

Resilient AI against Adversarial Interactions in Autonomous Vehicular Perception



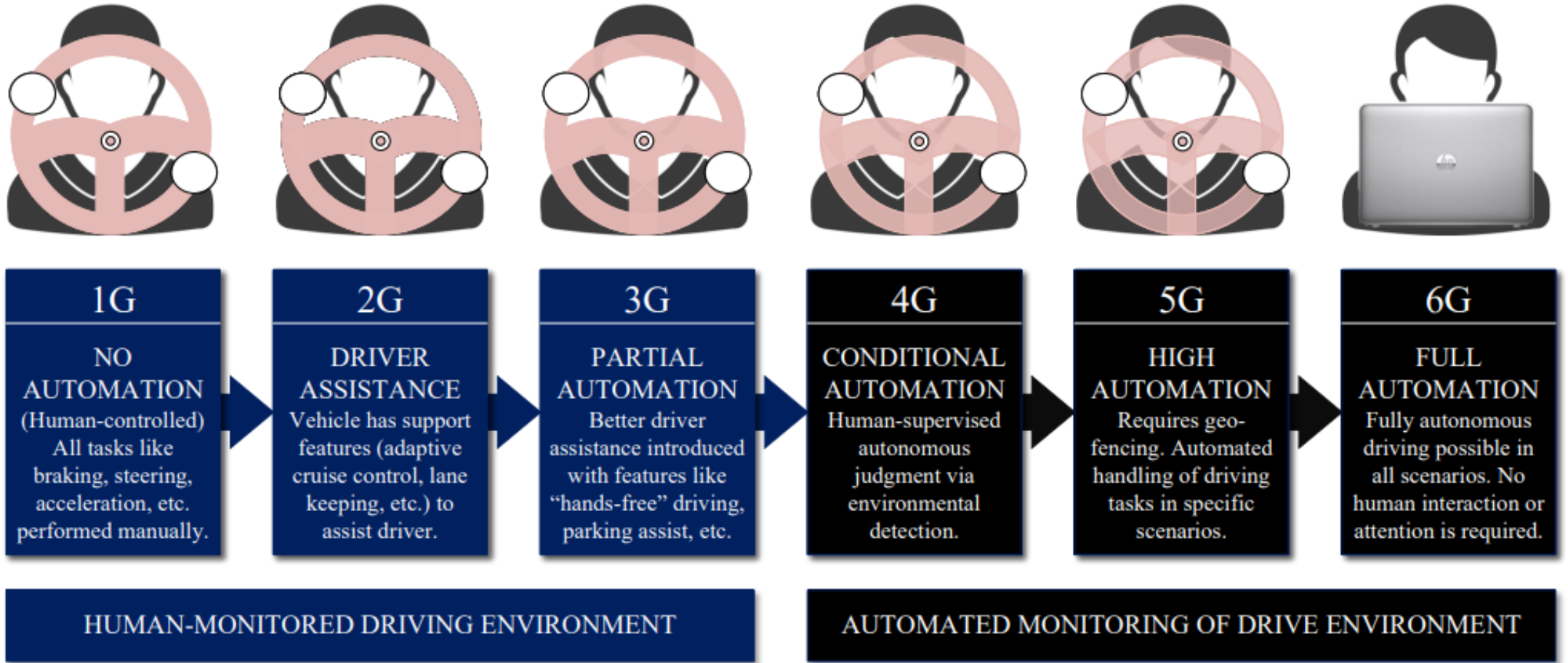
Tapadhir Das, Ph.D.
Assistant Professor
Department of Computer Science
University of the Pacific

University of the Pacific

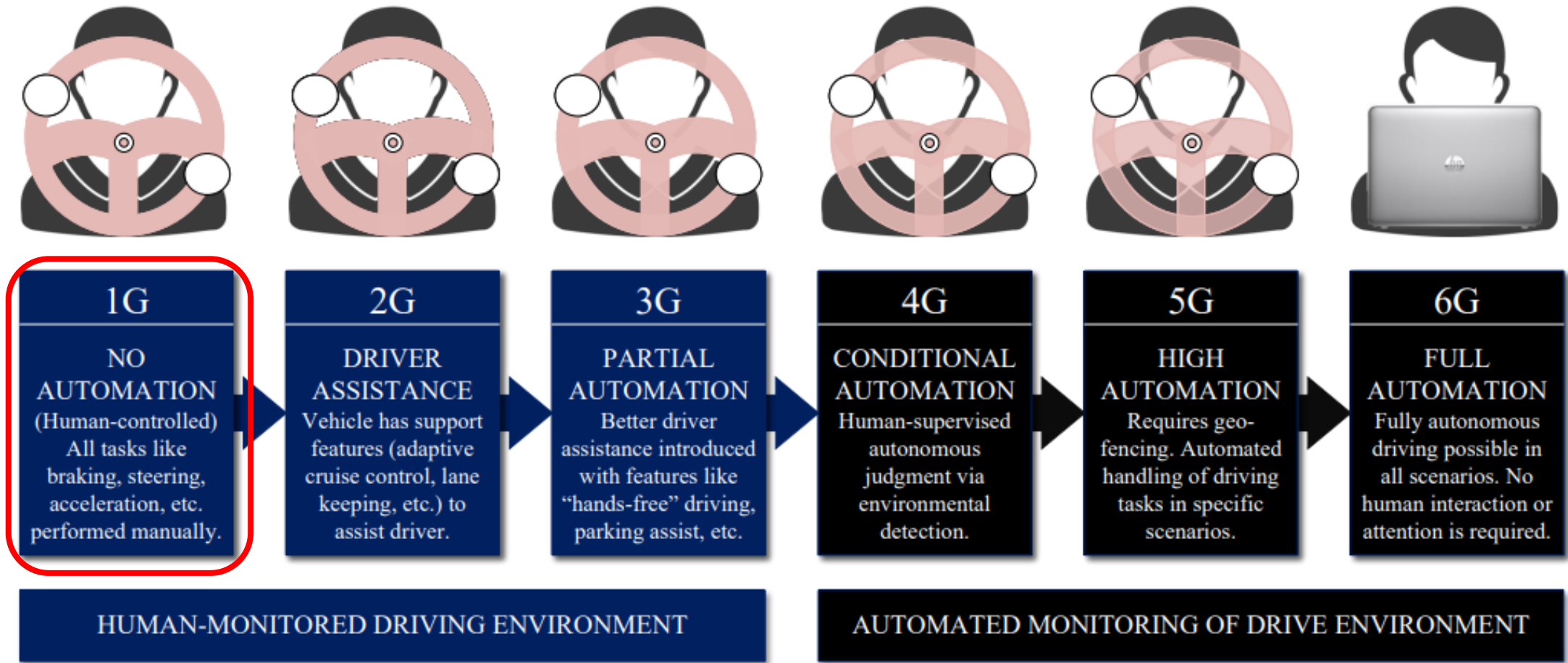
- California's Oldest University (1851)
- Student-centered institution, focused on teaching and experiential learning.
- Primarily Undergraduate Institution (PUI)
- Tremendous emphasis on undergraduate research

Dependability And Security Lab

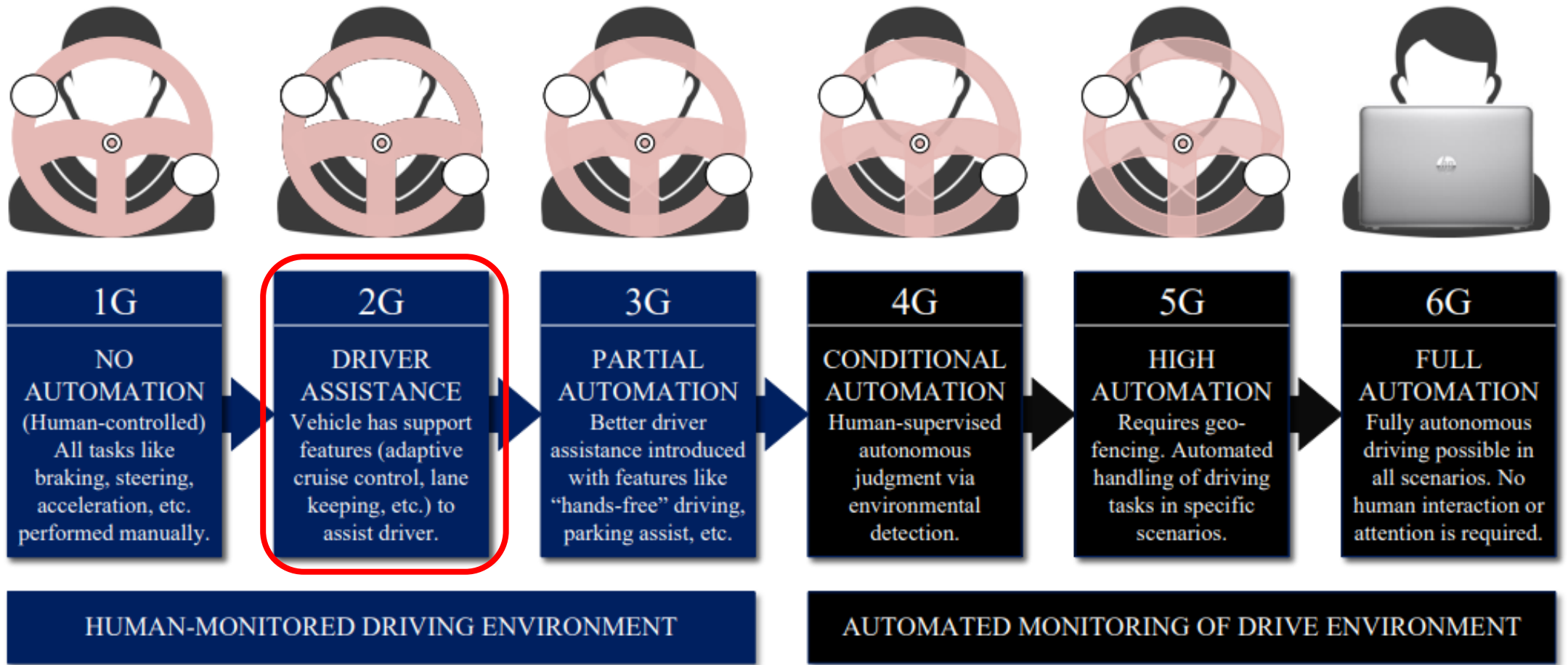
- Research Topics:
 - Smart Grid Security
 - Resilient and Trustworthy AI
 - Intelligent Vehicle Security
- Currently have 6 undergraduate and graduate researchers
- Alumni have gone on to multiple PhD positions at R1 schools and major industrial companies such as NVIDIA
- Since 2023, we have published 17 peer-reviewed publications in reputable IEEE conference and journal venues



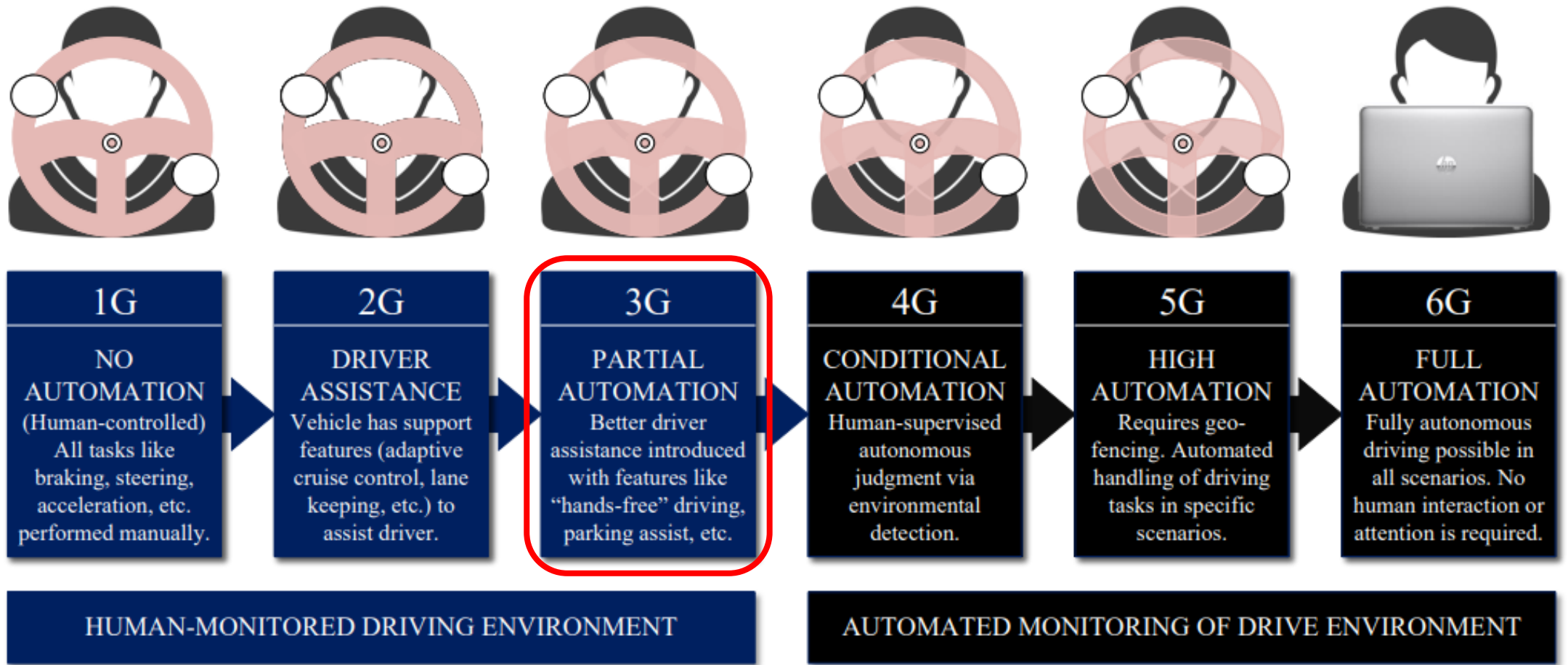
Autonomous Vehicle Evolution



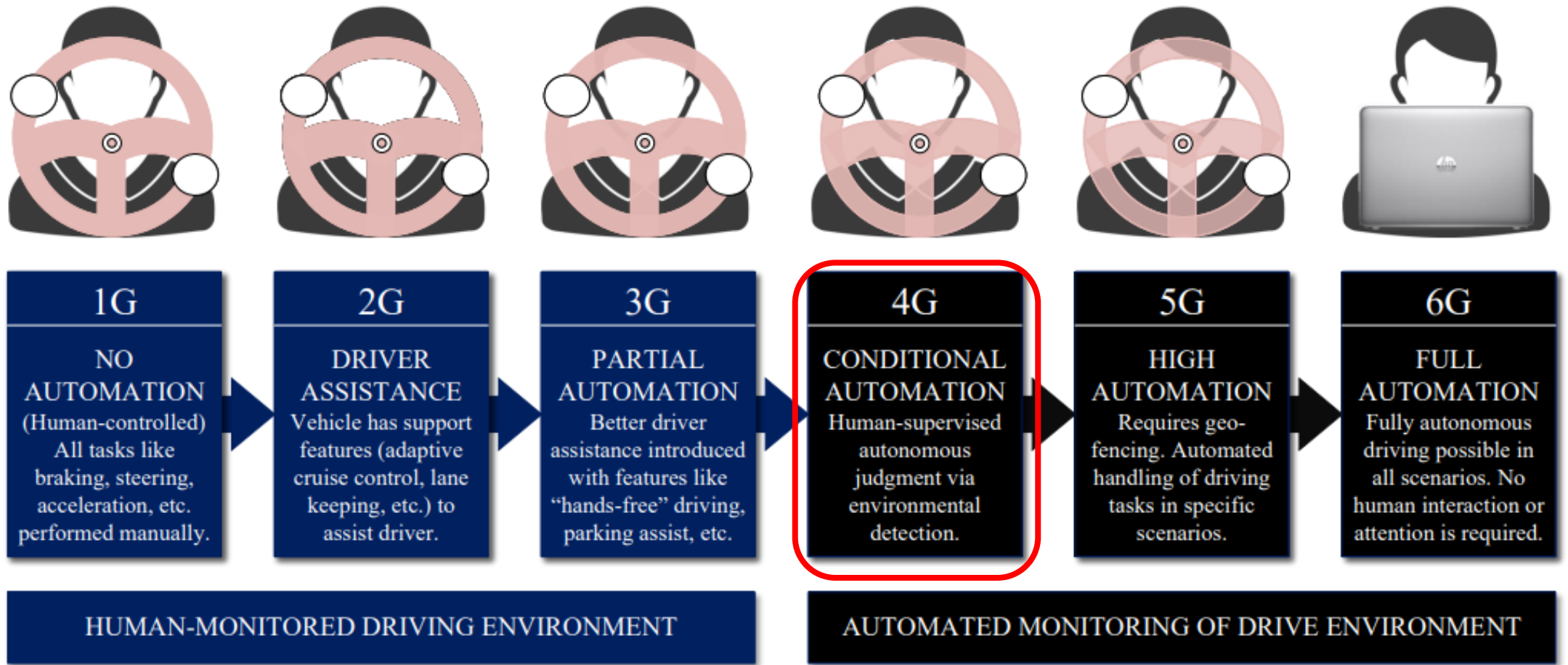
Autonomous Vehicle Evolution



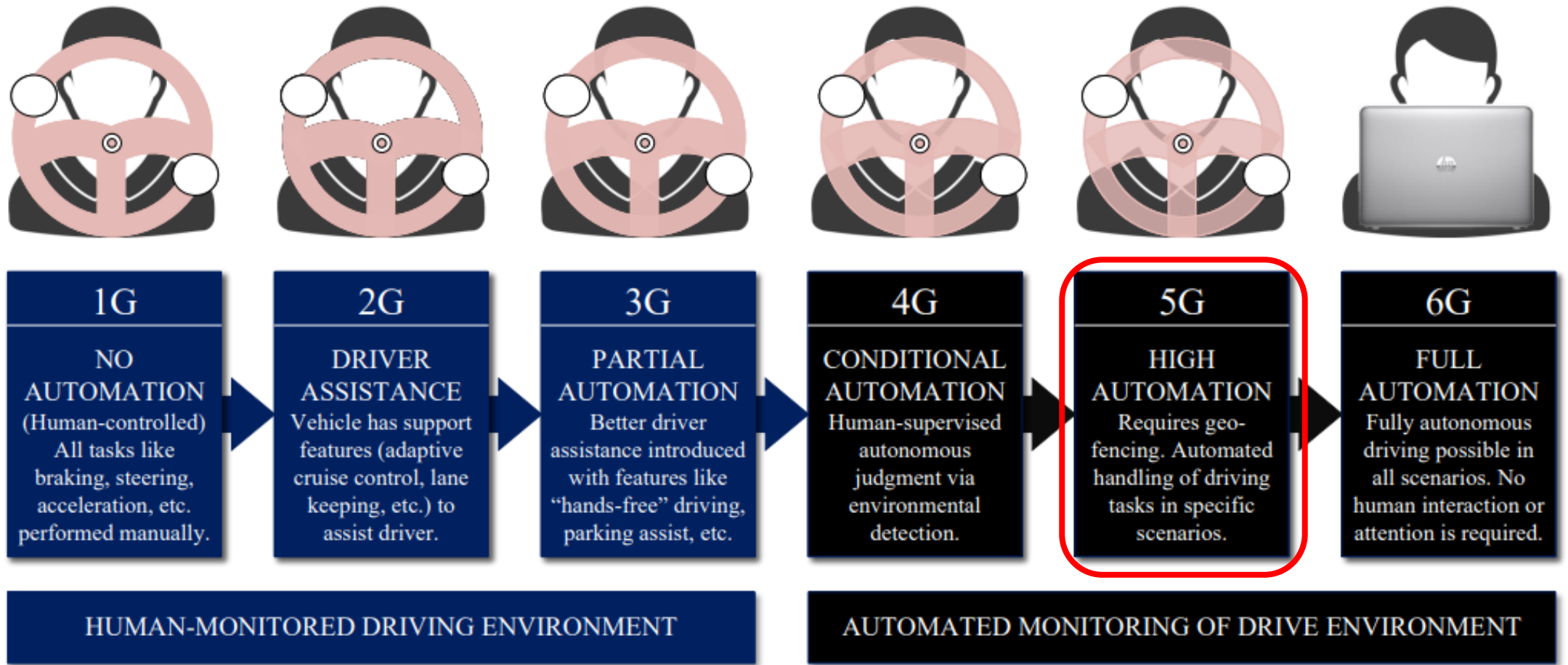
Autonomous Vehicle Evolution



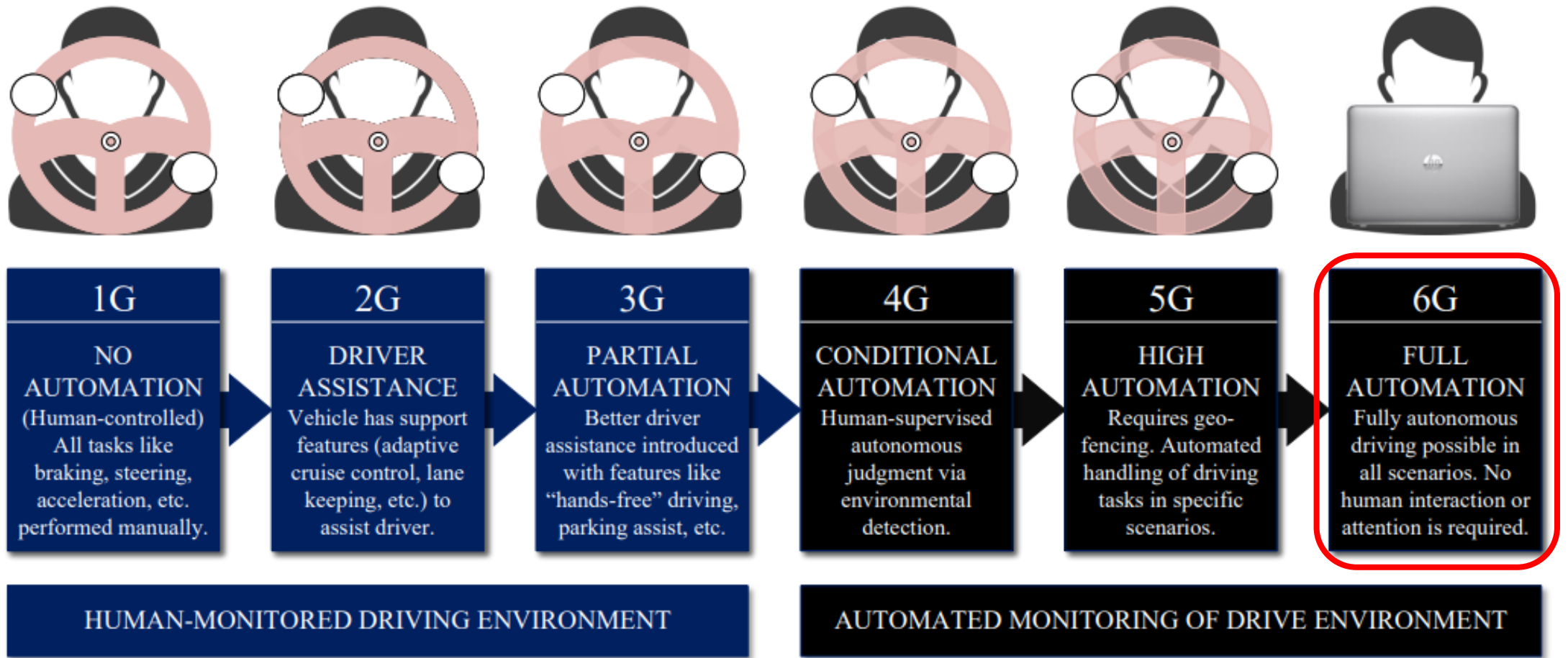
Autonomous Vehicle Evolution



Autonomous Vehicle Evolution



Autonomous Vehicle Evolution



Autonomous Vehicle Evolution

Autonomous Vehicles and Benefits

- Modern-day AVs are classified as Level 5G.
- Corporations: Waymo, Tesla, Zoox
- Benefits:
 - Driver assistance
 - Road Safety
 - Reduced human error





Most people hesitate to consistently ride an AV

Recent News

6TH STREET MASS SHOOTING

VIDEO: Waymo seen blocking first responders during West Sixth shooting

by: Esmeralda Zamora

Posted: Mar 1, 2026 / 12:55 PM CST

Updated: Mar 2, 2026 / 12:54 PM CST

TECHNOLOGY

Waymo self-driving taxi takes passenger through active police scene in downtown LA, video shows

Wednesday, December 3, 2025

TECH

Amazon’s Zoox robotaxi unit issues software recall after recent Las Vegas crash

PUBLISHED TUE, MAY 6 2025•1:23 PM EDT | UPDATED TUE, MAY 6 2025•2:15 PM EDT

Annie Palmer
@IN/ANNIERPALMER/

SHARE    

Videos show Waymo cars stuck at San Francisco intersections during massive power outage

Witnesses say the autonomous cars were apparently confused because the traffic lights were out



Sunday, December 21, 2025

Zoox robotaxi strikes parked car, injuring S.F. street ambassador

Zoox confirmed the collision the day after the incident, saying its vehicle 'tried to avoid' a car door opening but could not

by JOE RIVANO BARROS and JOE ESKENAZI
January 17, 2026, 3:47 pm

LOCAL NEWS

Waymo faces federal scrutiny after child is struck near Santa Monica school

by: Lily Dallow

Posted: Jan 30, 2026 / 07:21 AM PST

Updated: Jan 30, 2026 / 07:22 AM PST

Reasons for hesitancy

- Safety and Trust concerns:
 - Fear of Technology
 - Reported incidents
- Psychological Tension:
 - No Driver = No Control
 - Weirdness Factor
- Performance and Practicality:
 - Conservative driving
 - Limited-service areas



Reasons for hesitancy

- Safety and Trust concerns:
 - Fear of Technology
 - Reported incidents
- Psychological Tension:
 - No Driver = No Control
 - Weirdness Factor
- Performance and Practicality:
 - Conservative driving
 - Limited-service areas



Main reasons for Safety and Trust concerns

- Inconsistent performance of on-board AI models within perception systems.
 - Object detection
 - Scene Understanding
 - Response
- Inconsistent performance due to manipulated environment and sensors:
 - Adversarial samples
 - Occlusions
 - Accidents/vandalism
 - Fog/extreme weather



Inconsistent Object Detection



Inconsistent Scene Understanding



Inconsistent Sign Recognition



Accidents/Vandalism

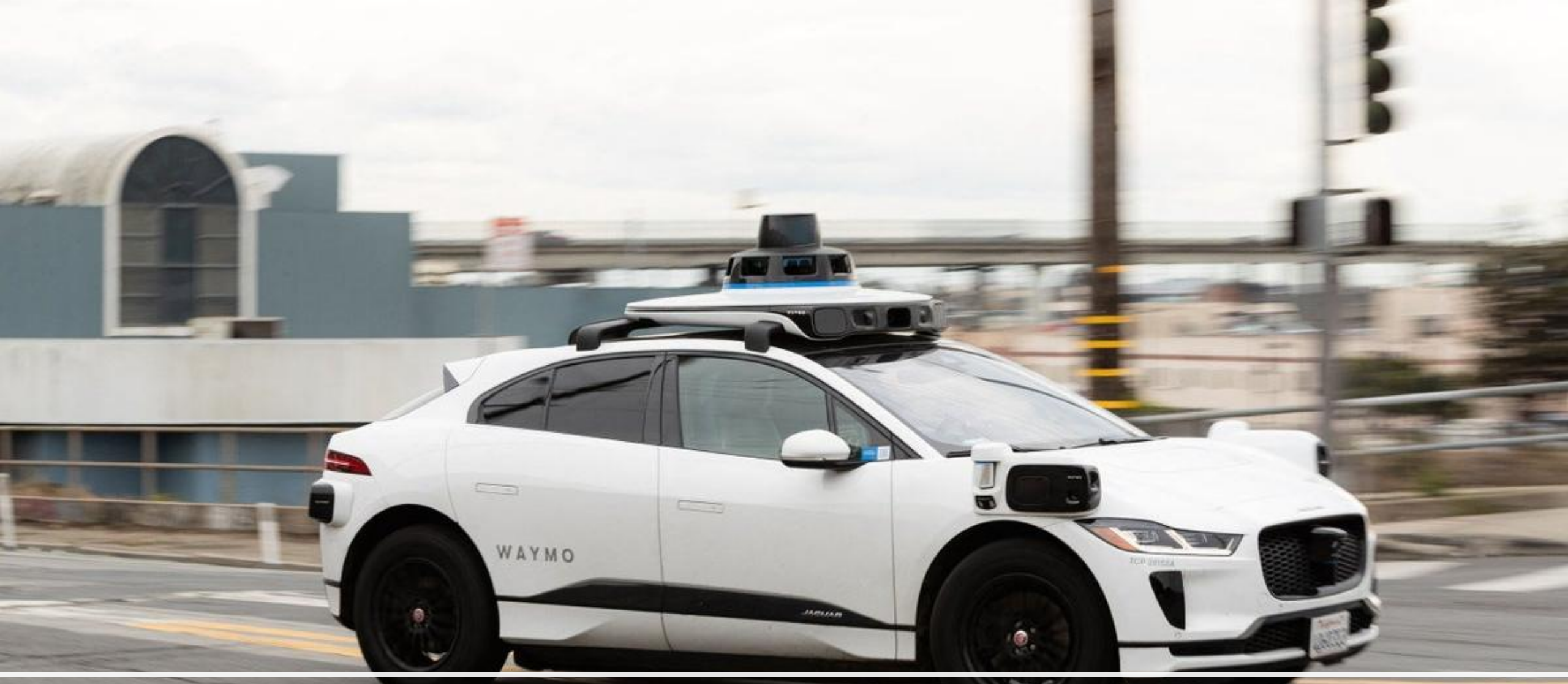




Current research direction

- Within DAS Lab, our current research directions include:
 - Resilient Traffic Sign Recognition
 - Real-time identification of adversarial samples in AV environments
 - Restoration of adversarial samples, digitally, using advanced AI frameworks
 - Real-time Vandalism Detection and Restoration
 - Sensor verification and validation under vandalism conditions
 - Real-time camera feed restoration under vandalism scenarios





Resilient Traffic Sign Recognition

Real-time identification of adversarial samples in AV environments

- Adversarial samples, specifically on AV camera modules, have received attention lately.
- Classic example: Traffic sign recognition
- Inefficient traffic sign recognition can put passengers, drivers, and pedestrians at risk.
- **Research Questions*:**
 - How effectively can AI models perform traffic sign recognition?
 - What happens when adversarial signs are introduced?
 - How effectively can AI models identify manipulated traffic signs to isolate them?

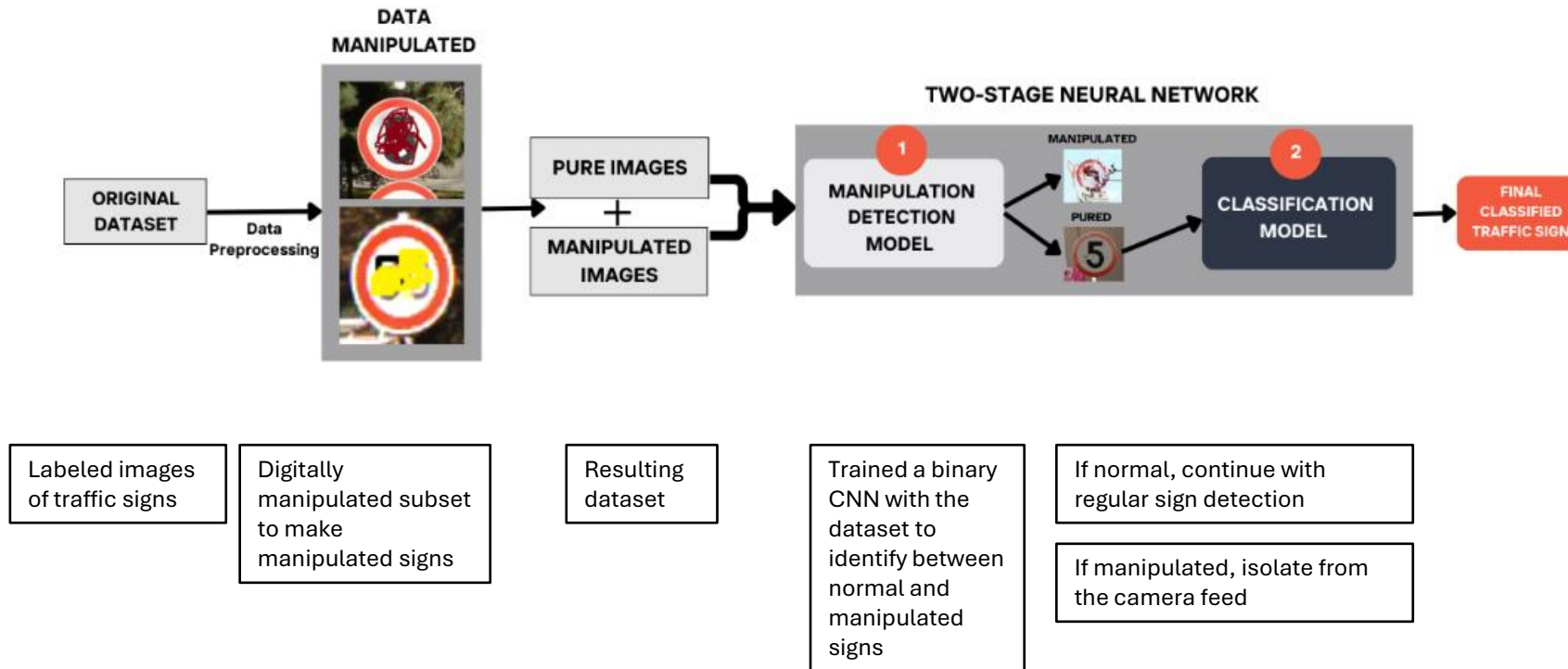


(a) Normal

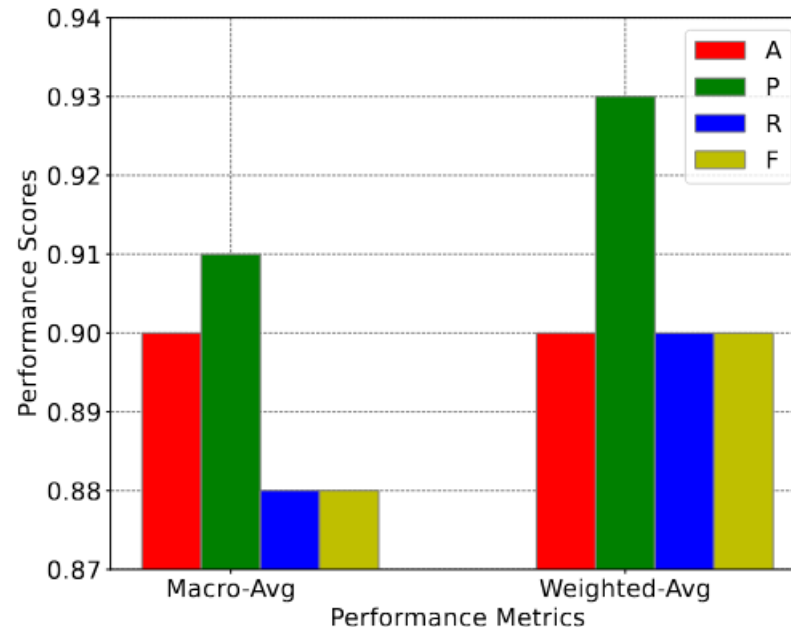


(b) Manipulated

Real-time identification of adversarial samples in AV environments

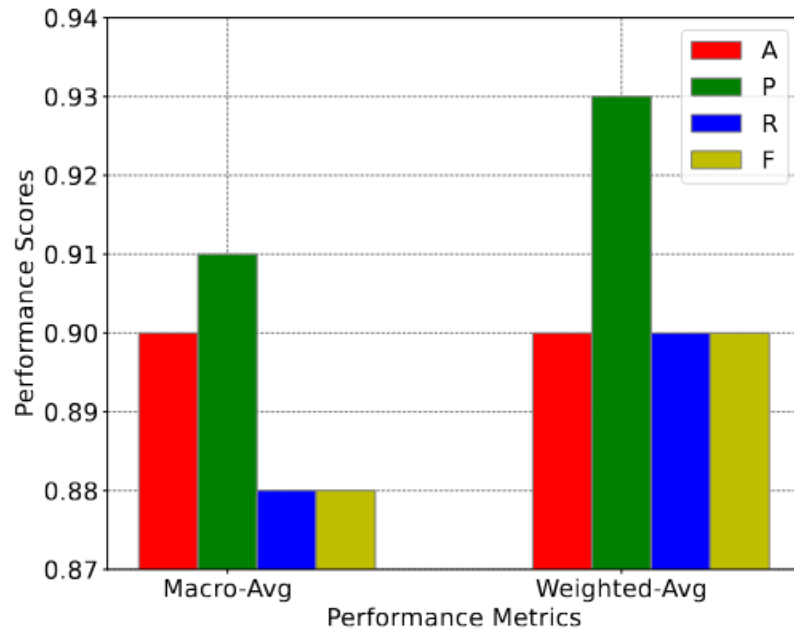


How effectively can convolutional neural networks (CNNs) perform traffic sign recognition?

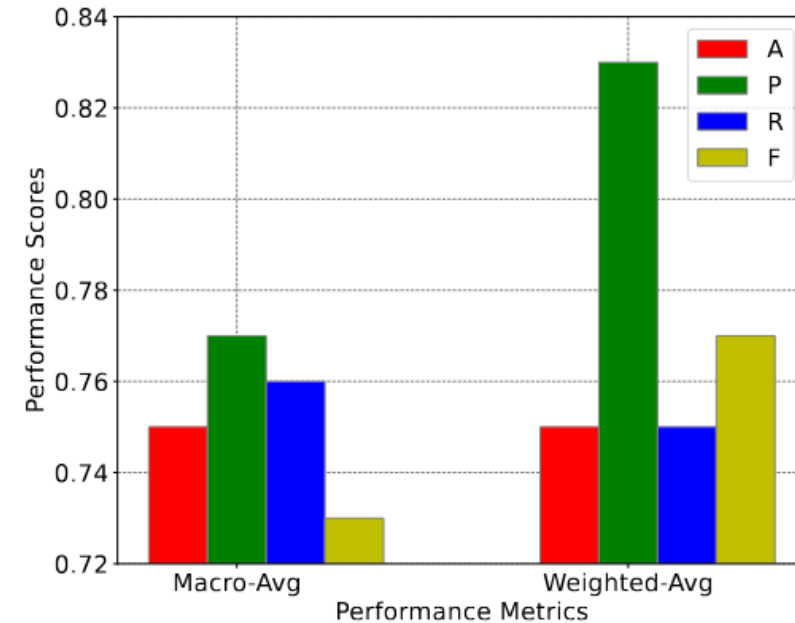


We observe that regular CNNs trained to do multi-class traffic sign recognition achieve high accuracy, precision, recall, and F1 scores at normal traffic sign recognition

What happens when adversarial signs are introduced?



