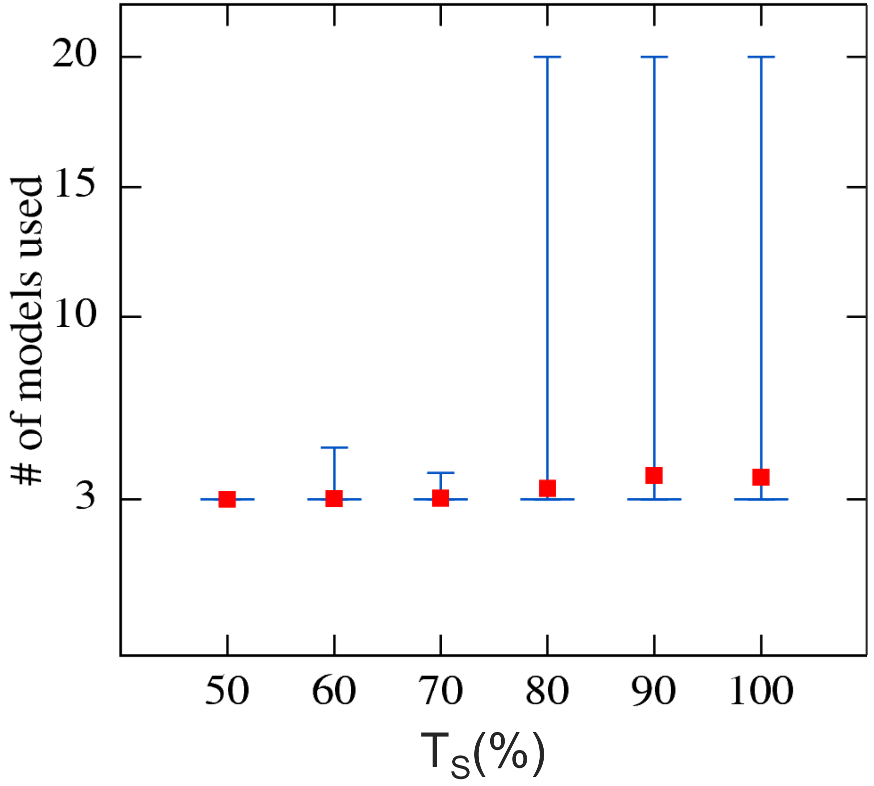
K Keras Serial mode

100% 100% 100%

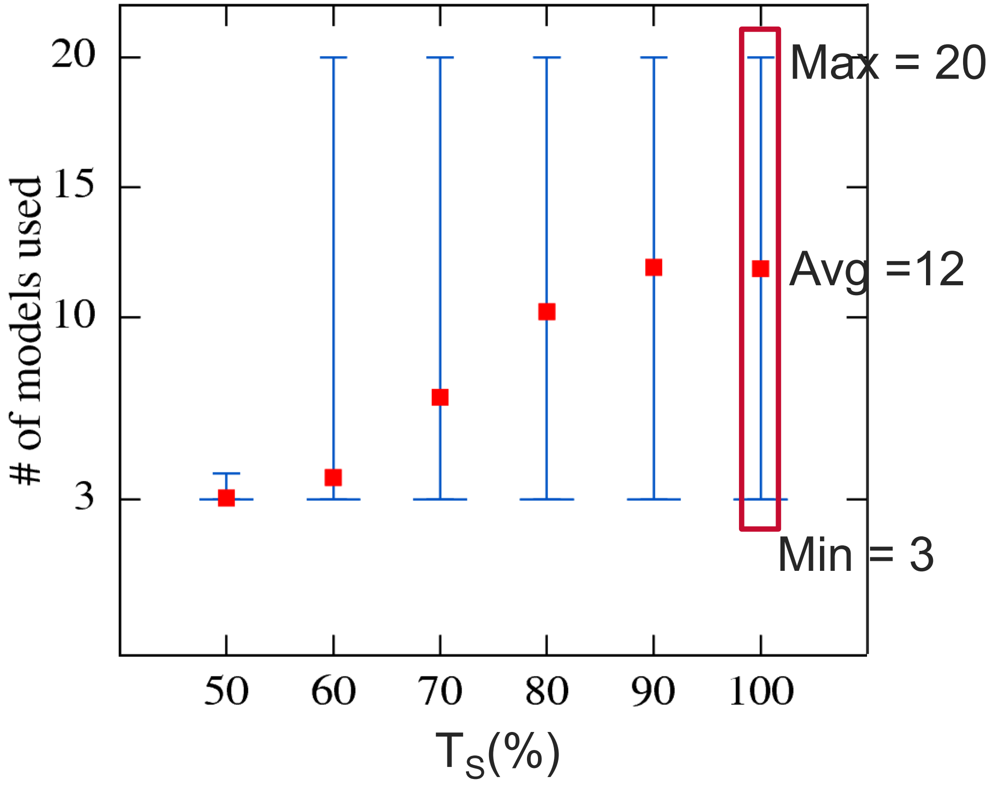
Serial DeepMTD with Early Stopping



Performance of Serial DeepMTD



No attack



With attack

Conclusion

- **DeepMTD** design to counteract adversarial examples
- **DeepMTD** performance evaluation against
  - Clean examples
  - Adversarial examples
- **DeepMTD** serial mode with early stopping
  - Reduces inference time while maintaining sensing performance

More details: Q. Song, Z. Yan, R. Tan, Moving Target Defense for Embedded Deep Visual Sensing against Adversarial Examples, ACM SenSys 2019, New York, USA.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE