AVIATE Safety

PI: Lui Sha

Ph.D. Researchers: Ayoosh Bansal and Simon Yu

System Model

Safe Air + Ground Vehicle (flying taxi) without landing pads. There are three key components

- Verifiably safe collision avoidance using physics model supervised ML vision
- Robust (statistically reliable) traffic sign recognition against noises and attacks.
- Safety-driven Integration
  - Safe perception (Perception Simplex, new)
  - Safe control (L1Simplex, existing work)

=> Synergistic Simplex Architecture

- **Approach:** extend our research on safe autonomous ground vehicles to flying taxi

Guiding Principles

- **For safety critical requirements** such as obstacle detections in air and on land, we shall have explainable and verifiable analytical model for system behaviors. This leads to physics model supervised machine vision for obstacle detection
- **For mission critical requirements** such as traffic sign classification in spite of noises and attacks, we are integrating robust multi-sensor fusion technologies with ML classification, where:
    - Sensors with uncorrelated fail modes will be replaced by features from ML, e.g., shape, color, symbol and text
    - We working on a new ML architecture where feature mis-classifications will be uncorrelated.
    - The first step is the development of an evaluation benchmark.

Perception Simplex

Ayoosh Bansal, Lui Sha

Cyber-Physical Systems Integration Lab

June 16, 2023

Source: Counterpoint Research

Source: NBAA

Current Status

Ground Vehicles

Verifiable Fail-Safe amidst Obstacle Existence Detection Faults

Ongoing work on best-effort Fail-Operational system

Validated using real-world datasets and software-in-the-loop simulation

Ongoing validation on real vehicle

Air Taxis

Ongoing Adaptation from ground to air

Future evaluation using PhotoRealistic Air Vehicle Simulator + NASA GUAM

References

Perception Simplex:
Verifiable Collision Avoidance in Autonomous Vehicles amidst Obstacle Detection Faults.
Ayoosh Bansal, Hunmin Kim, Simon Yu, Bo Li, Naira Hovakimyan, Marco Caccamo, and Lui Sha.
Software Testing, Verification & Reliability. (Under Review)

Verifiable Obstacle Detection.
Ayoosh Bansal, Hunmin Kim, Simon Yu, Bo Li, Naira Hovakimyan, Marco Caccamo, and Lui Sha.
In 2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE), 2022.

Risk ranked recall: Collision safety metric for object detection systems in autonomous vehicles.
Ayoosh Bansal, Jayati Singh, Micaela Verucchi, Marco Caccamo, and Lui Sha.
In 2021 10th Mediterranean Conference on Embedded Computing (MECO), IEEE, 2021.

System Model and Scope

Generalized Air + Ground Vehicle

VTOL is a subset

Full operation in cluttered environments

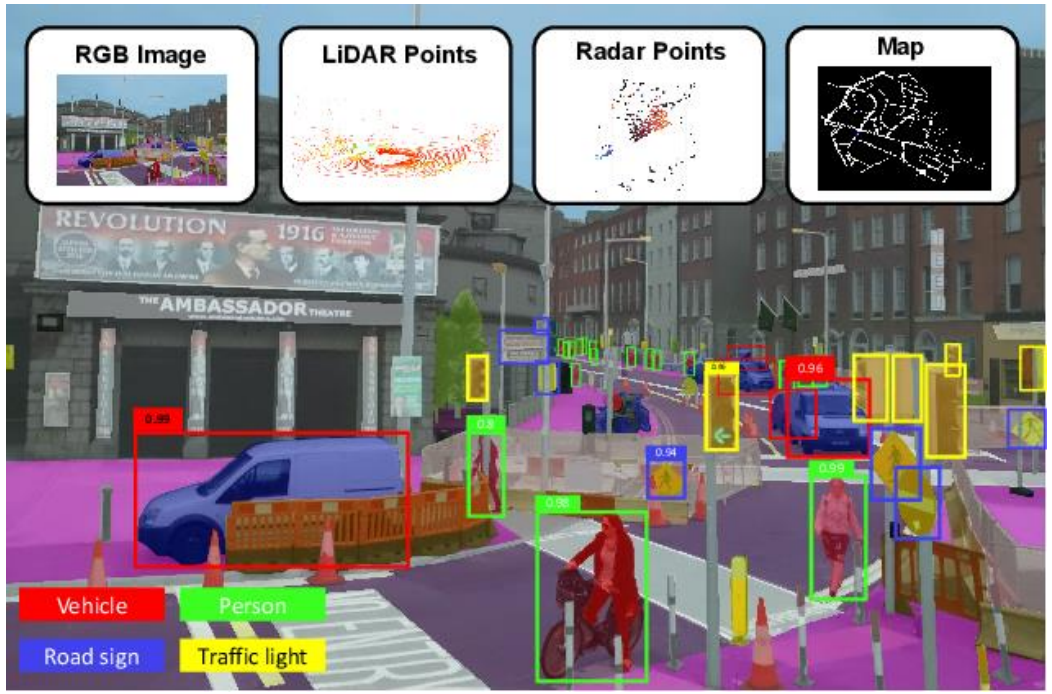
Physical System, Control, Planning, Perception

Mission / Safety

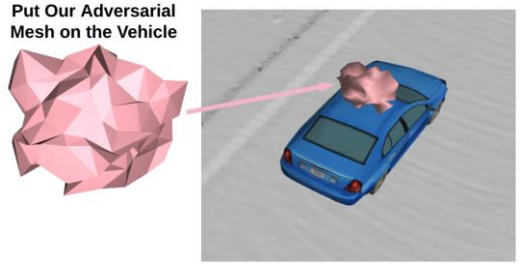
Source: information-age

Deep Learning

Capabilities



Safety

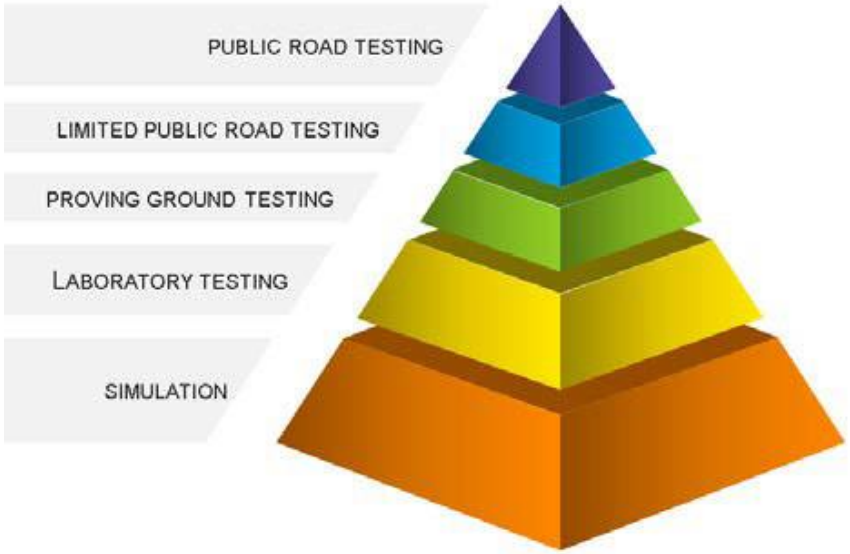


Perception & prediction present a uniquely difficult assurance challenge

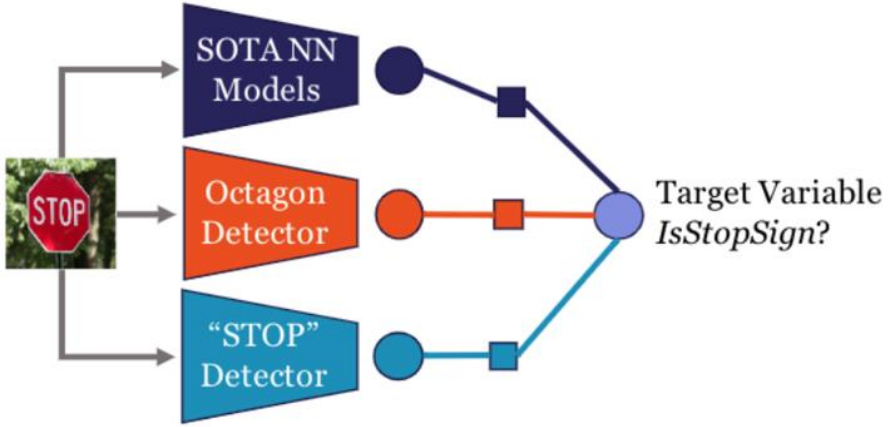
2017. Autonomous vehicle safety: An interdisciplinary challenge. IEEE Intelligent Transportation Systems Magazine.
2019. Why deep-learning AIs are so easy to fool. Nature.
2020. Physically realizable adversarial examples for lidar object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*
2020. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems.*
2022. Explainable deep learning: A field guide for the uninitiated. Journal of Artificial Intelligence Research.
2022. Autonomous Vehicle Standards & Open Challenges. P. Koopman.
2023. Dense reinforcement learning for safety validation of autonomous vehicles. Nature.

Approach I – Fault Prevention

Testing



Robust ML



2016. Challenges in autonomous vehicle testing and validation. SAE International Journal of Transportation Safety.
2018. Development of a test track for driverless cars: vehicle design, track configuration, and liability considerations. Periodica Polytechnica Transportation Engineering.
2021. Knowledge enhanced machine learning pipeline against diverse adversarial attacks. In International Conference on Machine Learning.

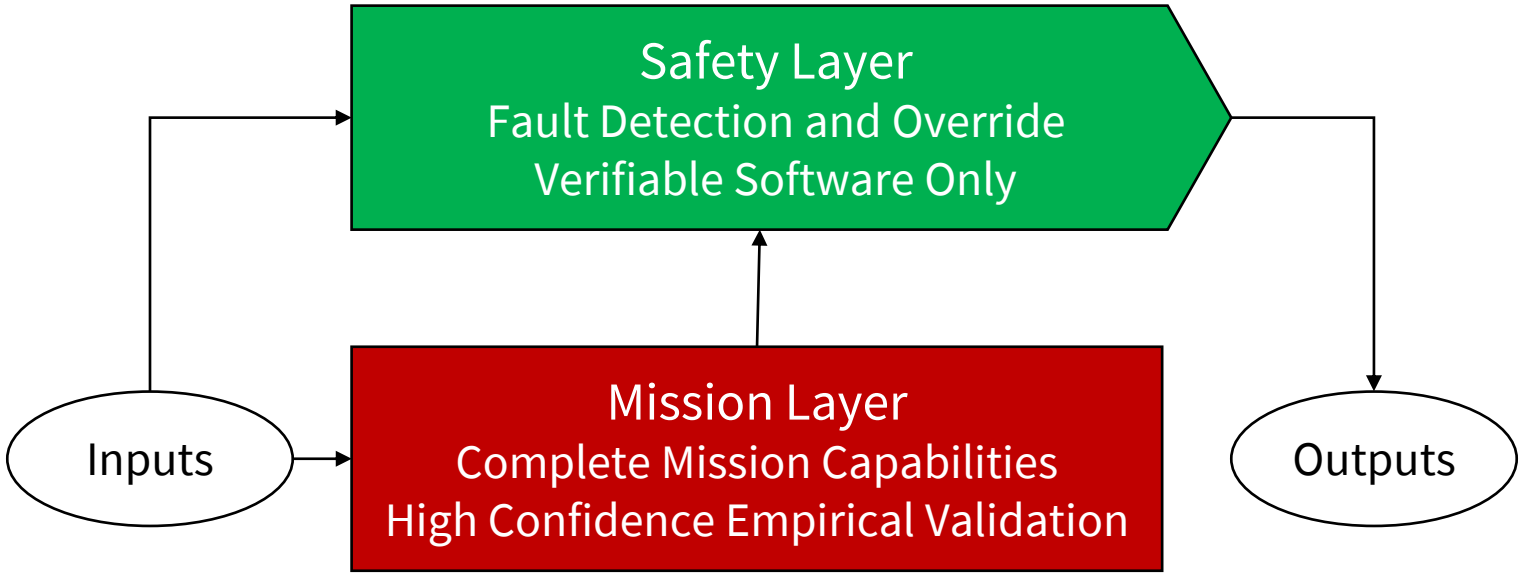
Approach II – Fault Tolerance

Goals

Selective Safety Guarantees

Utilize ML capabilities

Minimize Performance Loss



Procedure

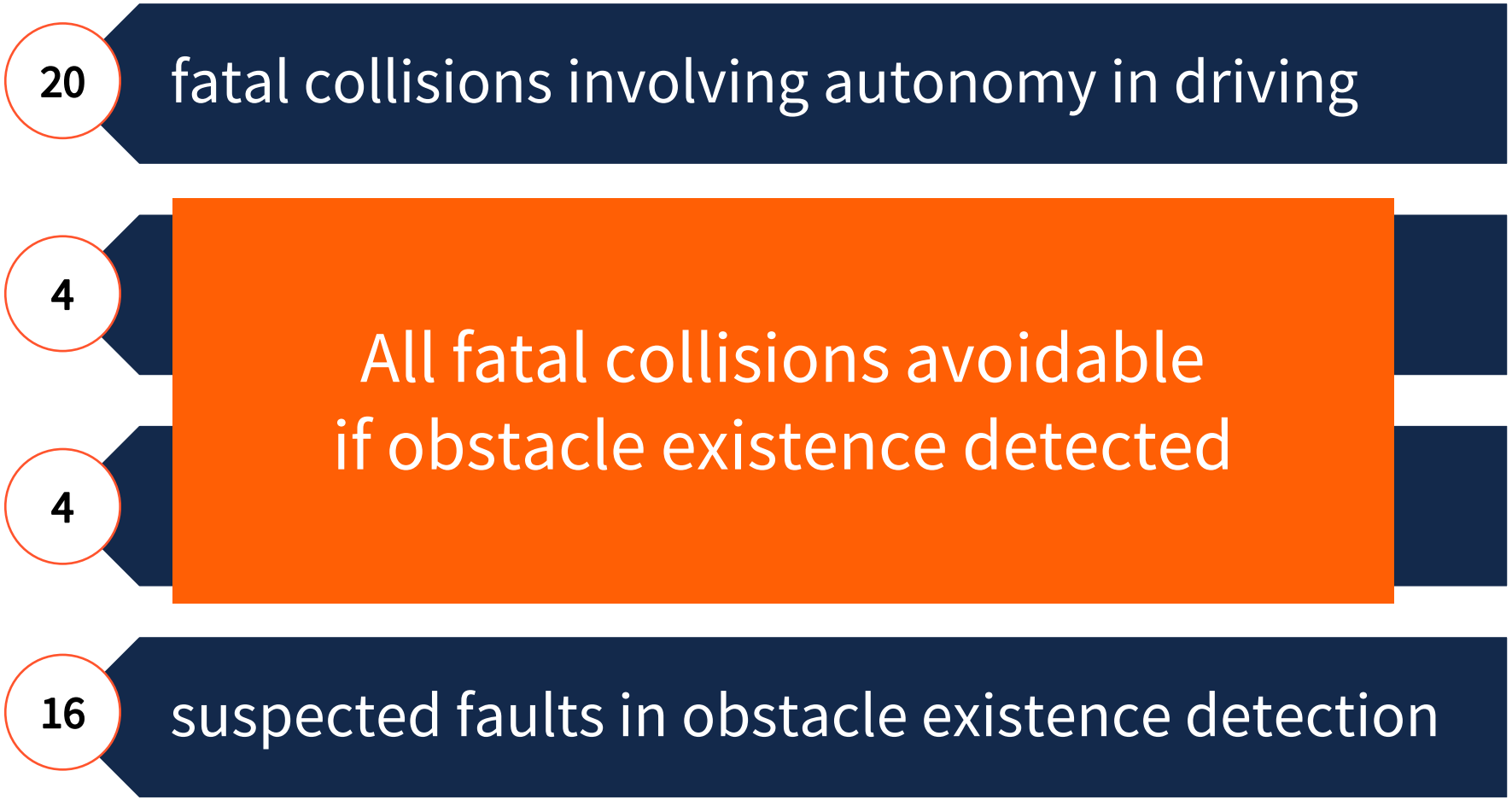
Faults

Selective

Survey of Collisions involving Autonomous Ground Vehicles

- 20 fatal collisions involving autonomy in driving
- 4 analyzed by NTSB, US Gov.
- 4 involved faults in obstacle existence detection
- 16 suspected faults in obstacle existence detection

Survey of Collisions involving Autonomous Ground Vehicles



Procedure

Faults

Requirements

Selective

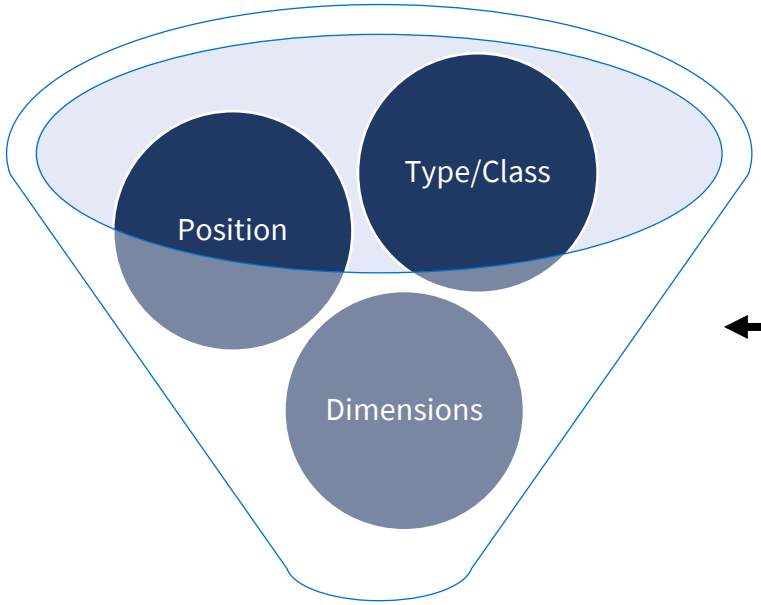
Minimal

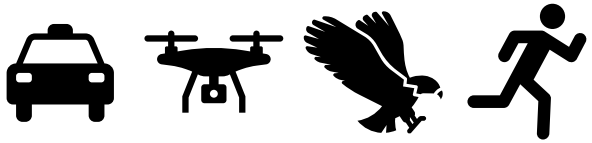
Object vs Obstacle Detection

Complex – Requires DNN

Simpler – Geometric Algorithms Suffice

✓ Explainable, Analyzable, Verifiable





Not required

Mission-Critical
Object Detection

Safety-Critical
Obstacle Detection

Minimal Obstacle Detection Requirements for Safety Critical Obstacle Avoidance

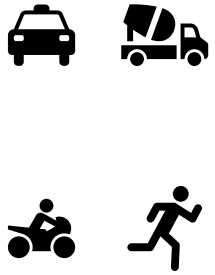
Class ✗ — Object class determination useful but not necessary

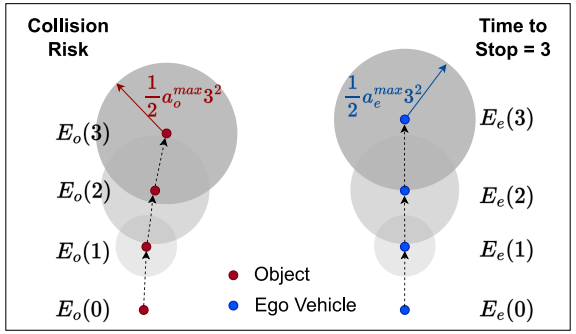
Risk ✗ — Only need to detect obstacles that pose a collision risk

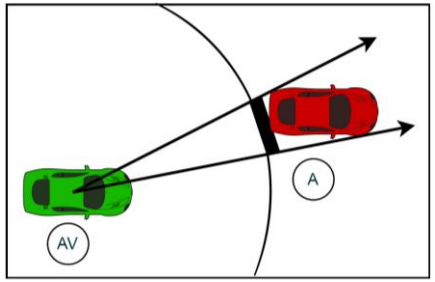
Height ✗ ✓ — Only detect if height within a collision range

Distance ✓ — Underestimation ok, overestimation bounded

Projection ✓ — Need to detect a projection towards the AV







$$r = \sum_{j=0}^{N} (\Pi_{t=0}^{j-1} \mathbf{1}_{X \setminus (\cup_{i \in M_O} O_i)}(x_t)) \mathbf{1}_D(x_j)$$

$$D_{Min}^{Detected} \leq 0.1 + 0.05 * D_{Min}^{GT}$$

2022. Verifiable Obstacle Detection. IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE).
2021. Risk Ranked Recall: Collision Safety Metric for Object Detection Systems in Autonomous Vehicles. 2021 10th Mediterranean Conference on Embedded Computing.

Procedure

Faults

Requirements

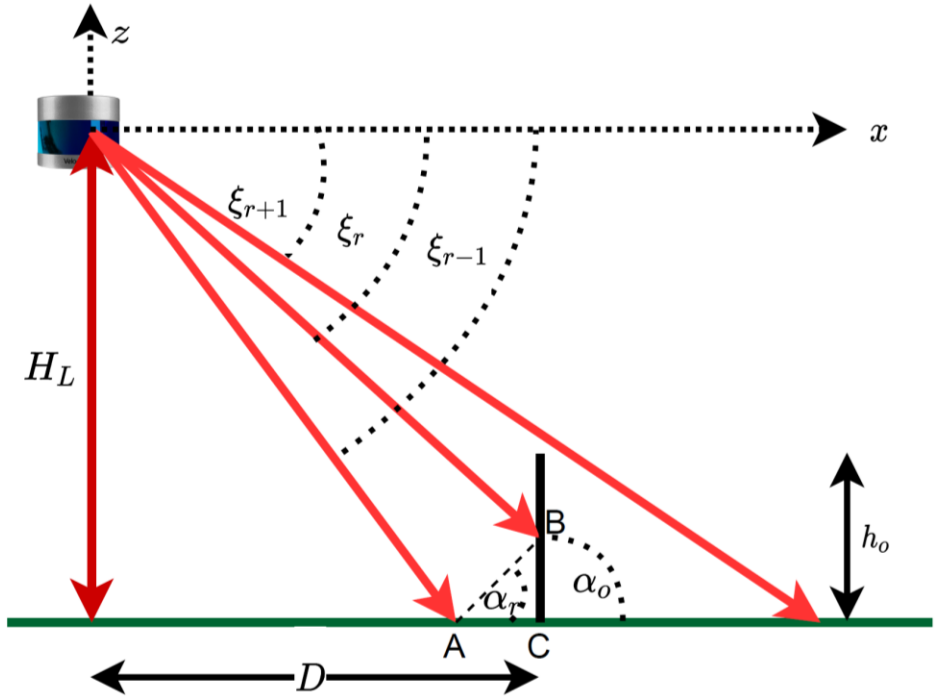
Algorithms

Selective

Minimal

Verifiable

Obstacle Detection Algorithm



Depth Clustering Algorithm
Ground Removal

$$\Delta\alpha_{r,c} = \left\{ \begin{array}{l} 0^o, \text{if } r = 1 \\ |\alpha_{r,c} - \alpha_{r-1,c}|, \text{otherwise} \end{array} \right\}$$

Obstacle detected if

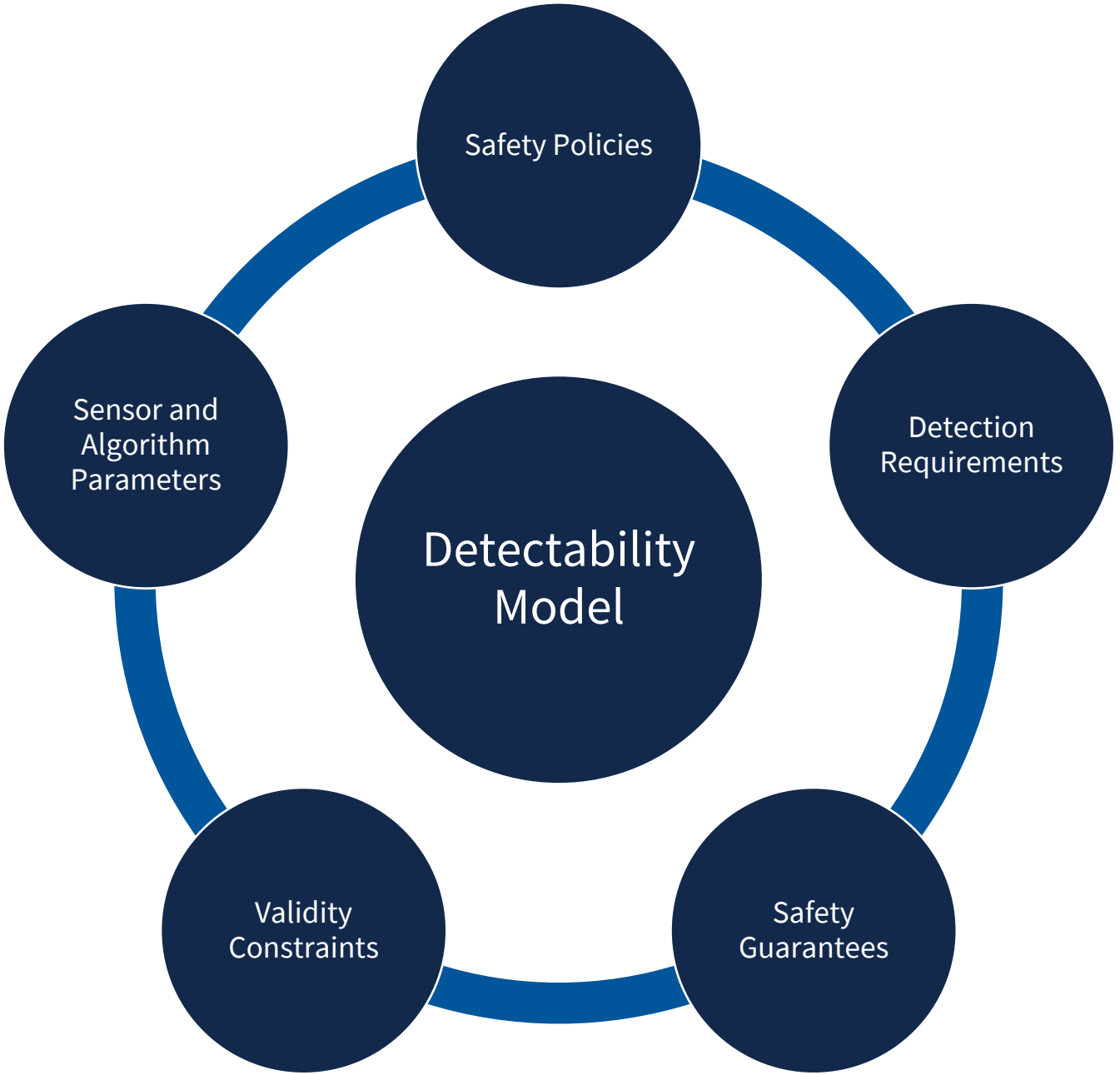
$$\Delta\alpha_r > \alpha_{th}$$

2016. Fast range image-based segmentation of sparse 3D laser scans for online operation. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).
2017. Efficient online segmentation for sparse 3D laser scans. PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science.