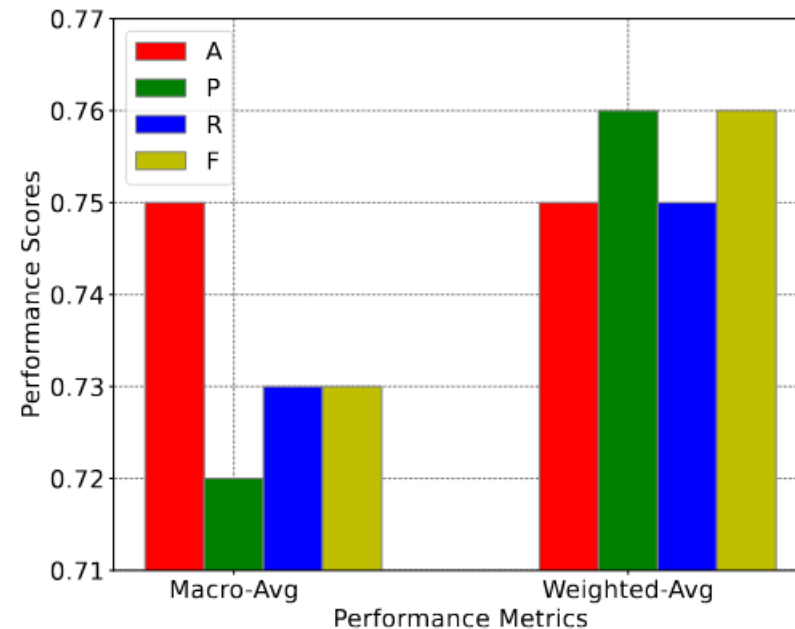
Normal Conditions



Manipulated Conditions

When adversarial signs are introduced, we observe that standard ML metrics like accuracy, precision, recall, and F1 scores take a hit for regular CNNs trained to do multi-class traffic sign recognition, as traffic signs are not so easy to recognize

How effectively can AI models identify manipulated traffic signs to isolate them?



Finally, we observe that our trained binary CNN can distinguish between normal and manipulated images with about a 75% accuracy. This provides us with a baseline for our future research efforts.

Restoration of adversarial samples, digitally, using advanced AI frameworks

- The previous research highlighted that AI models in AVs can distinguish between normal and manipulated traffic signs
- **Research Question:**
 - Is it possible to identify what the actual sign is underneath the manipulation?
 - Are some signs more difficult to identify than others?
 - What about if the level of manipulation is low vs high?



(a) Occlusion
(Sticker)



(b) Graffiti / Paint
Corruption



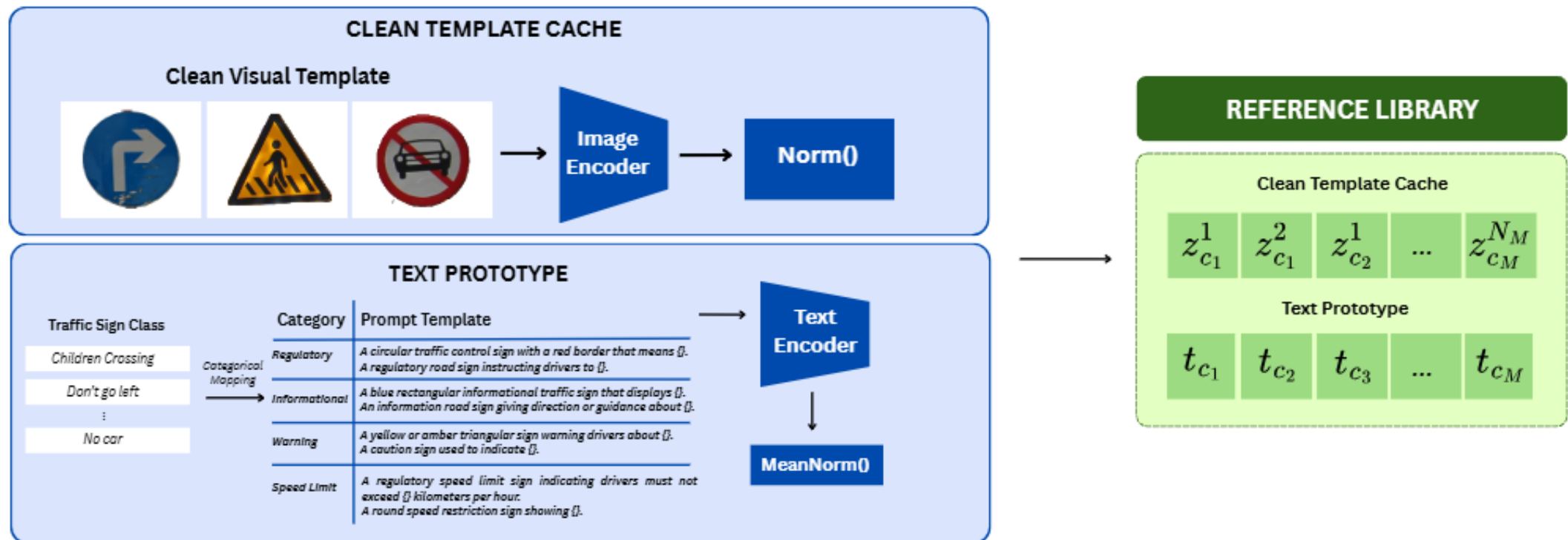
(c) Illumination /
Color Distortion

Restoration of adversarial samples, digitally, using advanced AI frameworks

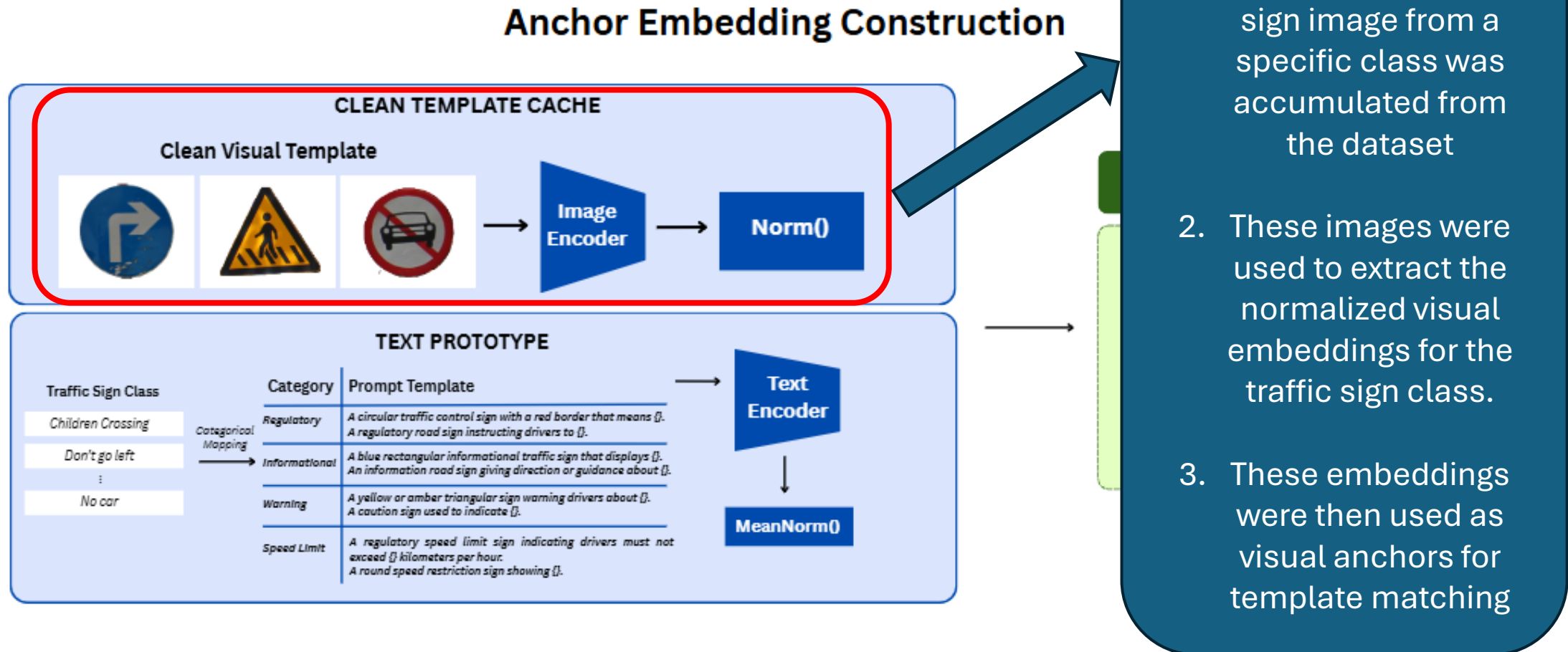
- Towards this, we developed **SignSight**, a hybrid framework that integrates **template matching** with **vision-language model (VLM)** reasoning to enhance the robustness of traffic sign recognition under visual distortions and adversarial conditions.
 - Template matching allowed for making templates for each potential traffic sign, to compare manipulated signs with.
 - VLM allowed us to enhance the robustness of the template matching by developing both visual and textual templates.

Restoration of adversarial samples, digitally, using advanced AI frameworks

Anchor Embedding Construction

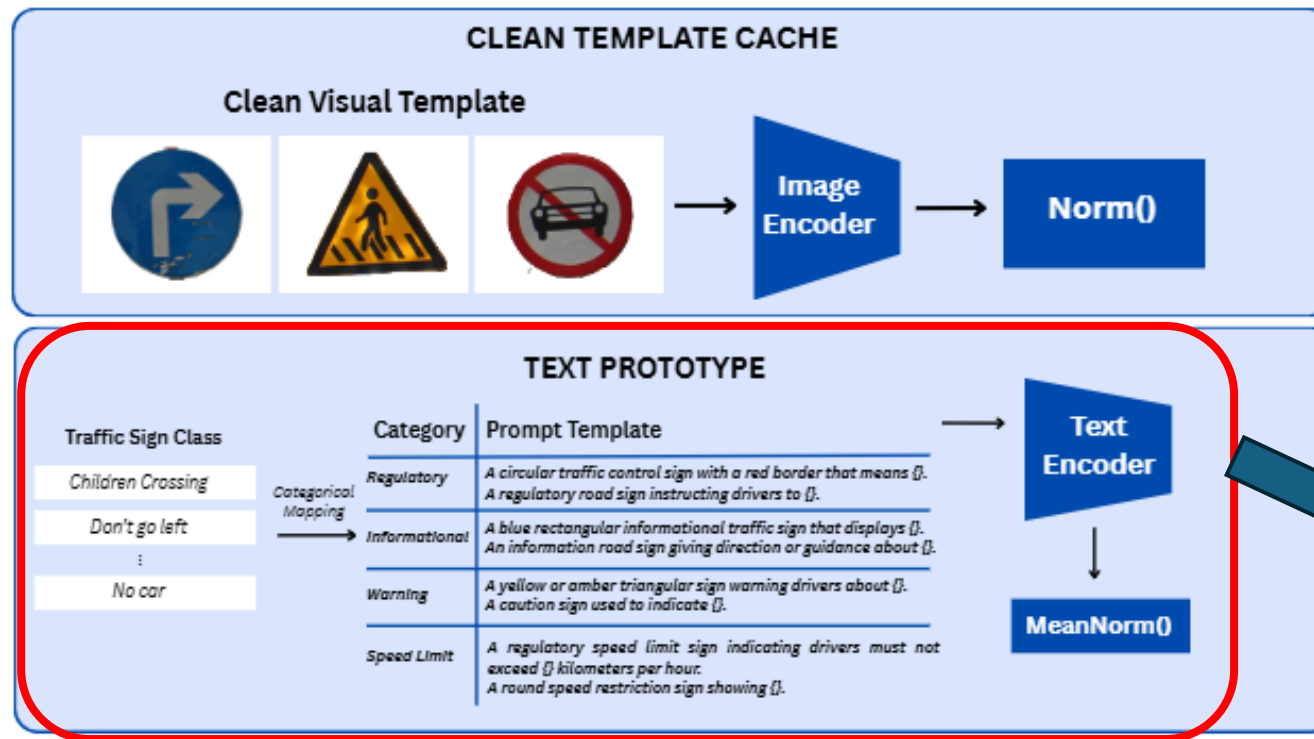


Restoration of adversarial samples, digitally, using advanced AI frameworks



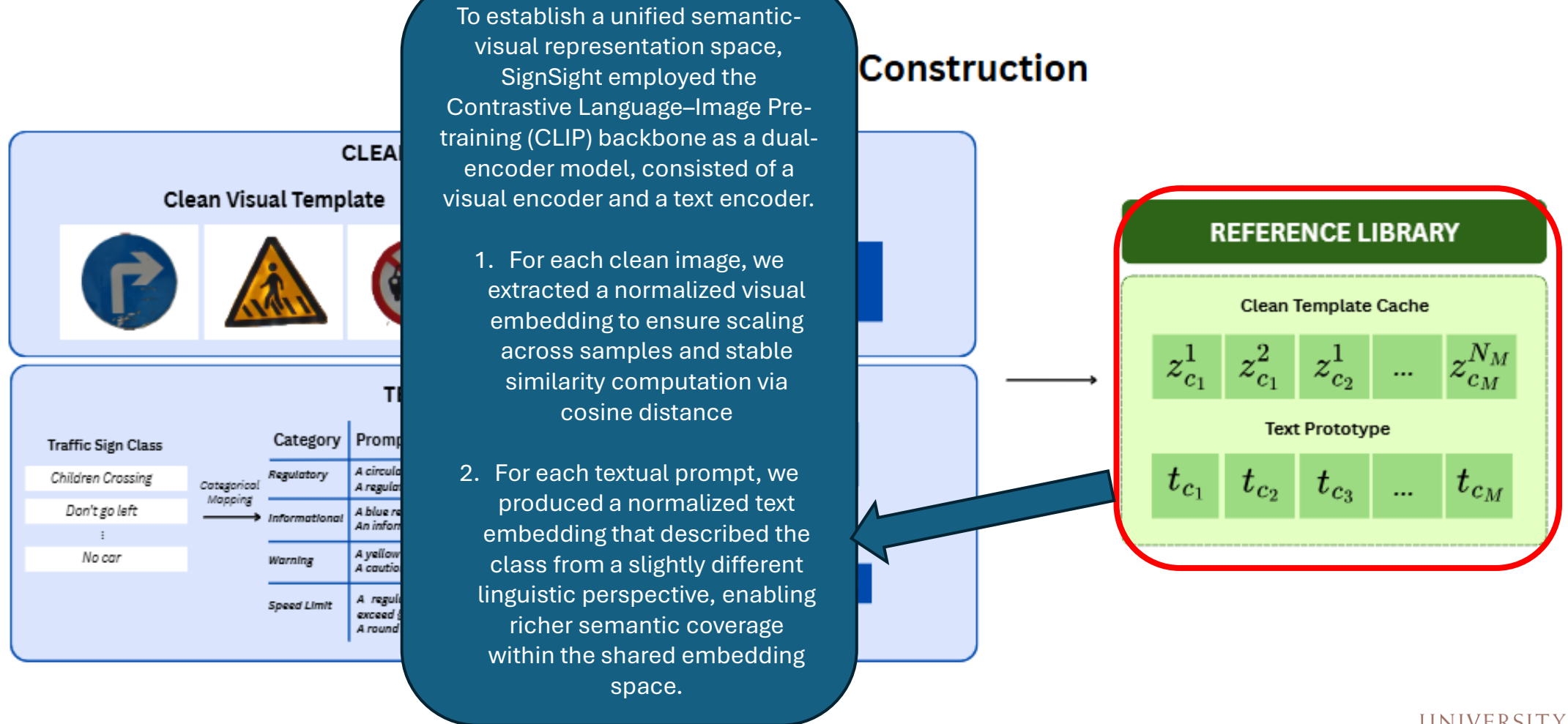
Restoration of adversarial samples, digitally, using advanced AI frameworks

Anchor Embedding Construction

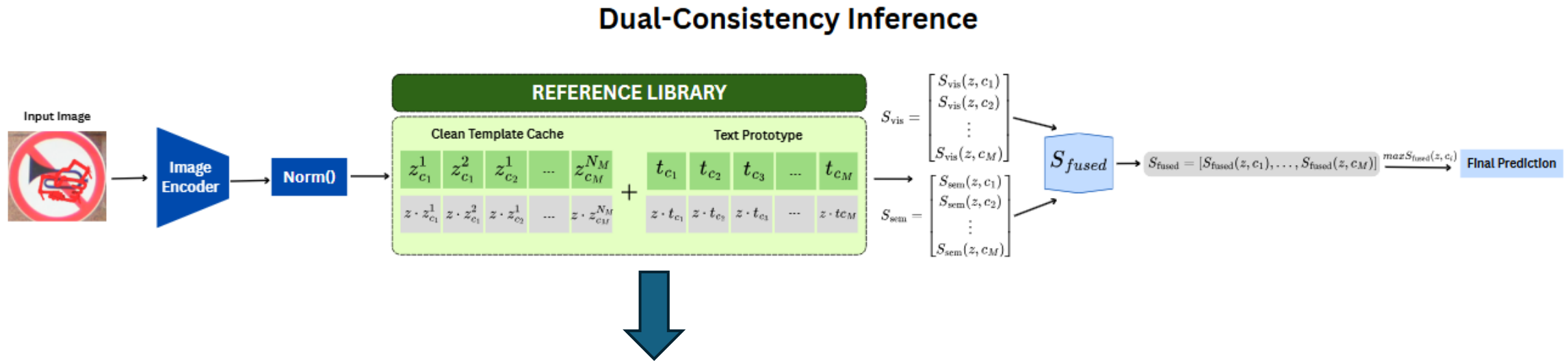


1. Correspondingly, we also employed a categorical-based prompting method to connect each traffic sign to a higher-level semantic class.
2. This captured both visual regularities and linguistic meanings
3. Each category had a small set of descriptive prompt templates, like color and shape
4. For each class, its class-specific prompt was constructed by substituting the class label into every template of its mapped category
5. This resulted in a textual foundation of the framework, representing the intended semantics of each traffic sign class.

Restoration of adversarial samples, digitally, using advanced AI frameworks



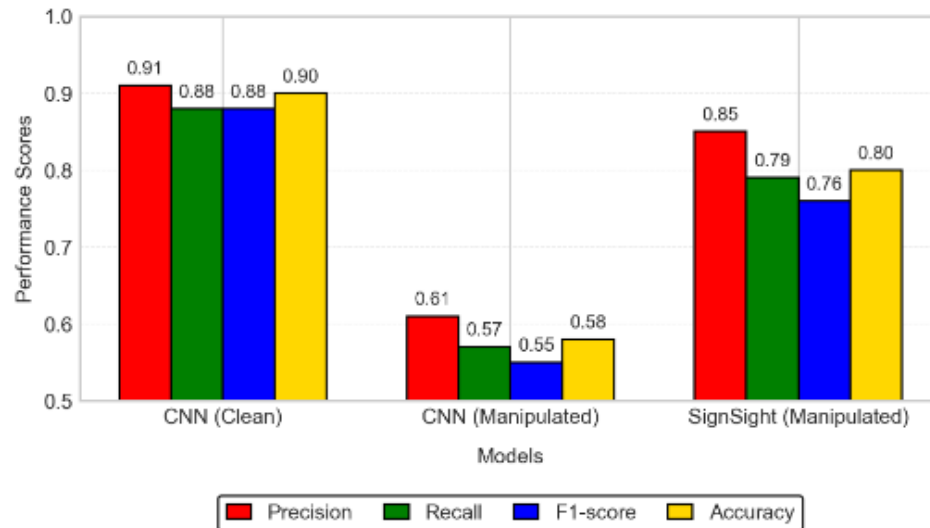
Restoration of adversarial samples, digitally, using advanced AI frameworks



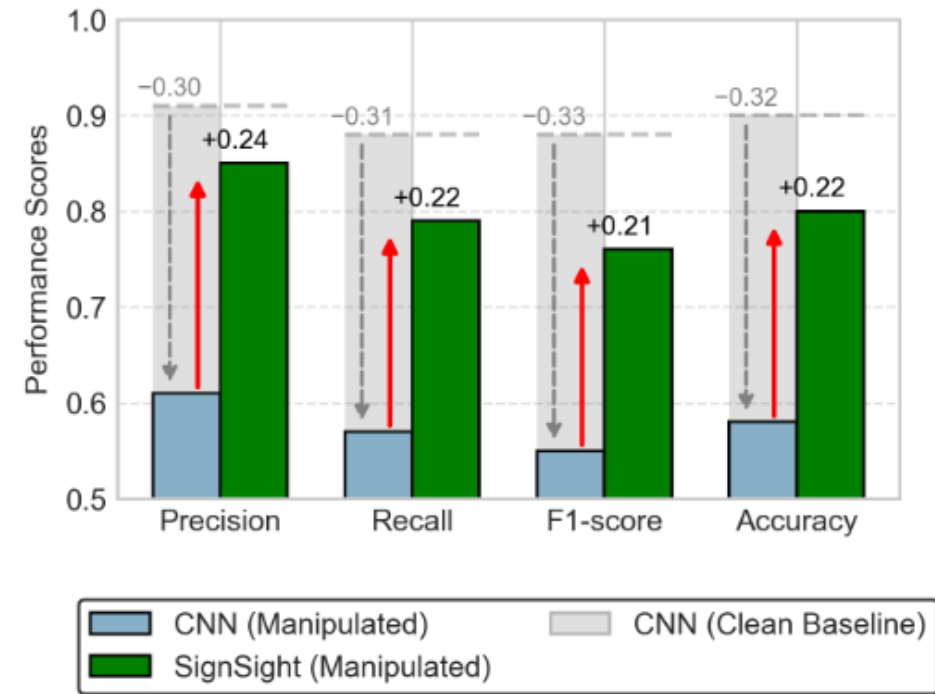
To ensure consistent recognition, SignSight introduced a dual-consistency verification mechanism that evaluated each input through two complementary pathways:

1. Visual-template consistency: Given an input image, its normalized visual embeddings were projected onto the unit hypersphere, allowing similarity to be measured purely in terms of cosine similarity.
2. Textual-semantic consistency: In parallel, for the same image, we measured how semantically aligned the input image was with the conceptual description of that class.

Is it possible to identify what the actual sign is underneath the manipulation?



(a) **Baseline comparison between CNN and SignSight.** Bar chart of precision, recall, F1-score, and accuracy showing that *SignSight* maintains higher consistency under visual distortions than the CNN baseline.



(b) **Degradation and robustness enhancement.** Gray bars indicate CNN performance on clean inputs, while blue bars show CNN performance on manipulated inputs. Red arrows highlight SignSight's robustness gain; gray arrows denote CNN degradation from clean to manipulated inputs.

Are some signs more difficult to identify than others?

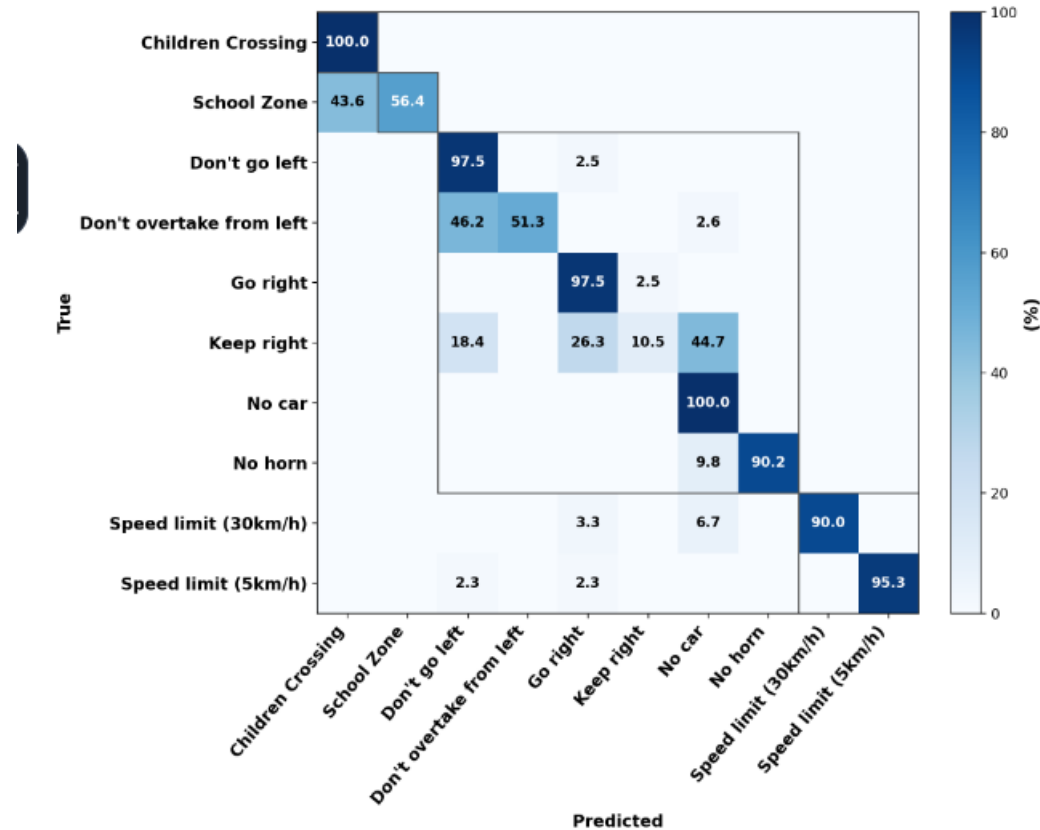


Figure 4: **Percentage confusion matrix.** Each row represents the distribution of predicted labels for a given true class. Classes are ordered by semantic category, with subtle block outlines indicating category boundaries.

Are some signs more difficult to identify than others?