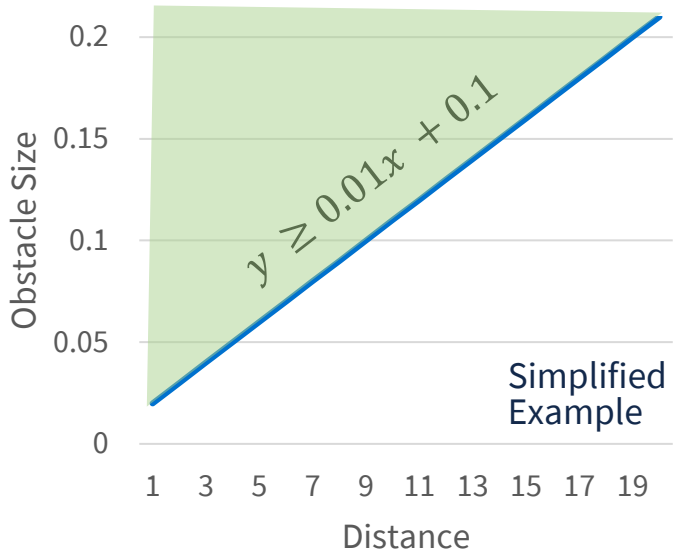
DNN *vs* Classical

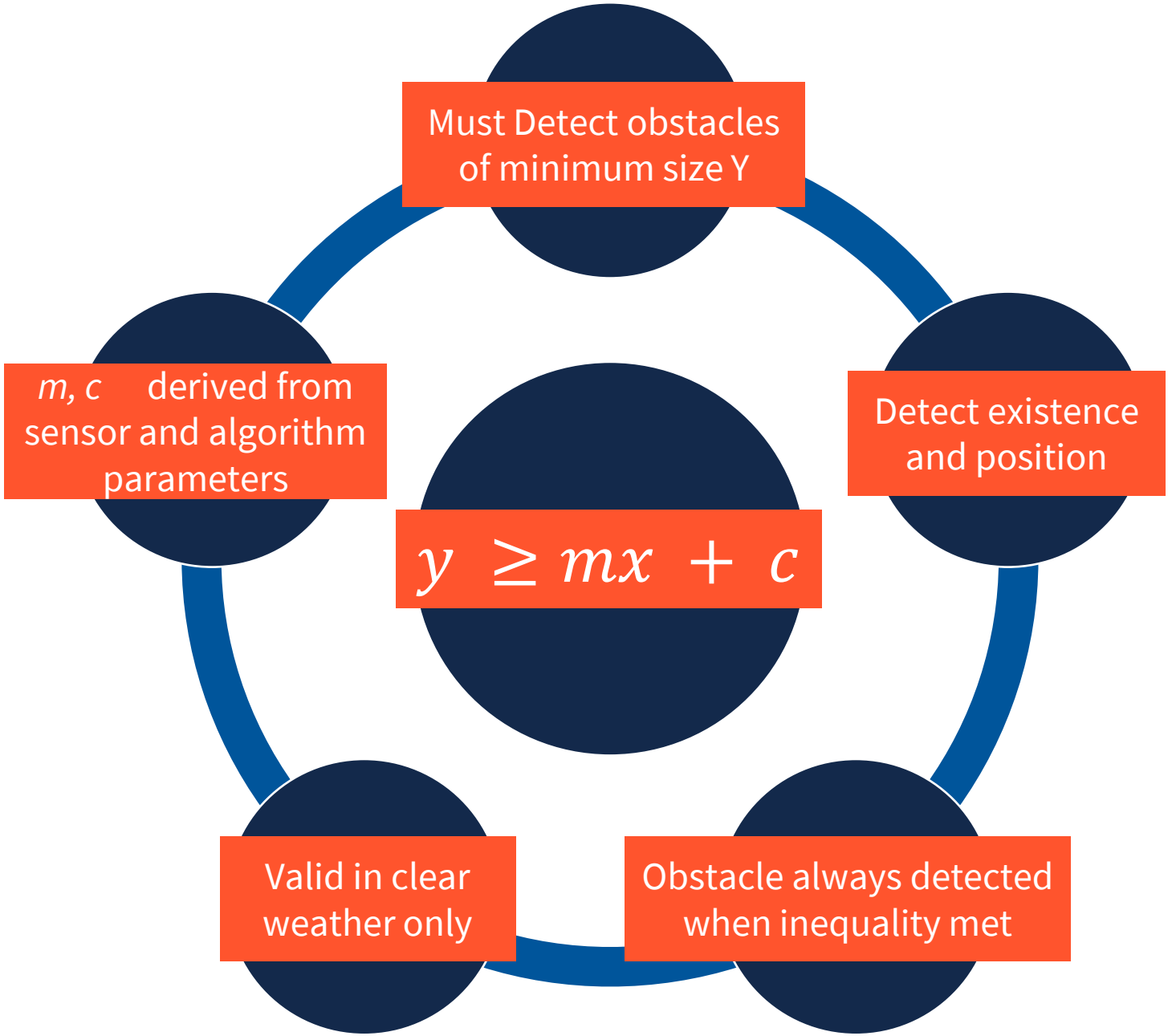
DNN	Classical
Learned Correlations	Causal Logic
Higher Capabilities and Performance	Limited Specific Capabilities
Empirical High Confidence Validation	Logical Analysis and Verification
Limited Feature Comprehensibility	Human Comprehensible Features and Limitations
Incomprehensible Features Inhibit Policy Support	Policy Support

Detectability Model

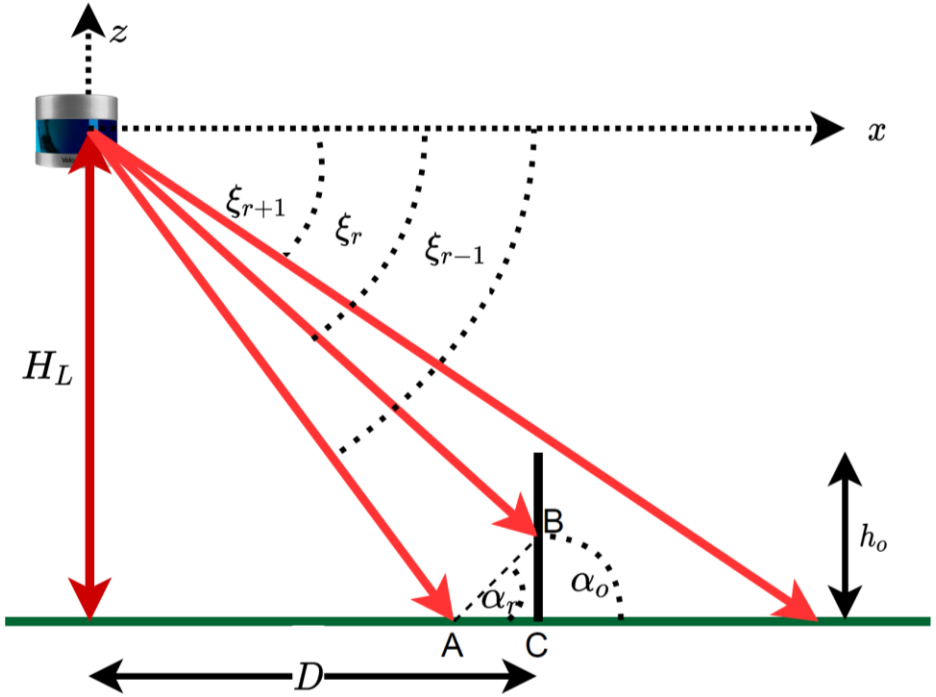


Detectability Model



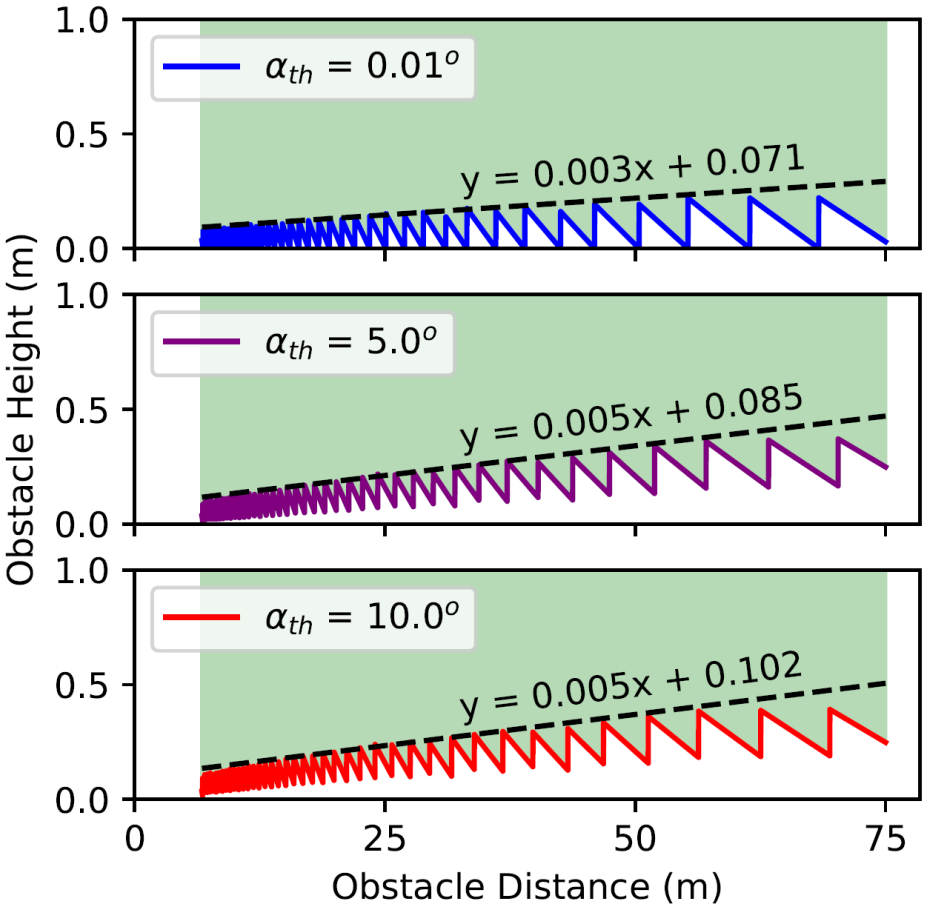


Detectability Model



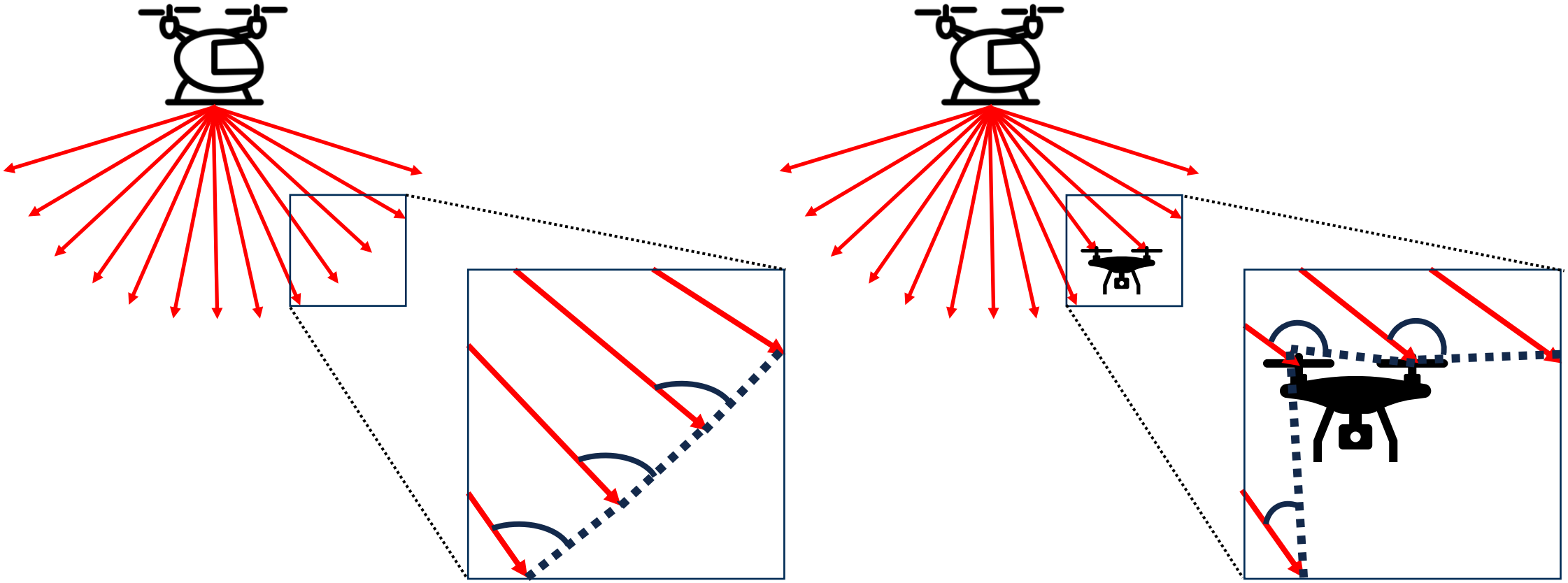
Theorem The obstacle is detected at distance D, if and only if one of the following conditions are true:

1) $H_r(D) \leq h_o < H_{r+1}(D)$ AND
   $\alpha_{th} < atan2(H_r(D), |D - \frac{H_L}{tan(\xi_{r-1})}|)$;
2) $h_o \geq H_{r+1}(D)$.

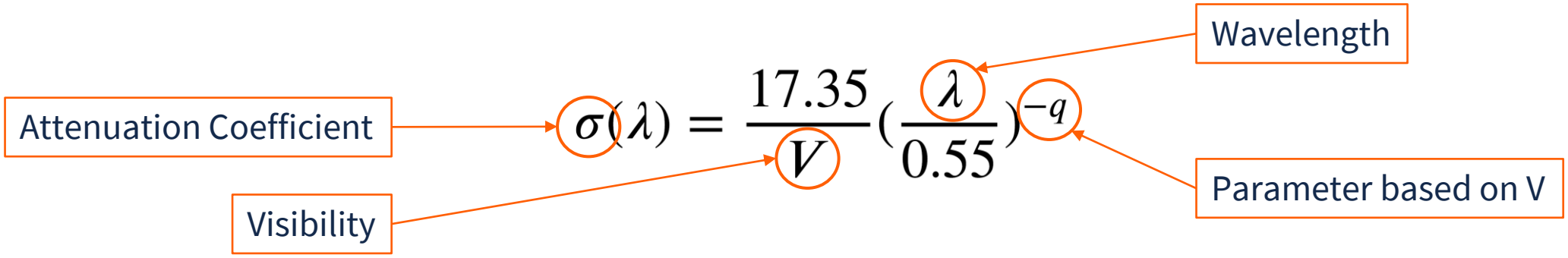


Source: Flaticon.com

Obstacle Detection Algorithm - Air



Weather Effect



$$R_{max}^{LiDAR} = \frac{\sigma_{clear}}{\sigma_{current}} R_{max,clear}^{LiDAR}$$

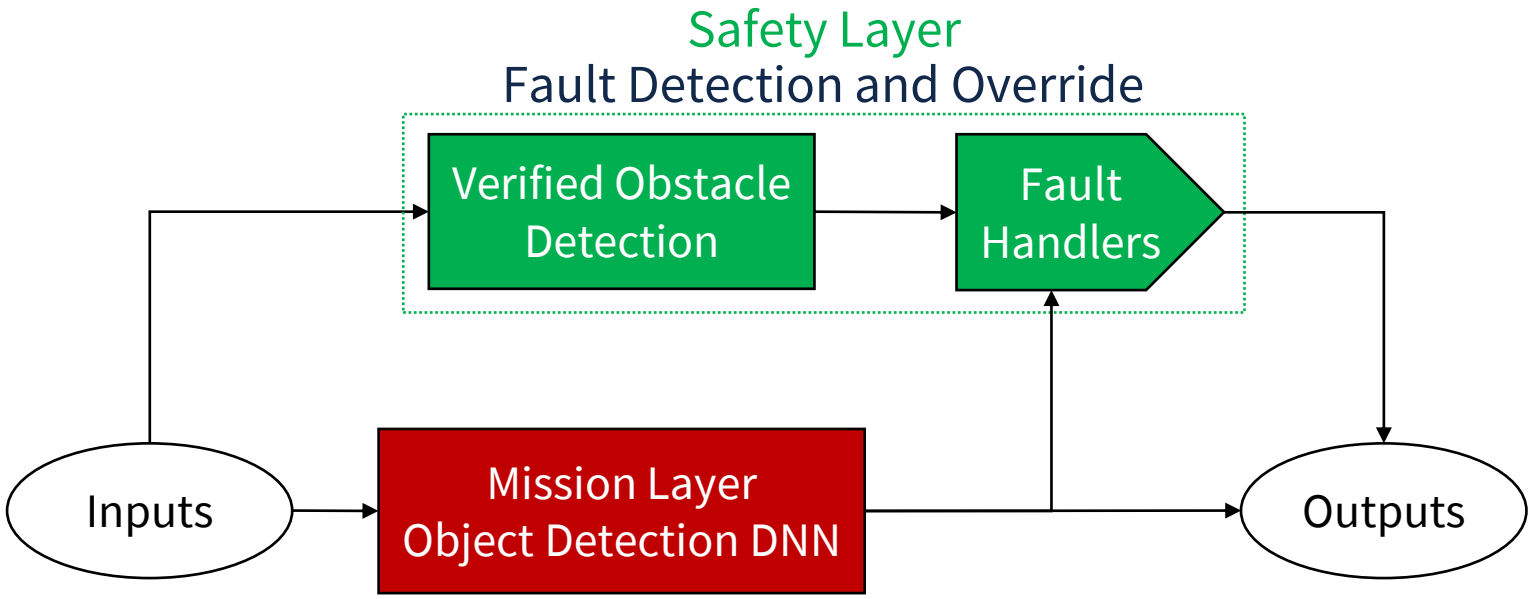
Reduction: 10x Haze, 100x Dense Fog

1990. Laser beam propagation in the atmosphere. SPIE press.
2001. Comparison of laser beam propagation at 785 nm and 1550 nm in fog and haze for optical wireless communications. In Optical wireless communications III.
2008. Radiometric calibration of LIDAR intensity with commercially available reference targets. IEEE-Transactions on Geoscience and Remote Sensing.
2020. Laser and LIDAR in a system for visibility distance estimation in fog conditions. Sensors.
2021. Visibility enhancement and fog detection: Solutions presented in recent scientific papers with potential for application to mobile systems. Sensors.

Procedure

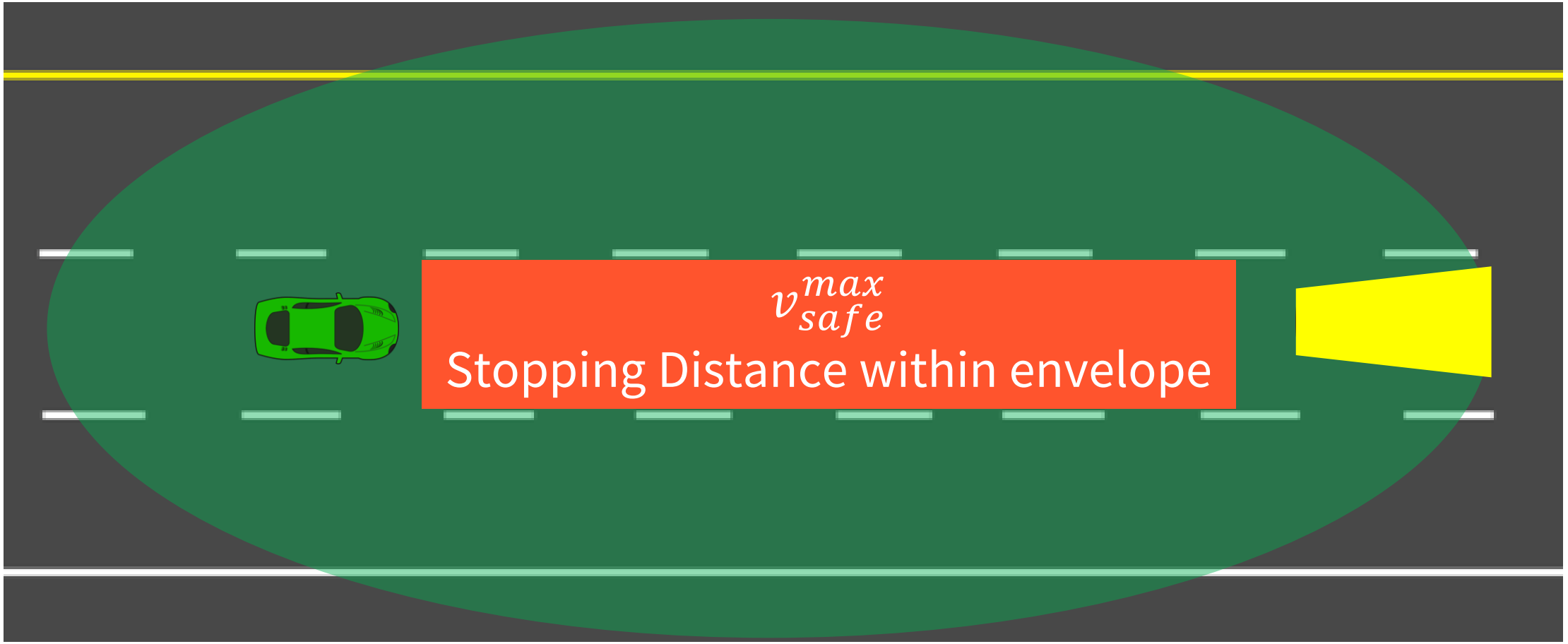
Faults	Requirements	Algorithms	Integration
Selective	Minimal	Verifiable	Synergy

Perception Simplex



Responsible for Mission (Navigation) while trying to maintain Safety (Collision Avoidance)

Safety Monitoring Envelope



Mission != Safety ? Override : Commands Passthrough

Velocity Limit

Detectability Limit

Lidar Range Limit

$$R_{max} = min(R^{O}_{max}, R^{LiDAR}_{max})$$

Detection Range

Safety Margin

$$D^{stop}_{max} = R_{max} - D^{Safety}$$

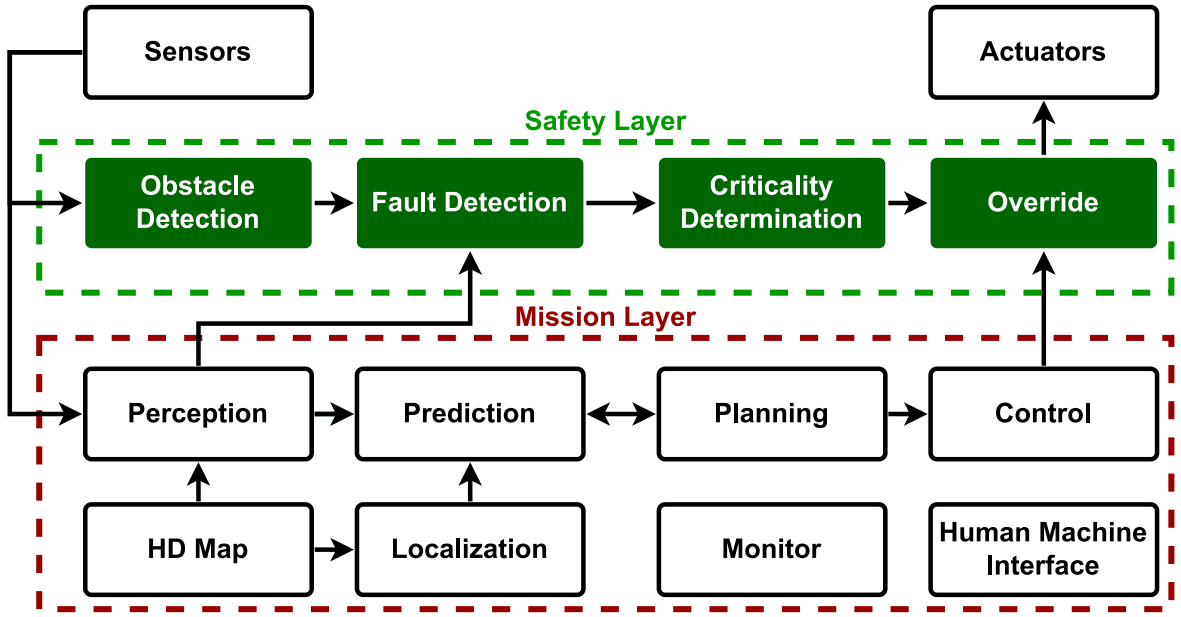
Reaction Distance

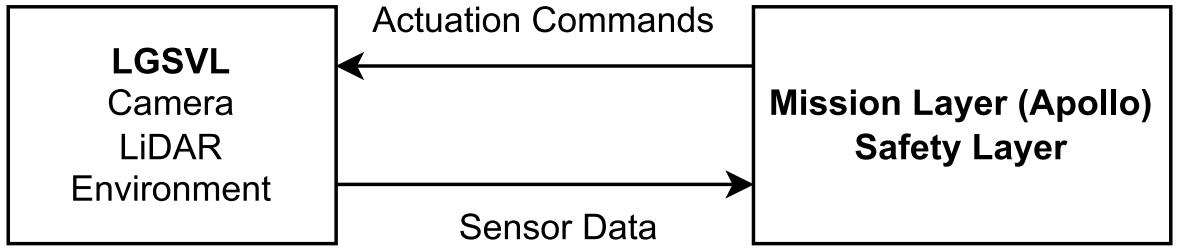
$$v^{safe}_{max} = \sqrt{(a^{av}_{max} L_{max})^2 + 2a^{av}_{max} D^{stop}_{max}} - a^{av}_{max} L_{max}$$

Max Deceleration

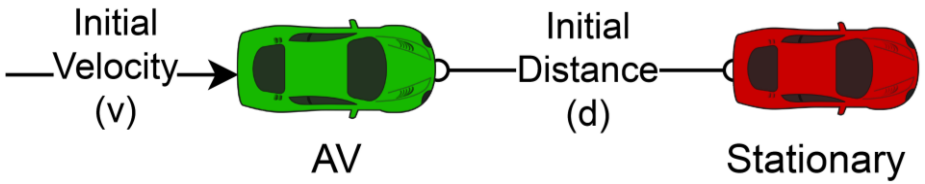
Computational Latency

Perception Simplex Implementation



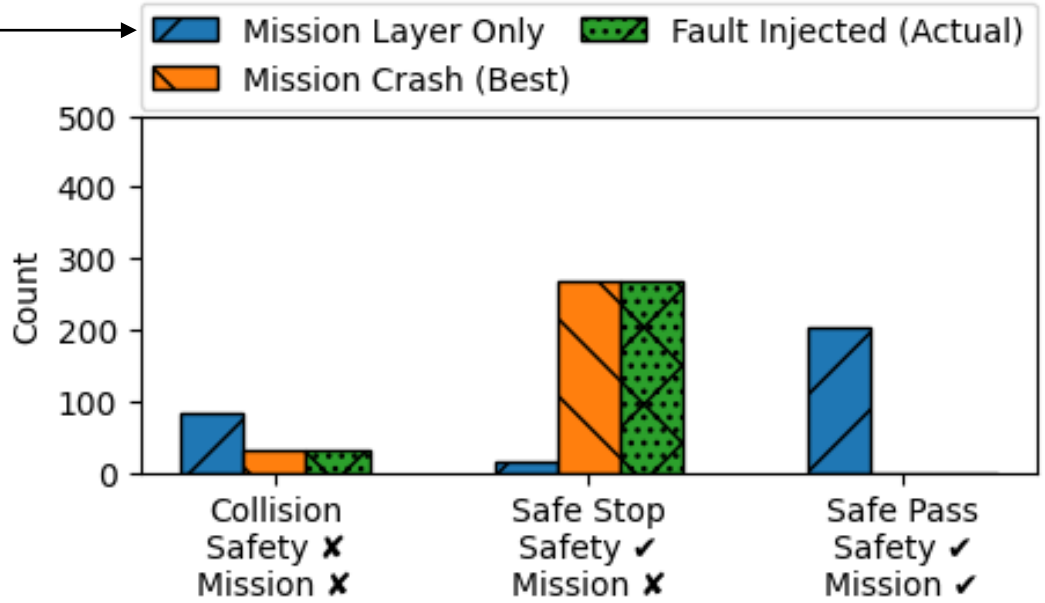


Perception Simplex Evaluation - Safety



Requirement	Collision	Safe Stop	Safe Pass
Safety	Violated ✗	Met ✓	Met ✓
Mission	Violated ✗	Abandoned ✗	Met ✓