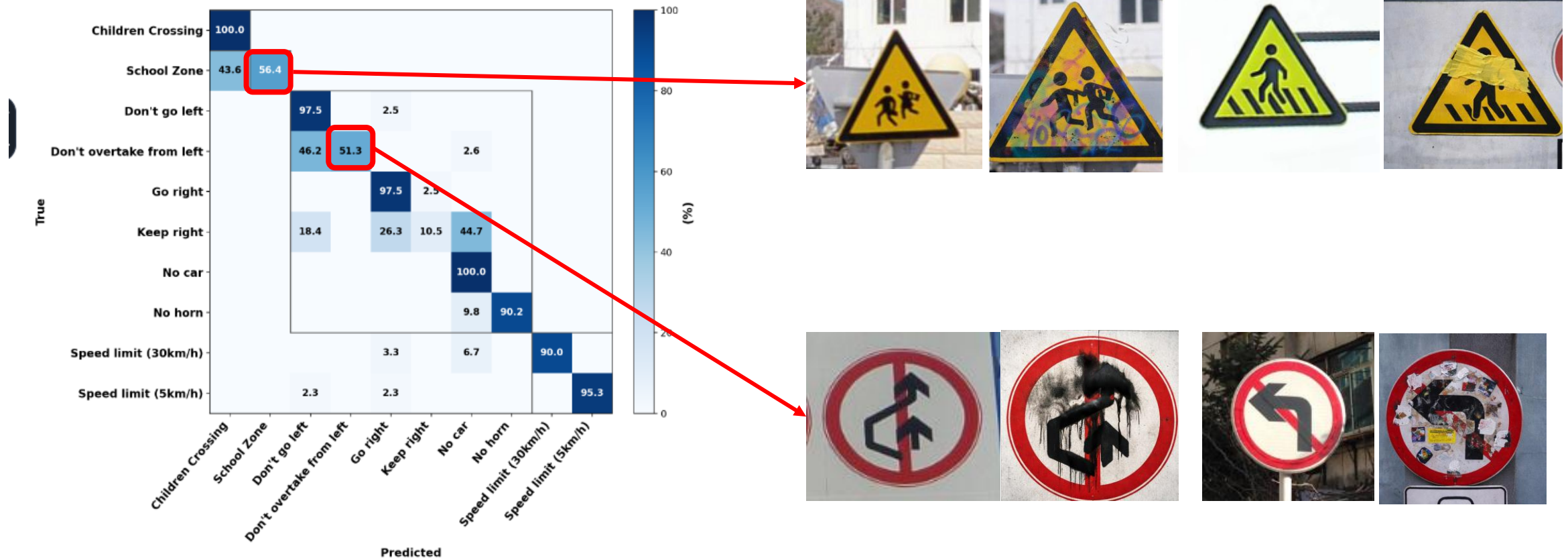


Figure 4: **Percentage confusion matrix.** Each row represents the distribution of predicted labels for a given true class. Classes are ordered by semantic category, with subtle block outlines indicating category boundaries.

What about if the level of manipulation is low vs high?



(a) **Light Manipulation.** Confident Score: 0.84

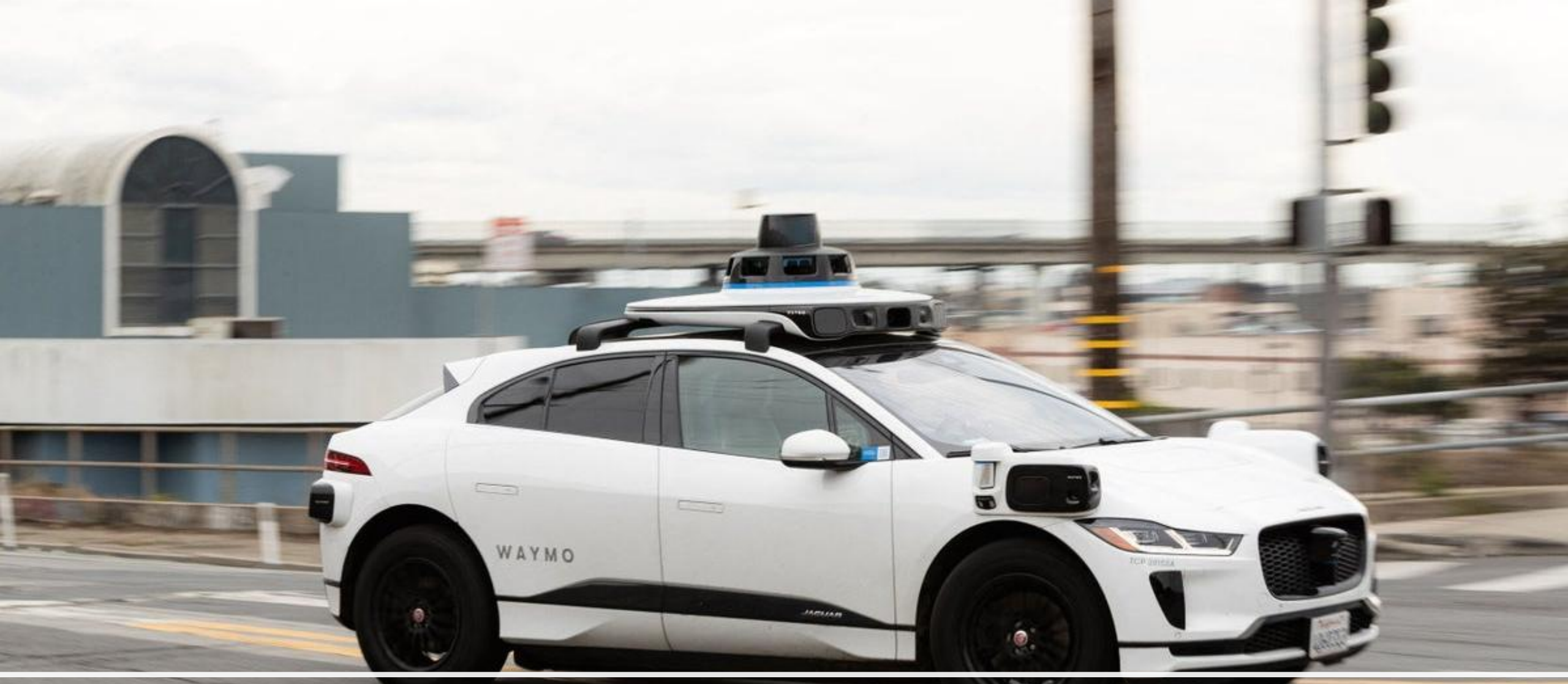


(b) **Medium Manipulation.** Confident Score: 0.61



(c) **Heavy Manipulation.** Confident Score: 0.53

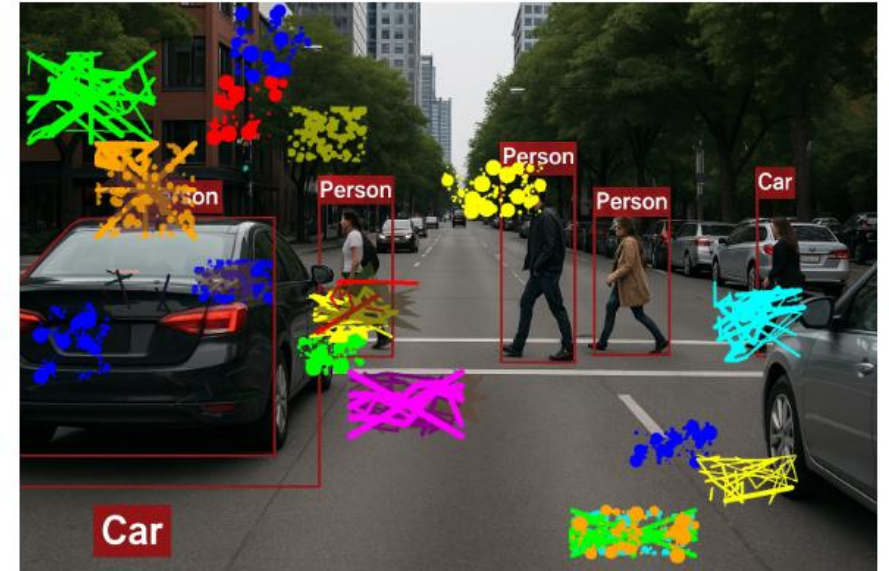
Figure 7: **Examples of physical perturbations at increasing severity levels.**



Real-time Vandalism Detection and Restoration

Sensor verification and validation under vandalism conditions

- Acts of vandalism, specifically on AV camera modules, have received attention lately.
- These can be motivated through political demonstrations, unruly crowds, or a fear of technology.
- Camera sensor malfunction can put passengers, drivers, and pedestrians at risk.
- **Research Questions*:**
 - How effectively can AI models identify vandalism on camera feeds?
 - Can it determine different types of vandalism?
 - Can it also differentiate between vandalism and other occlusions like fog, brightness, and snow?



*Exploratory research

Sensor verification and validation under vandalism conditions

- In this research, we introduce the Vandalism Identification and Sensor Trust Assessment (VISTA) framework for AV camera systems.
- VISTA consists of two specialized classifiers:
 - A binary classifier that distinguishes whether incoming camera frames are vandalized or non-vandalized
 - An Adversarial Training-enabled Multi-Class Classifier (ATMC) that identifies specific vandalism types among five categories (random, center-out, top-bottom, bottom-top, and targeted).

Sensor verification and validation under vandalism conditions

- Categories of vandalism



(a) Random

Random
spatter of
spray paint
or mud



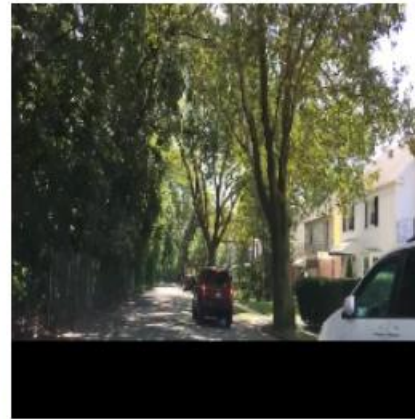
(b) Center-out

Sticker or tape
over the center
of the camera



(c) Top-Bottom

Paint or snow or
cloth on the top of
the sensor



(d) Bottom-Top

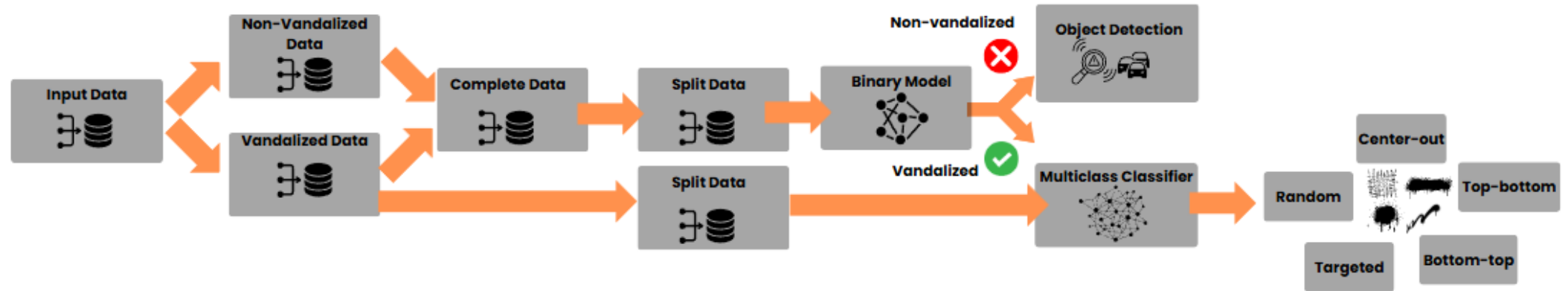
Mud, dirt, or debris
smeared on bottom
of camera



(e) Targeted

Randomly
positioned patches
over specific objects

Sensor verification and validation under vandalism conditions



Input data is camera images in perception system of the AV

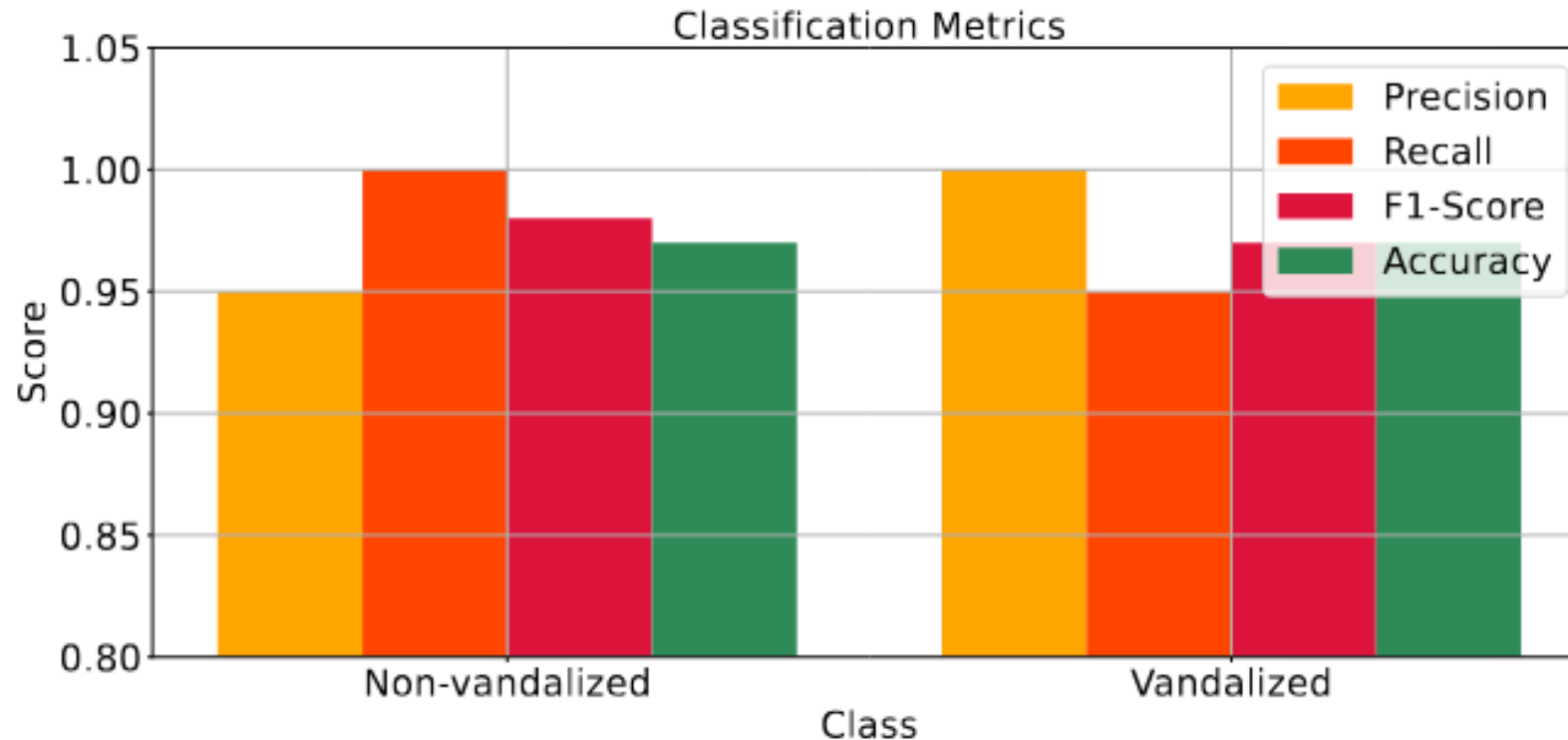
Vandalized data is synthetically generated vandalized images