Open-Source AV



Actual Response
==
Best Response

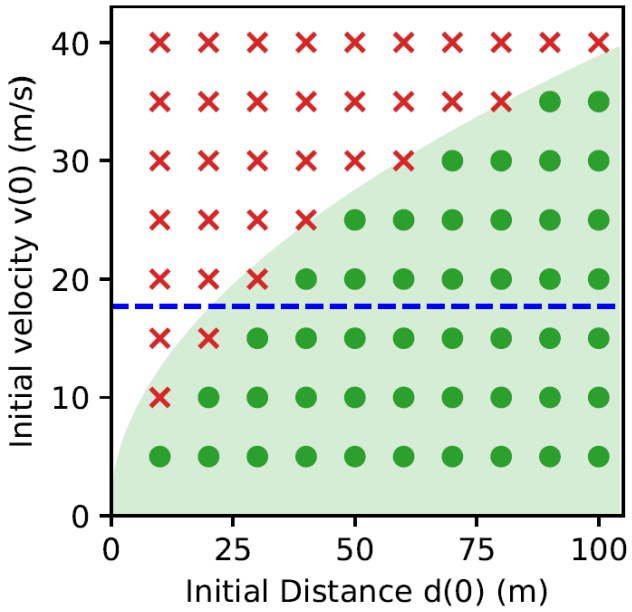
$$v \leq v_{safe}^{max}$$

Perception Simplex Evaluation - Safety

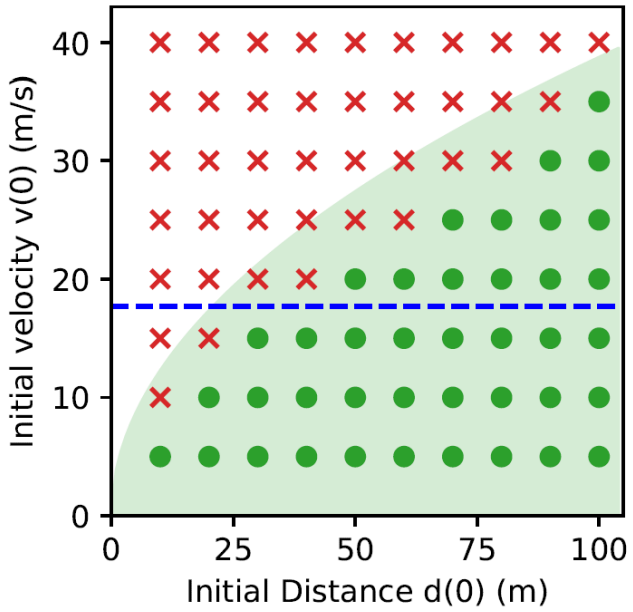
Safe Stop

Collision

v_{safe}^{max}

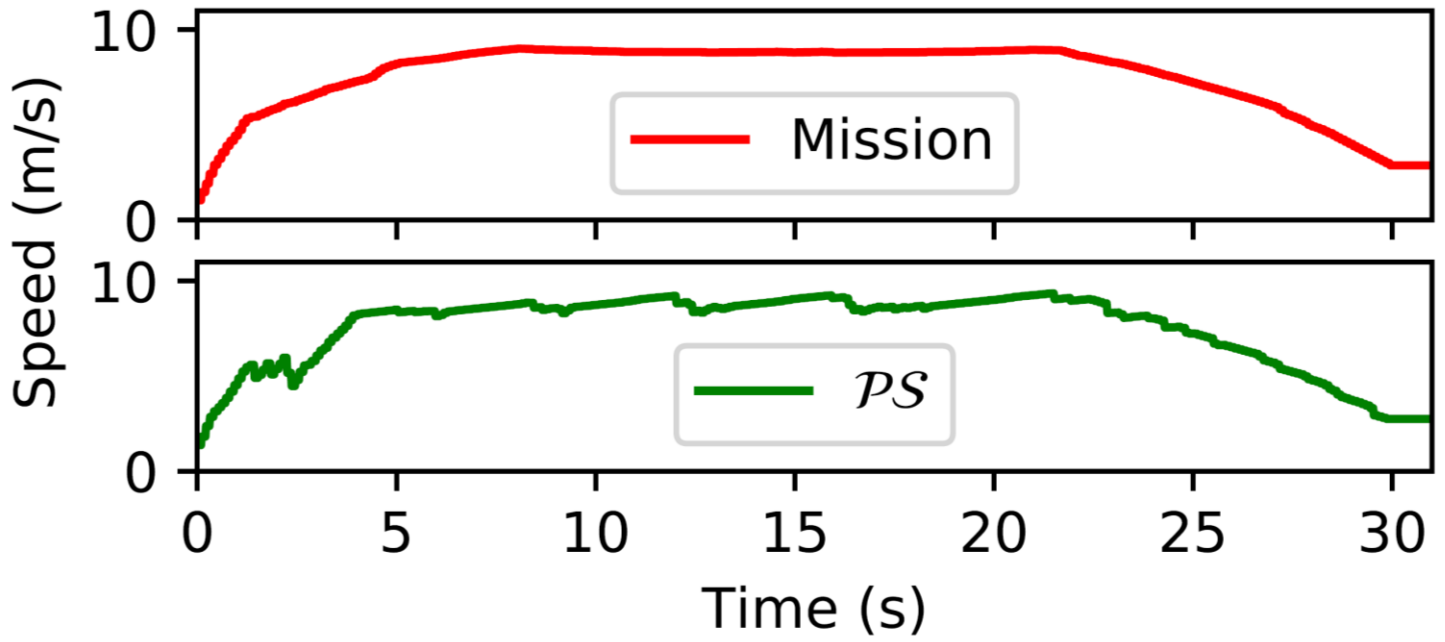


Mission Crash (Best)



Fault Injected (Actual)

Perception Simplex Evaluation - Performance

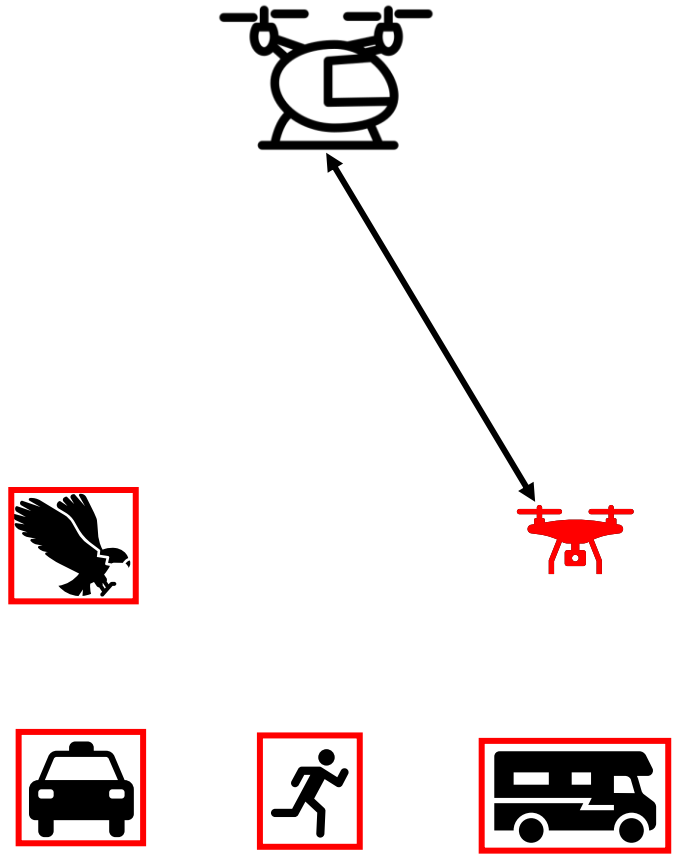


Source: Flaticon.com

Perception Simplex – Air Taxi

Goals

Safety Guarantees ✓



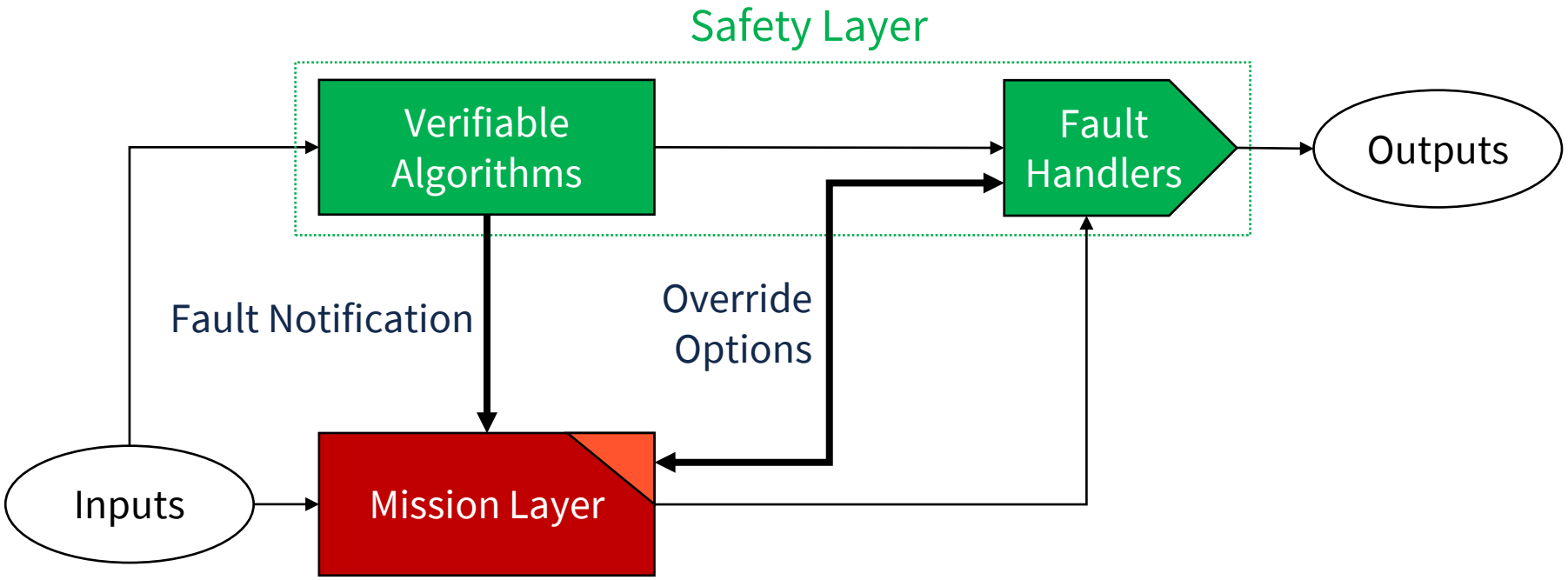
Override

Stop in place

External Support

Verifiable Fail-Safe Fail-Operational?

Synergistic Simplex – Across Layers



Better Fault Responses

Goals

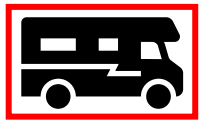
Safety Guarantees ✓

Utilize ML capabilities ✓

Responses

Fault Notification

Re-Plan to Evade



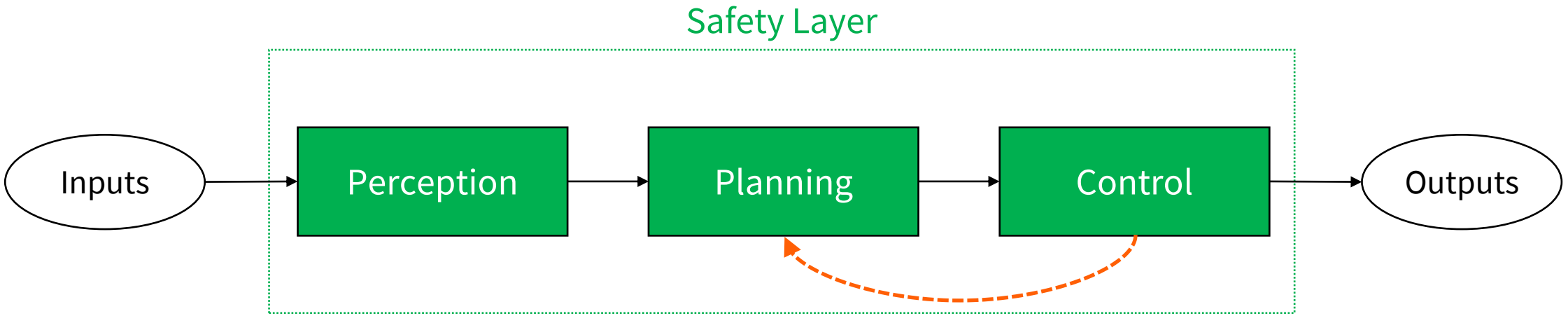
2014. Optical-flow based strategies for landing vtol uavs in cluttered environments. IFAC Proceedings Volumes.
2021. Enhanced potential field-based collision avoidance in cluttered three-dimensional urban environments. Applied Sciences.

Synergistic Simplex – Across Components

Control → Planning

Dynamic Confirmation

Control Capabilities

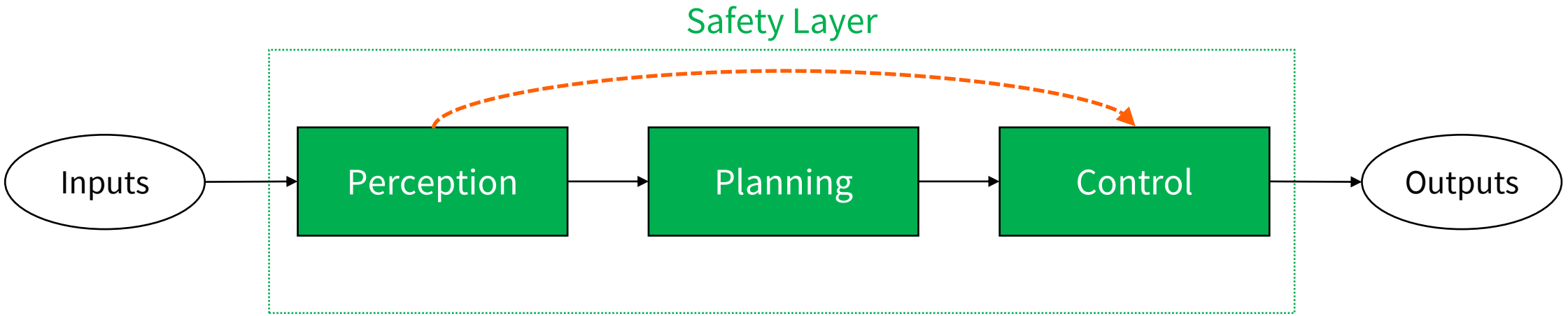


Synergistic Simplex – Across Components

Perception → Control

Proactive Adaptation

Upcoming environmental changes



Synergistic Simplex – Across Components

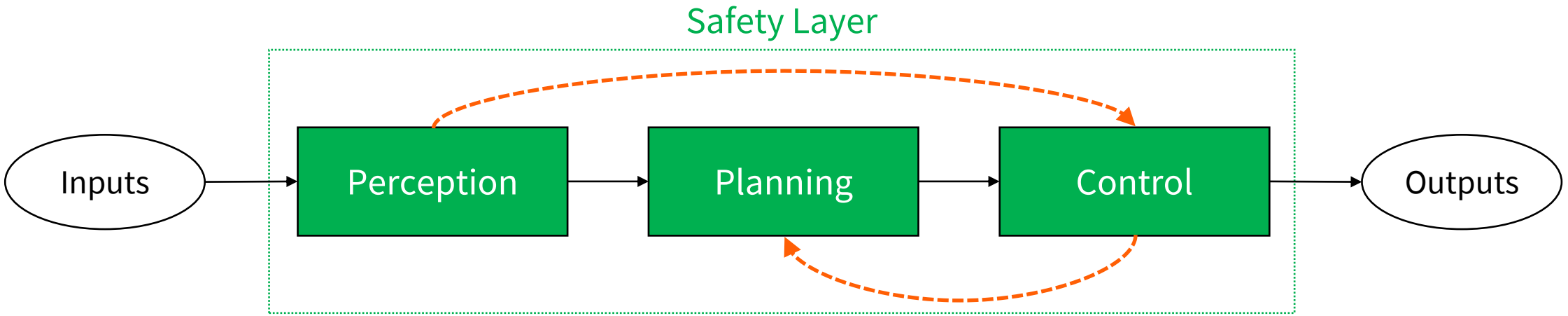
Goals

Safety Guarantees ✓

Utilize ML capabilities ✓

Minimize Performance Loss ✓

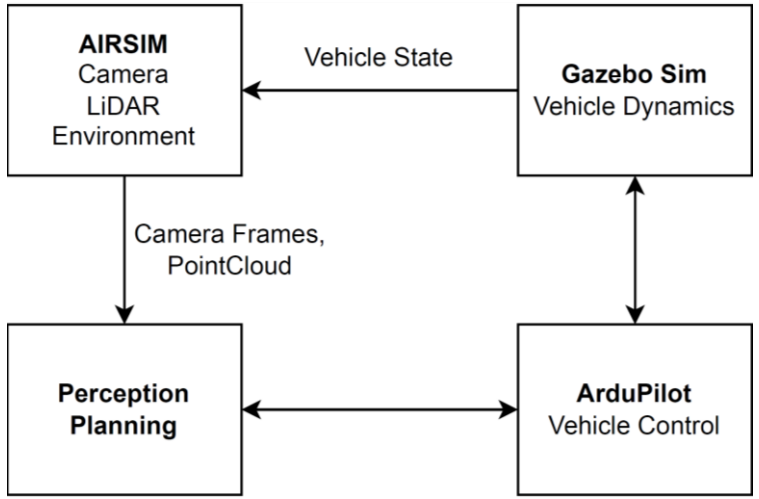
Verifiable Fail-Safe
Best-Effort Fail-Operational



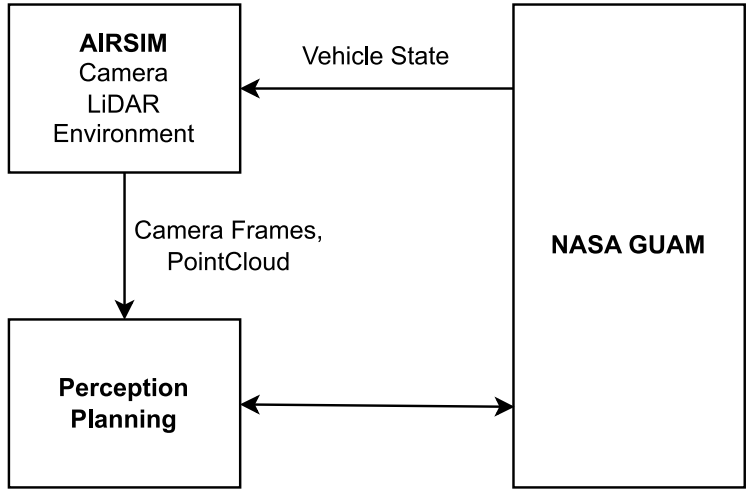
Platforms

Polaris GEM e2

PhotoRealistic Simualtor



NASA GUAM Integration



Source: Petros Voulgaris, Hyung-Jin Yoon

The Team

Lui Sha

Naira Hovakimyan

Bo Li

Tarek Abdelzaher

Petros Voulgaris

Marco Caccamo

Hunmin Kim

Sheng Cheng

Hyung-Jin Yoon

Ayoosh Bansal

Simon Yu

Yuliang Gu

Micaela Verucchi

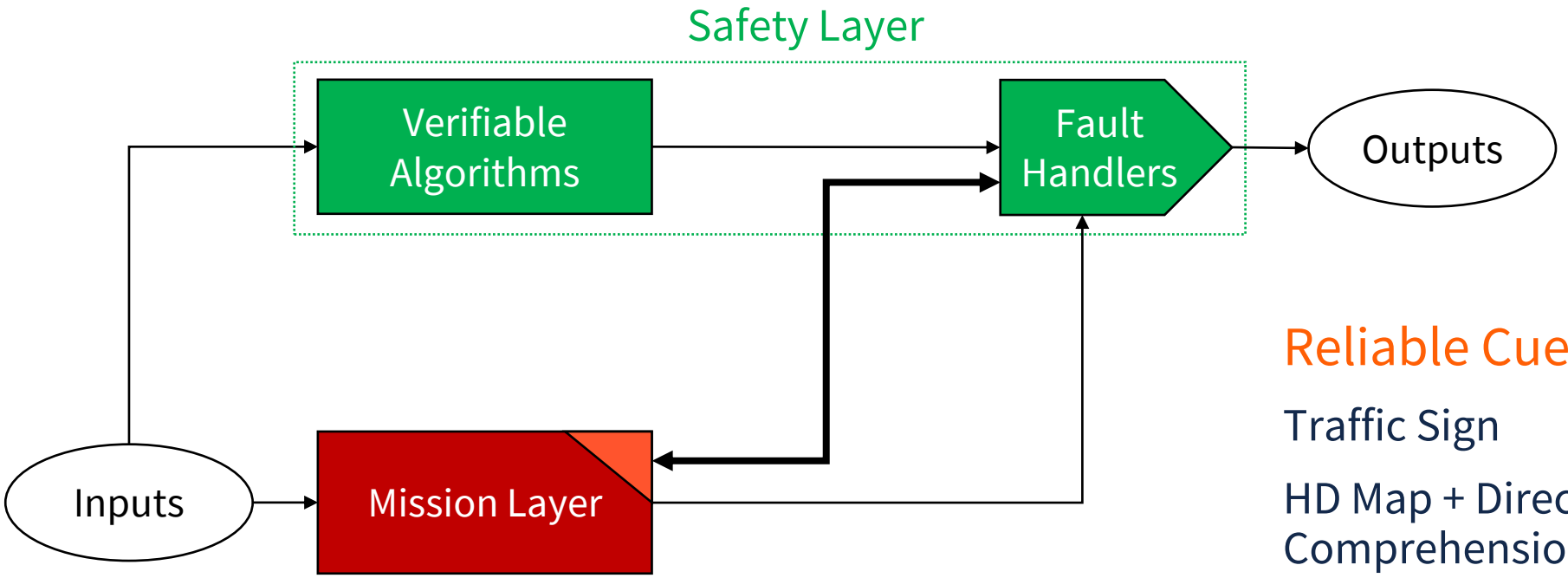
Jayati Singh

Yang Zhao

James Zhu

Ethan Pereira

Synergistic Simplex – Reliable Mission Capabilities



Reliable Cues

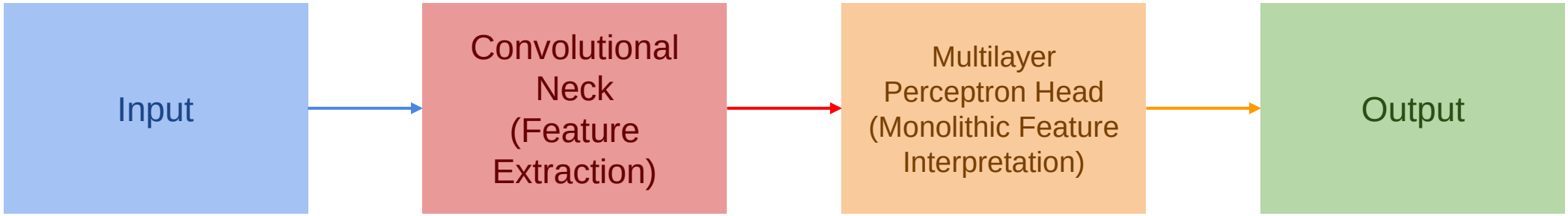
Traffic Sign

HD Map + Direct Comprehension

VISAT: Using Visual Attributes to Benchmark Image Recognition Robustness under Adversarial Attack and Distribution Shift

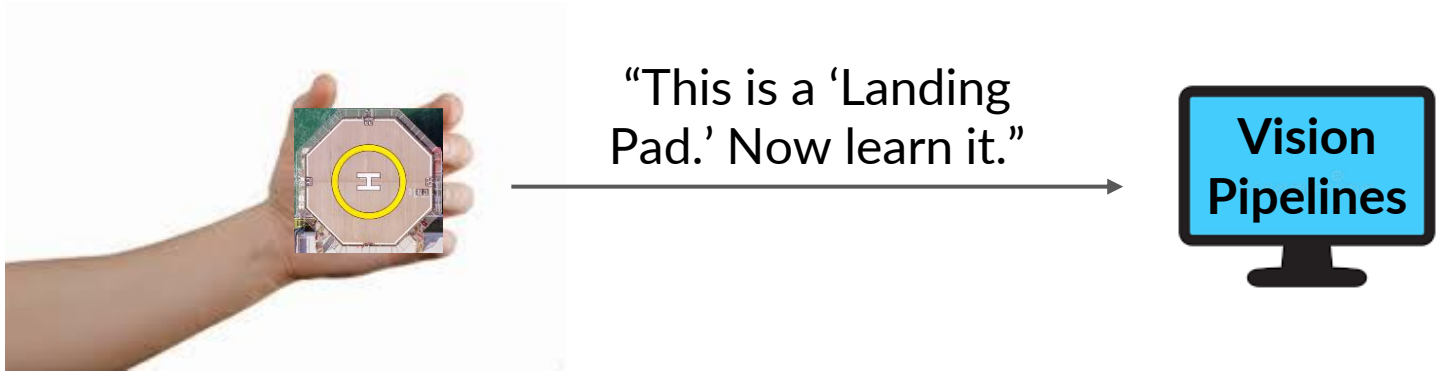
Simon Yu, Peilin Yu, Hongbo Zheng, Huajie Shao, Han Zhao, Lui Sha

Motivation: Purely Data-Driven Vision Pipelines



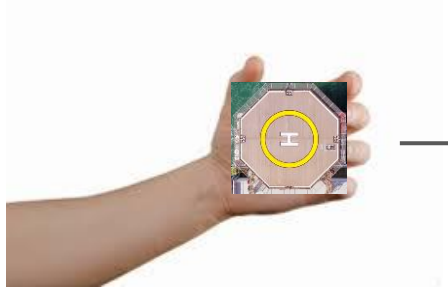
- Feature interpretations are trained into the MLP head using purely data-driven approach.
- The internal logic of the MLP head during training and inference remains as a black box.
- Limited backtrace and analyzability for any misclassifications made by the MLP head.

Motivation: Purely Data-Driven Vision Pipelines



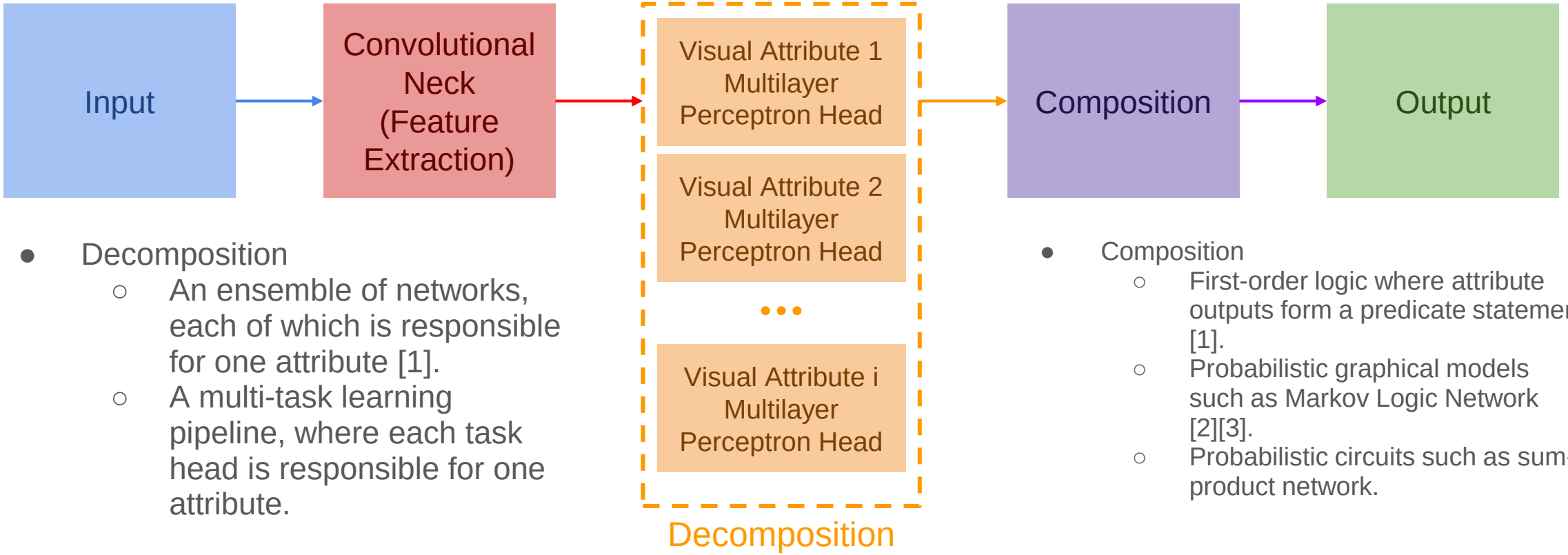
- Many existing deep learning vision datasets designate simple labels for each class.
  - with no explicit description of any features of the objects.
  - expects the vision pipelines to “figure out” corresponding features to the objects all on their own.
- We have limited knowledge on the features the vision pipelines learned for recognizing objects.
  - which might produce unbounded levels of spurious correlations causing confusions among objects.

Motivation: Decomposed Vision Pipelines using Visual Attributes



"This is a 'Landing Pad' with a brown background, a round circle, and an 'H' symbol."

"Now learn it."

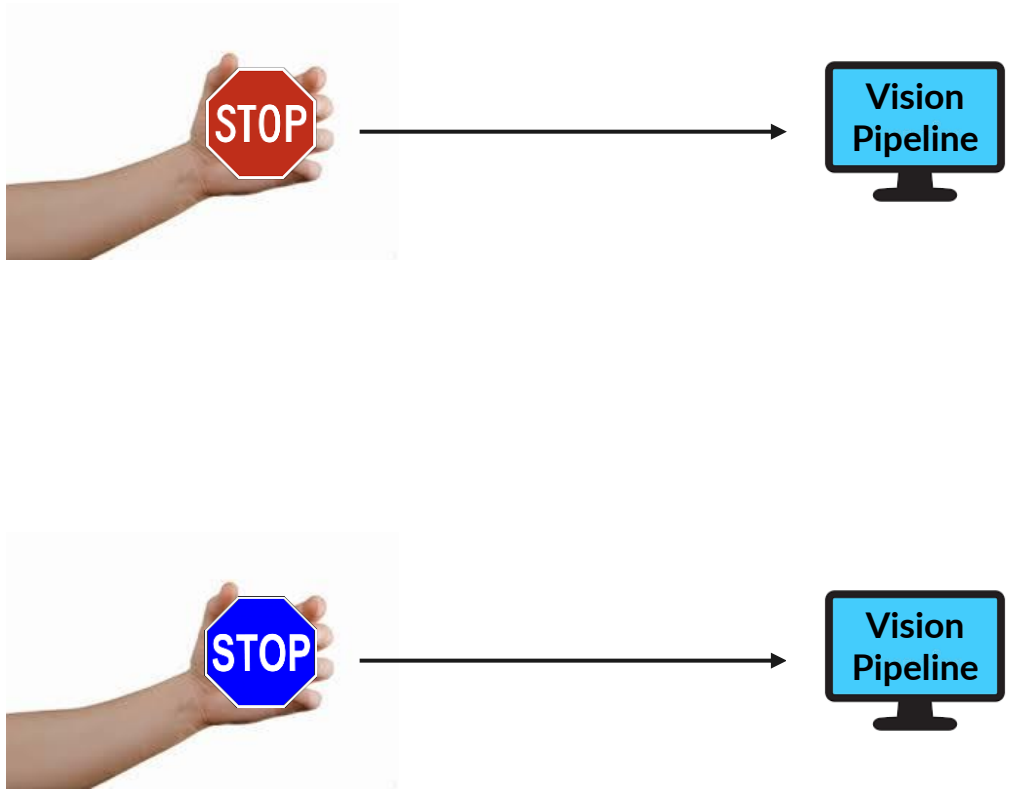


- Decomposition
  - An ensemble of networks, each of which is responsible for one attribute [1].
  - A multi-task learning pipeline, where each task head is responsible for one attribute.