The complete dataset is both vandalized and normal images

The binary model identifies whether an image is vandalized or not

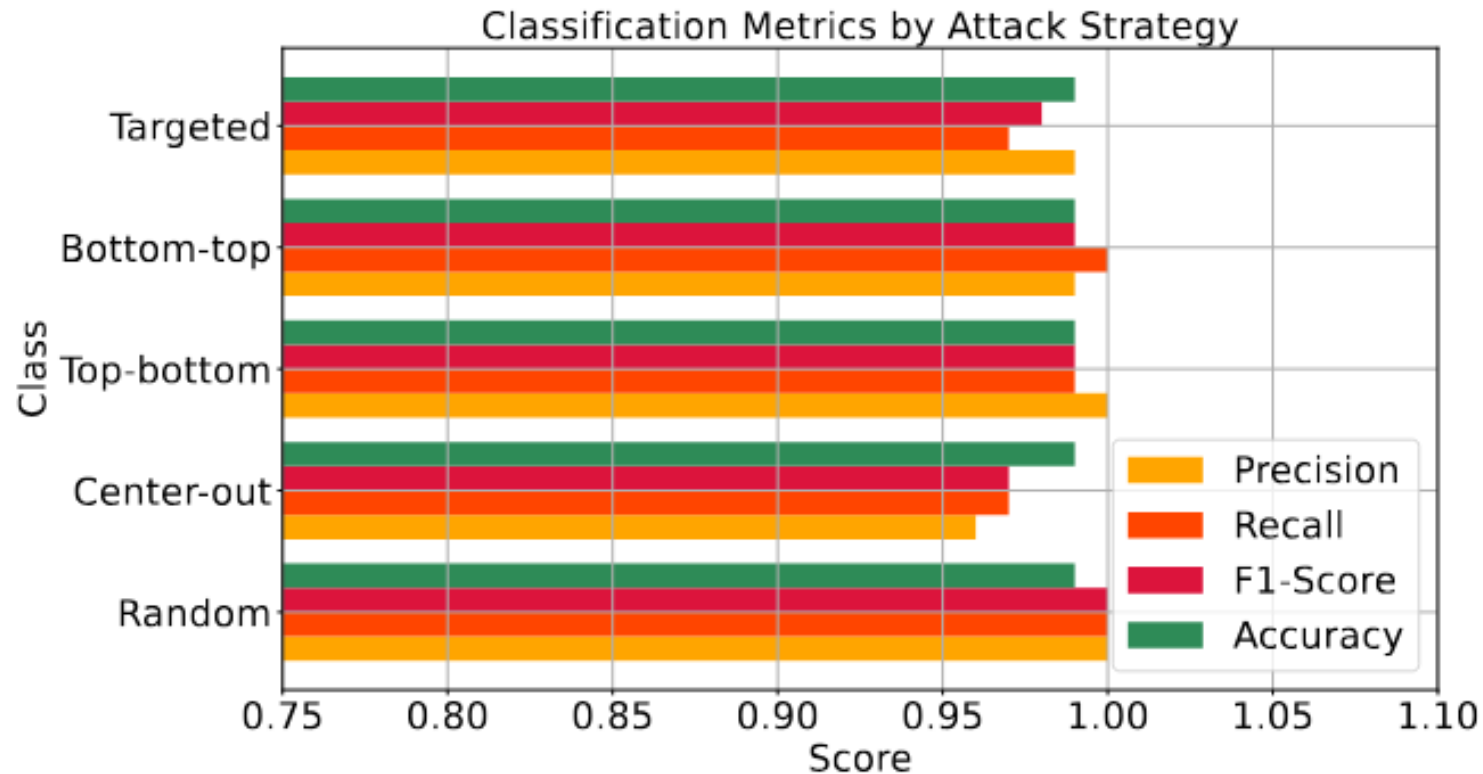
If yes, we identify the type of vandalism using the ATMC

How effectively can convolutional neural networks (CNNs) identify vandalism on camera feeds?



We observe that VISTA can identify normal and vandalized images with high accuracy, precision, recall, and F1 scores

How effectively can convolutional neural networks (CNNs) identify vandalism types on camera feeds?



We also observe that VISTA can identify different vandalism categories with high accuracy, precision, recall, and F1 scores

Can it also differentiate between vandalism and other occlusions like fog, brightness, and snow?



(a) Fog



(b) Brightness



(c) Snow



(d) Contrast



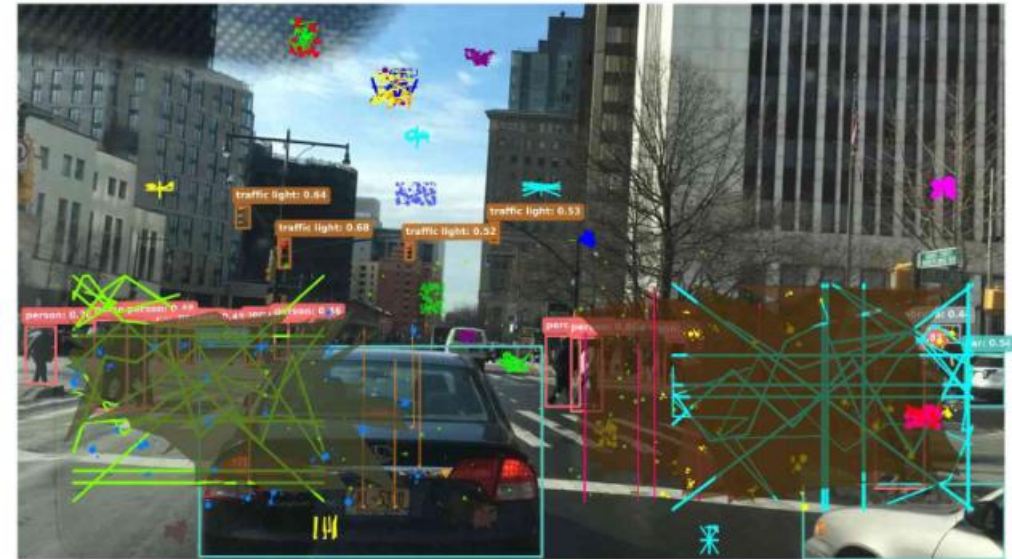
(e) Defocus Blur

Corruption Type	Images	Precision	Recall	F1-Score
Fog	20	100%	100%	100%
Brightness	20	100%	100%	100%
Snow	20	100%	100%	100%
Contrast	20	100%	100%	100%
Defocus Blur	20	100%	100%	100%
Average	100	100%	100%	100%

We synthetically generated images simulating fog, brightness, snow, contrast, and defocus. VISTA did not flag any of these images as vandalism types.

Real-time camera feed restoration under vandalism scenarios

- The previous research highlighted that AI models in AVs can distinguish vandalism and normal camera feeds
- **Research Question:**
 - Can they localize where the vandalism occurred on the camera screen?
 - Can they also recover vandalized portions of the image to help the AV operate?
 - What would be the latency of such operations?



Real-time camera feed restoration under vandalism scenarios

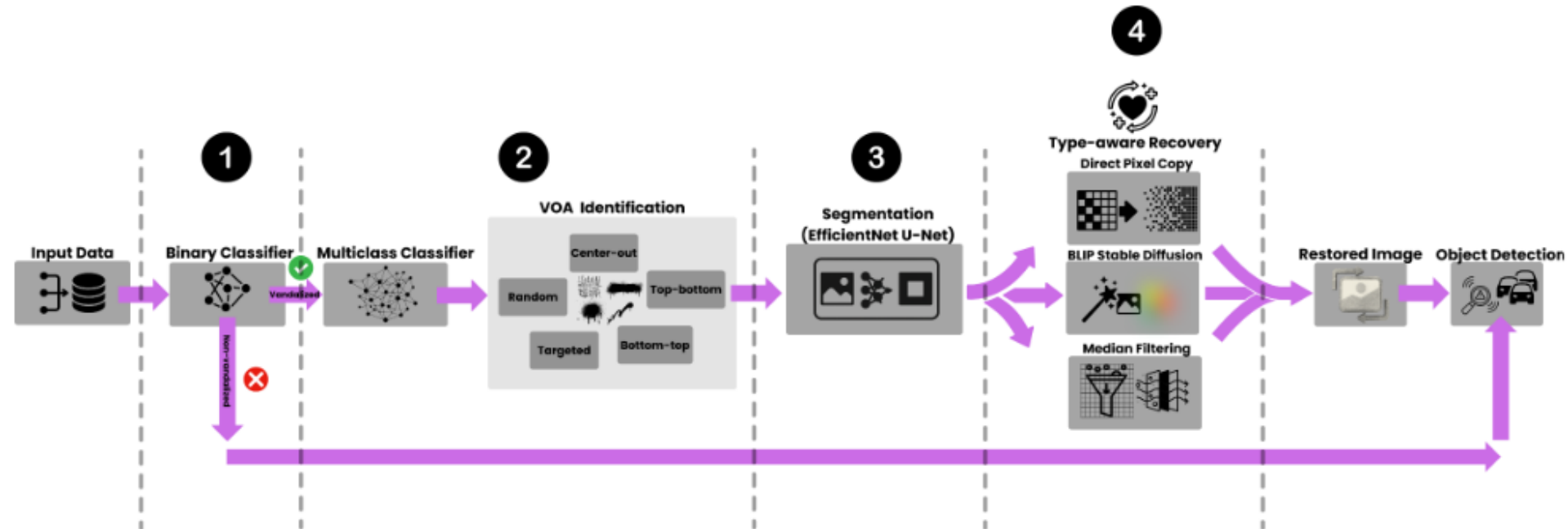
- In this research, we introduce the Recovery and Enhancement of Vandalized Images for Vision Excellence(REVIVE) framework.
- REVIVE consists of a 4-stage framework:
 - **A binary classifier** that distinguishes whether incoming camera frames are vandalized or non-vandalized
 - **An Adversarial Training-enabled Multi-Class Classifier (ATMC)** that identifies specific vandalism types among five categories (random, center-out, top-bottom, bottom-top, and targeted).
 - **EfficientNet-based U-Net** segmentation for vandalism localization
 - **Type-aware recovery** using statistical filtering, direct pixel replacement, or Bootstrapping Language-Image Pre-training (BLIP)-guided Stable Diffusion reconstruction

Real-time camera feed restoration under vandalism scenarios

Type-aware recovery:

- Statistical Parametric Filtering: We developed a sophisticated median filtering approach that dynamically adapts parameters based on image characteristics. The method employs histogram-derived adaptive thresholding, multi-scale morphological clustering analysis across kernel sizes, and conservative kernel selection with a quality-first philosophy.
- BLIP-guided Stable Diffusion: This pipeline integrates BLIP vision-language understanding with Stable Diffusion in-painting to create context-aware reconstructions.
- Direct Pixel Replacement: When original reference images are available, direct pixel replacement provides perfect reconstruction with minimal error.