- Composition
  - First-order logic where attribute outputs form a predicate statement [1].
  - Probabilistic graphical models such as Markov Logic Network [2][3].
  - Probabilistic circuits such as sum-product network.

Classification Error Correlation

- Since all visual attributes exist at once given any input, models might gain unintentional reliance on unrelated cues instead of the desire attributes.
- For example, since all stop signs are octagon and red, models might establish undesired classification error correlation between shape and color of stop signs due to their co-occurrence.
- We demonstrate the level of error correlations in decomposition pipelines by performing attacks targeted on one task while also observing their effects on all other tasks.
- We hypothesize that the error correlations exist not only in MTL network but also in other pipelines such as the ensemble networks due to the co-occurrence of visual attributes given any input during training.



Outline of VISAT

- Creation and Formulation of Visual Attributes
  - Formulation of visual attributes on a large-scale traffic sign recognition dataset.
  - Rapid labeling software enabling the efficient creation of visual attributes for large-scale dataset.
  - Enables robustness benefits demonstrated by [1][2][3].
- Robustness Benchmark
  - Instance-Wise: Adversarial Attack
    - Projected gradient descent [4].
  - Dataset-Wise: Distribution Shift
    - Method 1: color quantization.
    - Method 2: ImageNet-C [5] data corruption and natural variation.
  - An evaluation platform for gauging robustness of the decomposition components.
- Applications on the UAM/eVTOL Use Cases

Visual Attributes: Formulation

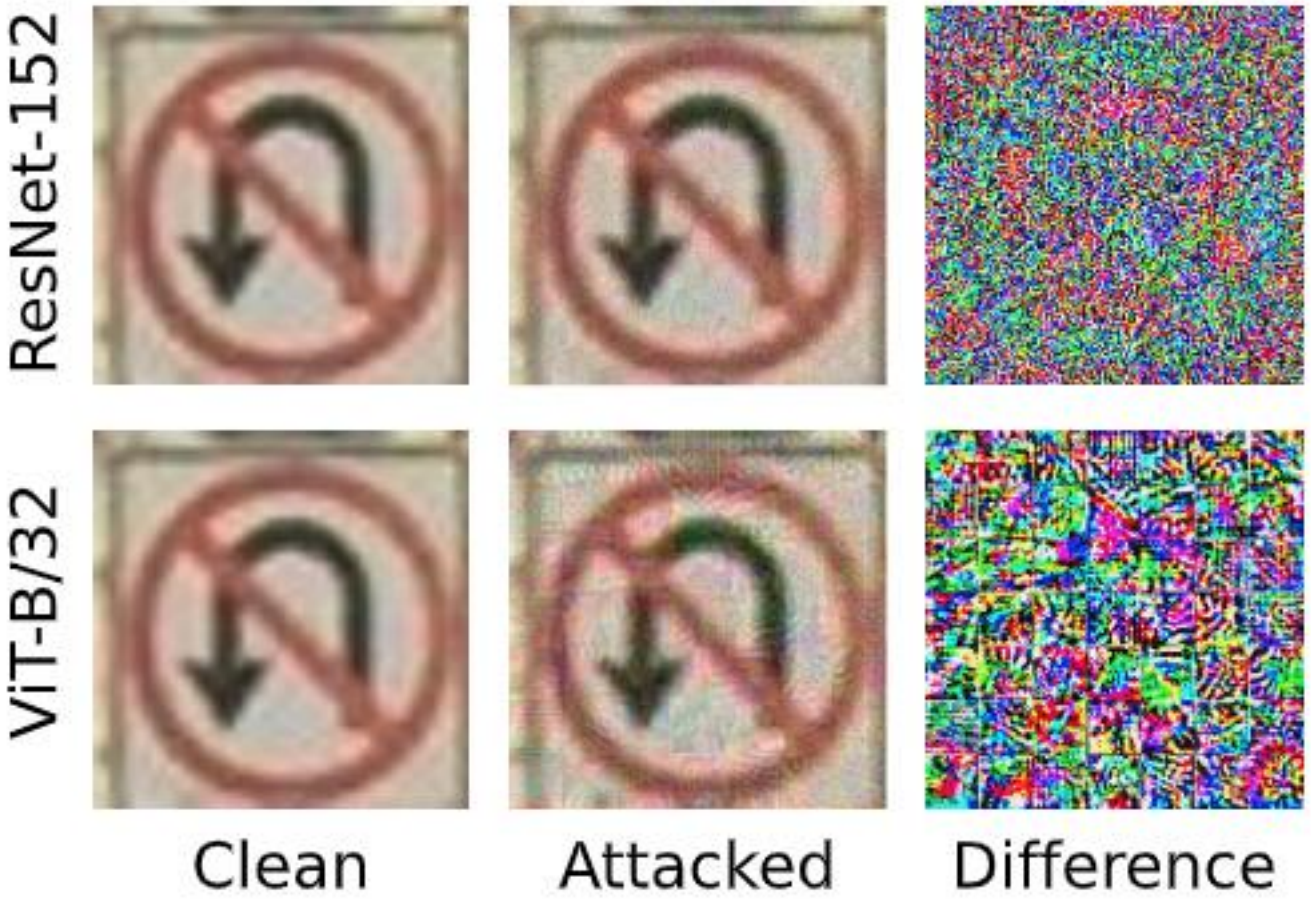
- 4 types of visual attributes (in the context of traffic sign recognition)
  - Color, shape, symbol, and text.
- Color attributes
  - Captures major background and foreground colors, e.g., “color--red--white” for (a).
- Shape attributes
  - Directly encode shapes of the signs, e.g., “shape--triangle--inverted” for (b).
- Symbol attributes
  - Described by words in geometric order, e.g., “symbol--arrow--down--arrow--up” for (c).
- Text attributes
  - Embed texts written on signs, e.g., “text--alphanumeric” for (d).

Visual Attributes: Rapid Labeling Software

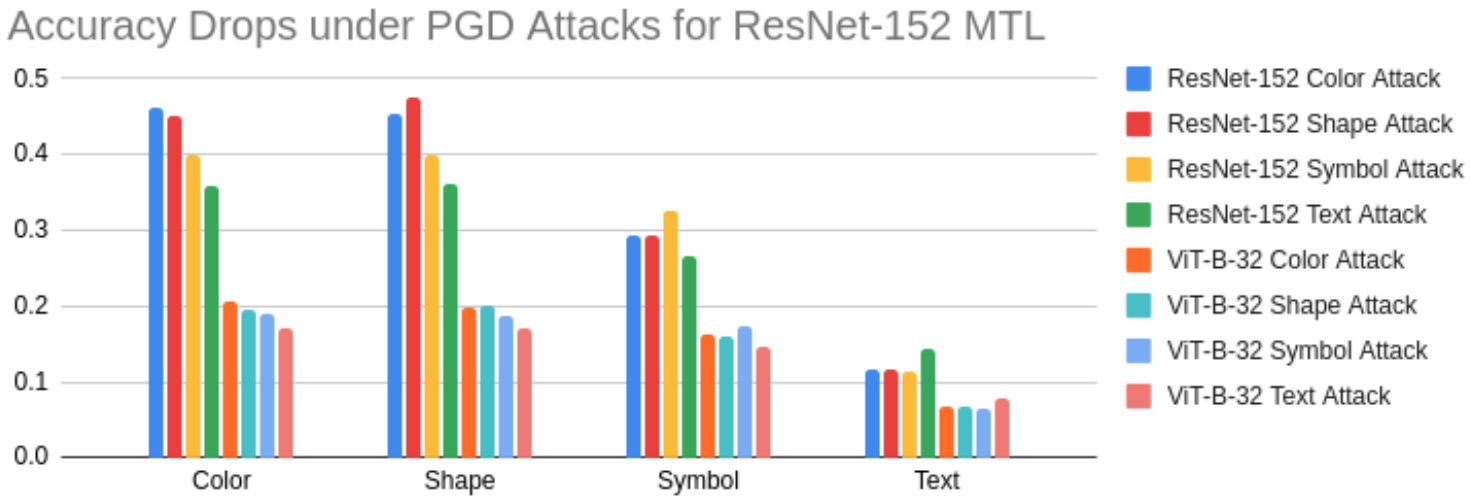
- Enables the efficient creation and labeling of visual attributes for large-scale datasets.
- View, Labeling, and Application Control divisions help human labelers observe and formulate visual attributes, and perform rapid labeling of visual attributes over a large set of classes.
- Fully customizable to accommodate various screen sizes as well any any type of user-defined attribute.
- Dataset-agnostic and can be used to facilitate creations of visual attributes for any dataset so long as they can be conceptualized and formulated by humans.

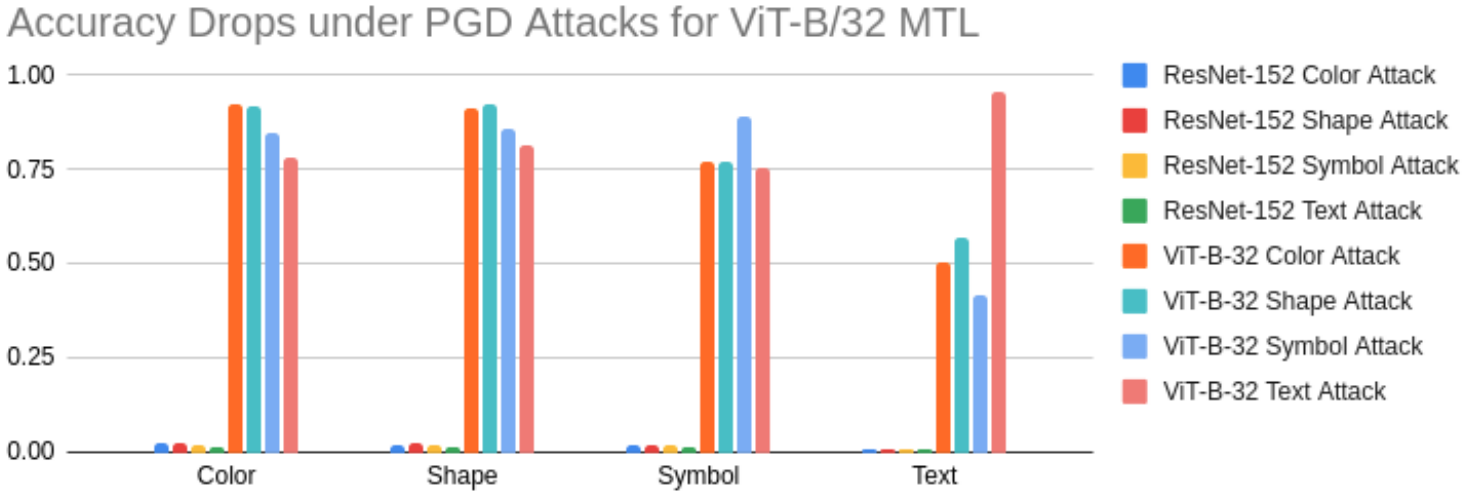
Robustness Benchmark: Adversarial Attack (PGD)

- Model-Specific
  - Generated attacks target one specific model.
- Projected Gradient Descent
  - Attacks generated based on model parameters.
  - Maximize model loss function by manipulating input within certain bound.
- Resulting Attacks
  - Limited alterations of original inputs.
  - Effective against attacked models.



Robustness Benchmark: Adversarial Attack (PGD)

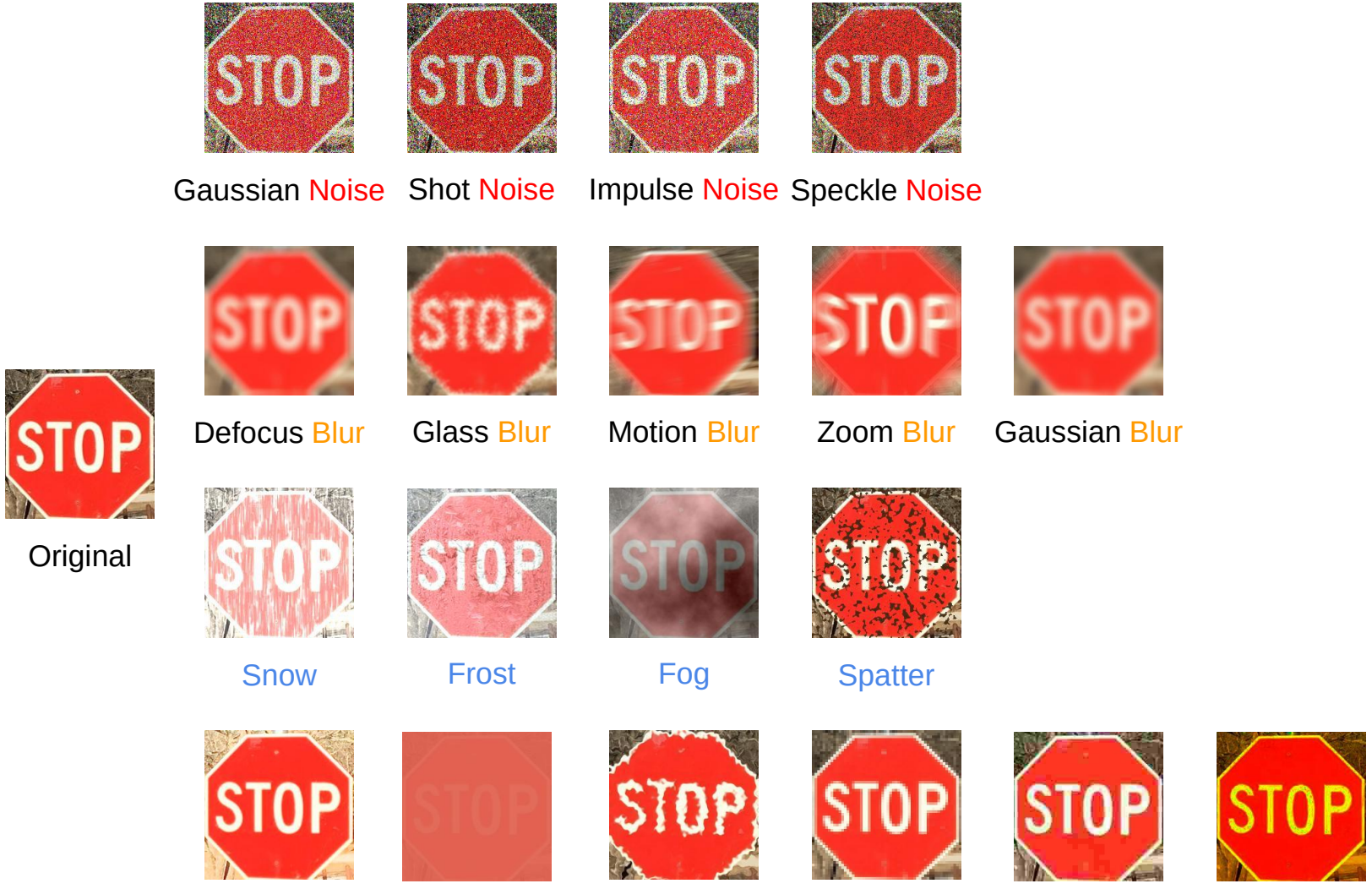




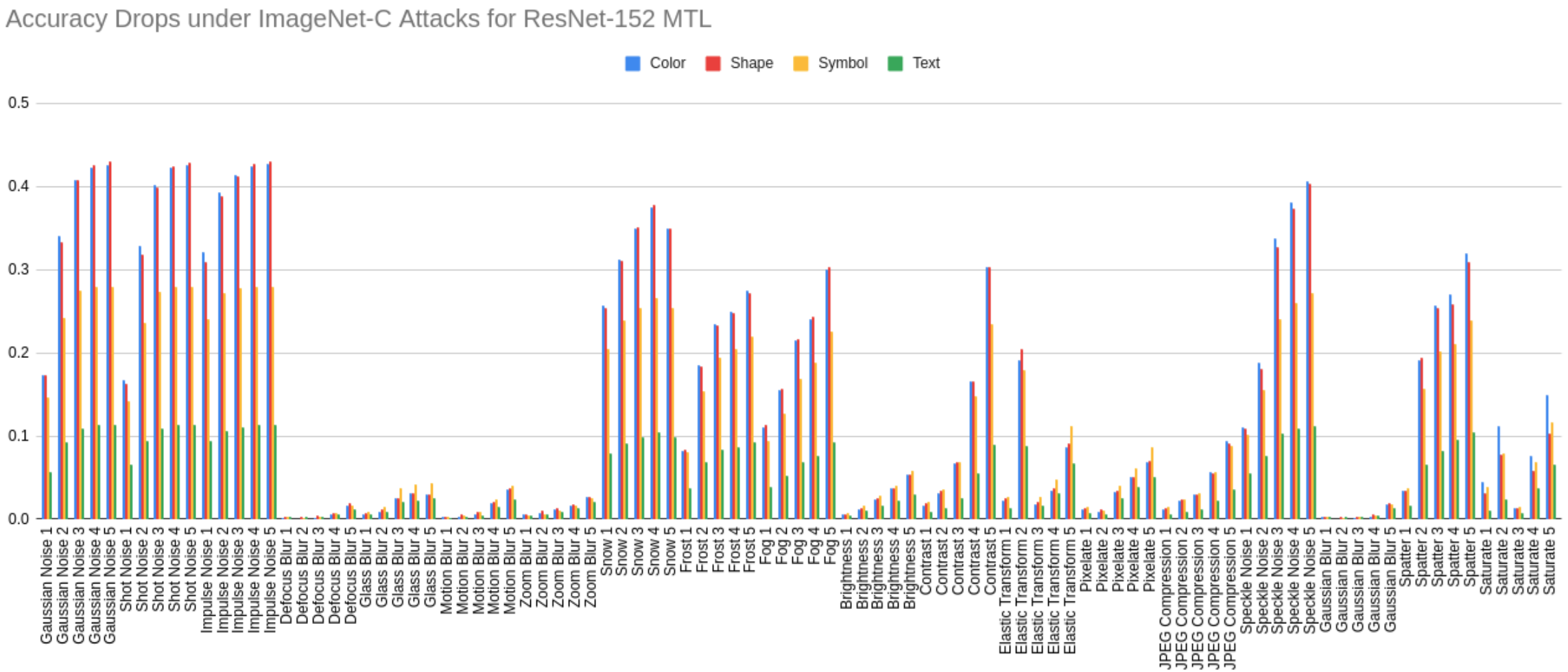
- We evaluate the effect of PGD attacks on the MTL models training using our visual attributes.
- The ViT-B/32-based attacks are effective on both ResNet-152 and ViT-B/32.
- The spurious correlations among the MTL tasks is demonstrated by performing an attack specific to one task, and observing its effects on remaining tasks.
- Given the results, we see noticeable spurious correlations between color and shape task heads.

Robustness Benchmark: Distribution Shift (ImageNet-C)

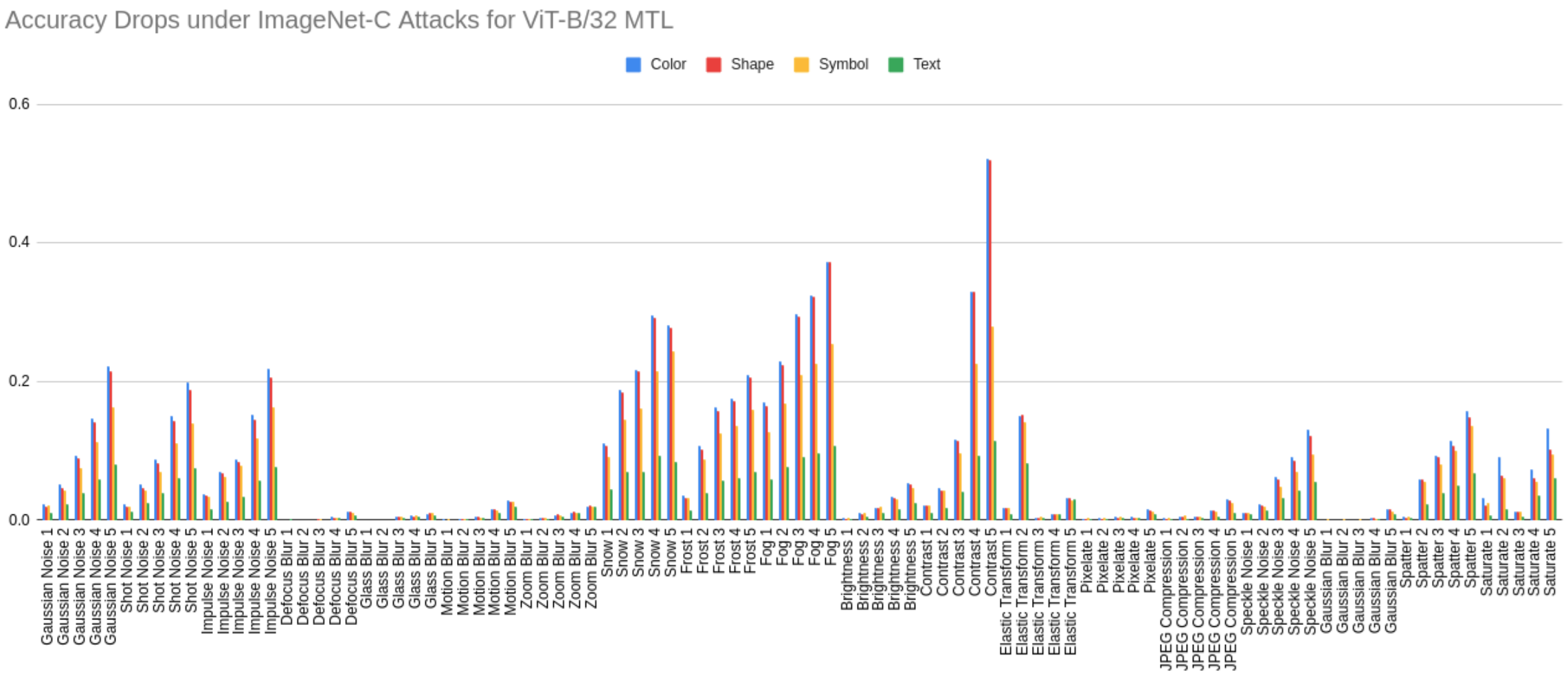
- Model-Agonistic
  - Generated attacks do not depend on model details.
- ImageNet-C
  - 19 individual data corruption and variations, with 5 severity for each variation.
- There are 4 different major variation types: noise, blur, weather, and digital.



Robustness Benchmark: Distribution Shift (ImageNet-C)



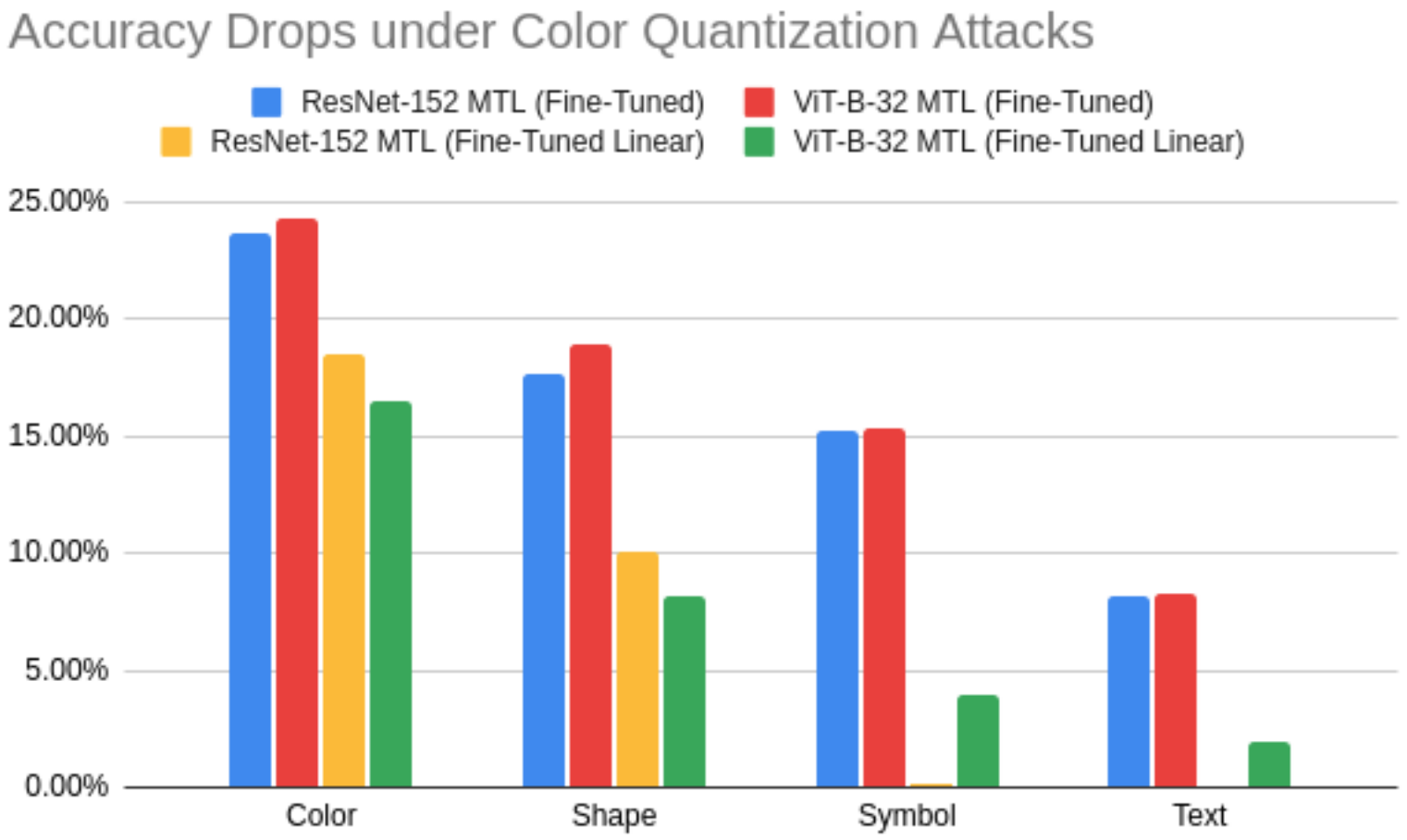
Robustness Benchmark: Distribution Shift (ImageNet-C)



Robustness Benchmark: Distribution Shift (Color Quantization)

- Model-Agonistic
  - Generated attacks do not depend on model details.
- Color Quantization
  - Quantize or cluster elements of color for each sign, enabling modifications on their color palettes.
- Attribute-Specific
  - Generated attacks specifically targets color attributes with minimal impacts on others.

Robustness Benchmark: Distribution Shift (Color Quantization)



- The level of spurious correlations among the MTL tasks is seen by observing the effects of the attack on the remaining tasks.
- Ideally, if the MTL tasks are truly independent, performing attacks on color should have no effect on the remaining tasks.

Applications on the UAM/eVTOL Use Cases

- We performed comprehensive studies on model robustness using visual attributes in the context of traffic sign recognition.
- Nevertheless, the concept of visual attributes is universal for any visual recognition tasks.
    - Visual attributes can be easily conceptualized and formulated for many vision tasks in the use case of UAM/eVTOL, e.g., ground area recognition, airport runway sign recognition, urban area insertion, etc.
    - Our rapid visual attribute labeling software is dataset-agnostic and therefore can be used for many other vision datasets.
- Our robustness benchmarking pipeline also remain useful, can be adapted to urban aerial use cases, and serve as a robustness evaluation platform for decomposition pipeline designs.
- Our VISAT dataset and benchmarks are submitted to NeurIPS 2023 and is available to the public: http://rtsl-edge.cs.illinois.edu/visat/
- In the near future, we plan to explore specific decomposition and composition pipeline designs by employing multi-task learning networks and probabilistic circuits.

System Model Summary

Safe Air + Ground Vehicle (flying taxi) without landing pads.
There are three key components :

- Verifiably safe collision avoidance using physics model supervised ML vision
- Robust (statistically reliable) traffic sign recognition against noises and attacks.
- Safety-driven integration of
  - Safe perception (Perception Simplex, new)
  - Safe control (L1Simplex, existing work)
- => Synergist Simplex Architecture
- **Approach:** extend our research on safe autonomous ground vehicles to flying taxi

References

- [1] Gürel, Nezihe Merve, et al. "Knowledge enhanced machine learning pipeline against diverse adversarial attacks." International Conference on Machine Learning. PMLR, 2021.
- [2] Zhang, Jiawei, et al. "CARE: Certifiably Robust Learning with Reasoning via Variational Inference." 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). IEEE, 2023.
- [3] Yang, Zhuolin, et al. "Improving certified robustness via statistical learning with logical reasoning." Advances in Neural Information Processing Systems 35 (2022): 34859-34873.
- [4] Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).
- [5] Hendrycks, Dan, and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." arXiv preprint arXiv:1903.12261 (2019).