Real-time camera feed restoration under vandalism scenarios



Input data is camera images in perception system of the AV

Vandalized data is synthetically generated vandalized images

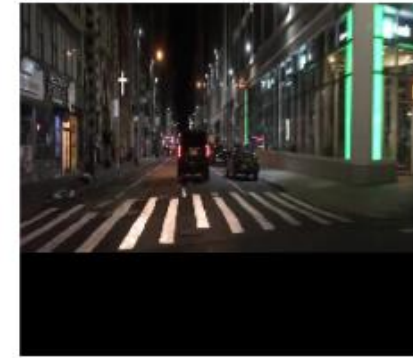
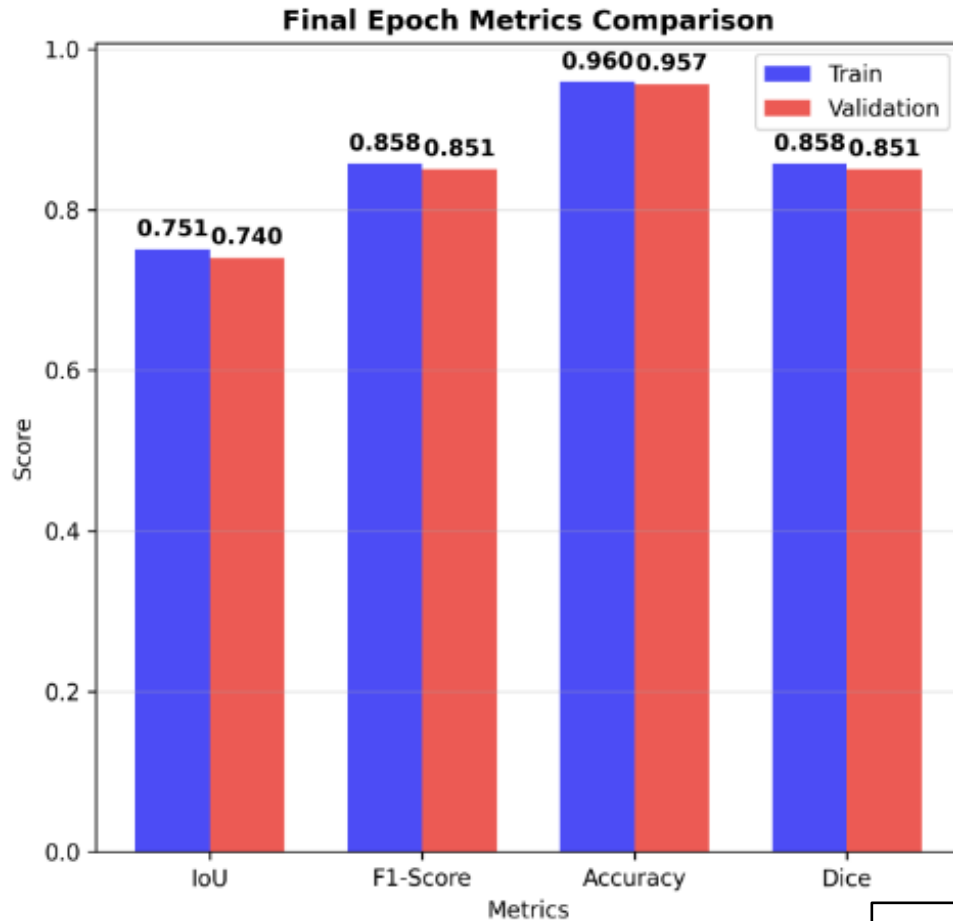
The binary model identifies whether an image is vandalized or not

If yes, we identify the type of vandalism using a multi-class classifier

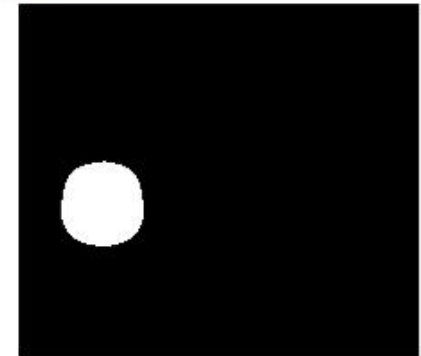
Next, we employ an EfficientNet-based U-Net architecture to localize vandalism on the feed.

Finally, we restore the vandalized images using type-aware recovery

Can AI Models localize where the vandalism occurred on the camera screen?



(a) Bottom-top (IoU: 0.992)



(b) Targeted (IoU: 0.852)

Yes, they can, and with pretty good performance across accuracy, precision, recall, F1, and Intersection over Union (IoU) [common metric used for object detection]

Can AI Models also recover vandalized portions of the image to help the AV operate?



Figure 8: Bottom-top VOA recovery showing the complete pipeline (left to right: Original, Vandalized, GenAI Recovery, Pixel Copy Recovery). The Original image is the reference for all subsequent recovery examples.



(a) Center-out VOA recovery (Vandalized → GenAI → Pixel Copy)



(b) Targeted VOA recovery (Vandalized → GenAI → Pixel Copy)

Yes, they can, but it is necessary to use the right recovery method for each.

Can AI Models also recover vandalized portions of the image to help the AV operate?



Figure 8: Bottom-top VOA recovery showing the complete pipeline (left to right: Original, Vandalized, GenAI Recovery, Pixel Copy Recovery). The Original image is the reference for all subsequent recovery examples.



(a) Center-out VOA recovery (Vandalized → GenAI → Pixel Copy)



(b) Targeted VOA recovery (Vandalized → GenAI → Pixel Copy)

Vandalism at the periphery of the camera feed will work with GenAI methods like our BLIP-guided Diffusion model, since there is no contextual information missing

Can AI Models also recover vandalized portions of the image to help the AV operate?



Figure 8: Bottom-top VOA recovery showing the complete pipeline (left to right: Original, Vandalized, GenAI Recovery, Pixel Copy Recovery). The Original image is the reference for all subsequent recovery examples.



(a) Center-out VOA recovery (Vandalized → GenAI → Pixel Copy)



(b) Targeted VOA recovery (Vandalized → GenAI → Pixel Copy)

Vandalism at the periphery of the camera feed will work with GenAI methods like our BLIP-guided Diffusion model, since there is no contextual information missing

Can AI Models also recover vandalized portions of the image to help the AV operate?



Figure 8: Bottom-top VOA recovery showing the complete pipeline (left to right: Original, Vandalized, GenAI Recovery, Pixel Copy Recovery). The Original image is the reference for all subsequent recovery examples.



(a) Center-out VOA recovery (Vandalized → GenAI → Pixel Copy)



(b) Targeted VOA recovery (Vandalized → GenAI → Pixel Copy)

Vandalism at the periphery of the camera feed will work with GenAI methods like our BLIP-guided Diffusion model, since there is no contextual information missing

Can AI Models also recover vandalized portions of the image to help the AV operate?



Figure 8: Bottom-top VOA recovery showing the complete pipeline (left to right: Original, Vandalized, GenAI Recovery, Pixel Copy Recovery). The Original image is the reference for all subsequent recovery examples.



(a) Center-out VOA recovery (Vandalized → GenAI → Pixel Copy)



(b) Targeted VOA recovery (Vandalized → GenAI → Pixel Copy)

Vandalism at the periphery of the camera feed will work with GenAI methods like our BLIP-guided Diffusion model, since there is no contextual information missing

Can AI Models also recover vandalized portions of the image to help the AV operate?



Figure 8: Bottom-top VOA recovery showing the complete pipeline (left to right: Original, Vandalized, GenAI Recovery, Pixel Copy Recovery). The Original image is the reference for all subsequent recovery examples.



(a) Center-out VOA recovery (Vandalized → GenAI → Pixel Copy)



(b) Targeted VOA recovery (Vandalized → GenAI → Pixel Copy)

Vandalism in the middle of the camera feed will work with pixel copy since vandalism destroys scene context, and GenAI may not successfully recover it

Can AI Models also recover vandalized portions of the image to help the AV operate?



Figure 8: Bottom-top VOA recovery showing the complete pipeline (left to right: Original, Vandalized, GenAI Recovery, Pixel Copy Recovery). The Original image is the reference for all subsequent recovery examples.



(a) Center-out VOA recovery (Vandalized → GenAI → Pixel Copy)



(b) Targeted VOA recovery (Vandalized → GenAI → Pixel Copy)

Vandalism in the middle of the camera feed will work with pixel copy since vandalism destroys scene context, and GenAI may not successfully recover it

Can AI Models also recover vandalized portions of the image to help the AV operate?



Figure 8: Bottom-top VOA recovery showing the complete pipeline (left to right: Original, Vandalized, GenAI Recovery, Pixel Copy Recovery). The Original image is the reference for all subsequent recovery examples.



(a) Center-out VOA recovery (Vandalized → GenAI → Pixel Copy)



(b) Targeted VOA recovery (Vandalized → GenAI → Pixel Copy)

Vandalism in the middle of the camera feed will work with pixel copy since vandalism destroys scene context, and GenAI may not successfully recover it

Can AI Models also recover vandalized portions of the image to help the AV operate?



Figure 8: Bottom-top VOA recovery showing the complete pipeline (left to right: Original, Vandalized, GenAI Recovery, Pixel Copy Recovery). The Original image is the reference for all subsequent recovery examples.



(a) Center-out VOA recovery (Vandalized → GenAI → Pixel Copy)



(b) Targeted VOA recovery (Vandalized → GenAI → Pixel Copy)

Vandalism in the middle of the camera feed will work with pixel copy since vandalism destroys scene context, and GenAI may not successfully recover it

Can AI Models also recover vandalized portions of the image to help the AV operate?



(a) Random VOA recovery (Vandalized $\rightarrow$ Median Filtering)

Vandalism randomly sprayed in the camera feed would need median filtering to be filtered out

What would be the latency of such operations?

Method	SSIM	PSNR	Time(s)	mAP
Stable Diffusion	0.69-0.98	10.9-26.4dB	7.0	0.78
Pixel Replace	1.00	∞	0.001	0.89
Median Filter	0.13	11.8dB	0.010	0.67

The inference fastest latency is Pixel Replacement, followed by Median Filtering. Stable Diffusion, due to its intense computational requirements, has the highest latency.

Conclusion

- AI Resiliency in AV Perception Systems is critical for efficient operations, especially as we head to Level 6G
- Continued efforts must be made to ensure perception systems actively predict and avoid areas where vandalized traffic signs and accidents are potential risks.

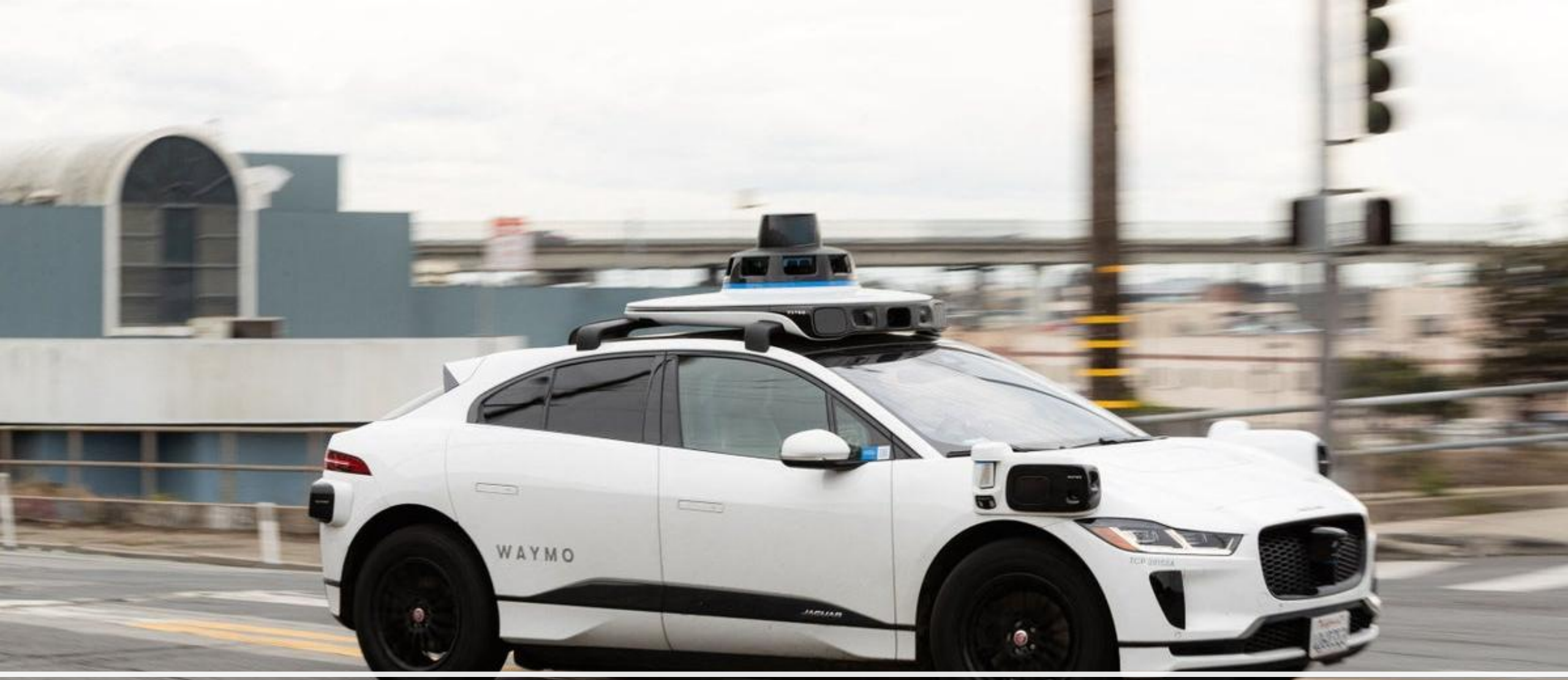




Future Work

We plan to augment the developed frameworks by:

- Improving latencies for camera feed restoration involving GenAI when encountering vandalism
- Ensuring consistent performance across different vandalism categories
- Making our traffic sign recognition solution portable across geographical regions.
- Performing practical testing on autonomous vehicular testbeds that simulate real-world driving conditions.



Thank you! Questions?



If you think of
something
later.....

Tapadhir Das

Email: tdas@pacific.edu

Website: tapadhirdas.com