

Understanding Human-AI Interactions in Complex AI-Driven Autonomous Systems: Implications of Trust, Interpretability, and Ethical Reliability in Critical Missions

Vineetha Menon, Daniel Pham, Howard Chen, Bryan Mesmer, Kristin Weger,
and Nicholas Loyd

Motivation

- Artificial Intelligence (AI) is part of our lives.
- AI in Combat Search and Rescue (CSAR)
- CSAR Scenario: Hostage Rescue
 - High urgency
 - Vast, dynamic geographical area
 - Obscured ground vision
- CSAR Objectives
 - Gather information on environment quickly and accurately
 - Determine optimal path to hostages and back to safe zones
 - Correctly identify hostages and obstacles
 - Minimal to no casualties
- Can AI systems be trusted to make the right decisions in a hostage rescue scenario?

Introduction

- AI-driven/Assistive Autonomous (AI/AA) systems present in CSAR
 - Unmanned Aerial Vehicles (UAVs)
- Challenge: AI integration in critical missions
 - What are the implications of risk?
- Challenge: Human-AI collaboration under uncertainty
 - How can AI-driven autonomous systems be employed in hazardous environments?
- Challenge: Safeguard lives of soldiers in complex, dynamic, & high-risk environments

How can XAI improve human-AI trust and teaming?

Research Objectives

AI Tasks:

- Map optimal paths to various hostage location points
- Safely navigate optimal paths and extract hostages with minimal to no casualties

Challenges:

- Traditional black-box AI/ML techniques
- Real-world data variations: Obstacles, lighting, shadows,...
 - Trust and liability challenges
 - Ethical reliability concerns

Current Objective:

- Develop explainable AI framework

Long Term Objectives:

- Enhance interpretability and trust calibration
- Validate ethical reliability

Simulation Environment

- Unity simulation for CSAR environment
- Human agent and drone agent (AI/AA system)
- 61 participants, split between drone and human agents
- Goal: Rescue all 16 hostages
- Two modes are toggleable (Human or Drone)
- Limitations
 - Human agent can only navigate along the ground
 - Drones automatically follow a static path

Human-AI/AA System Interaction

Human Mode

- Ground movement limited
- View rotation independent
- Terrain constraints
- Not all hostages visible or reachable



Figure 1: Simulation in Human Perspective

Drone Mode

- Movement in the air
- Sees further
- Easily passes over obstacles



Figure 2: Simulation in Drone Perspective

This presentation focuses on how the drone explains to the human what the drone detects.

Proposed PCA-Laplacian-CNN System

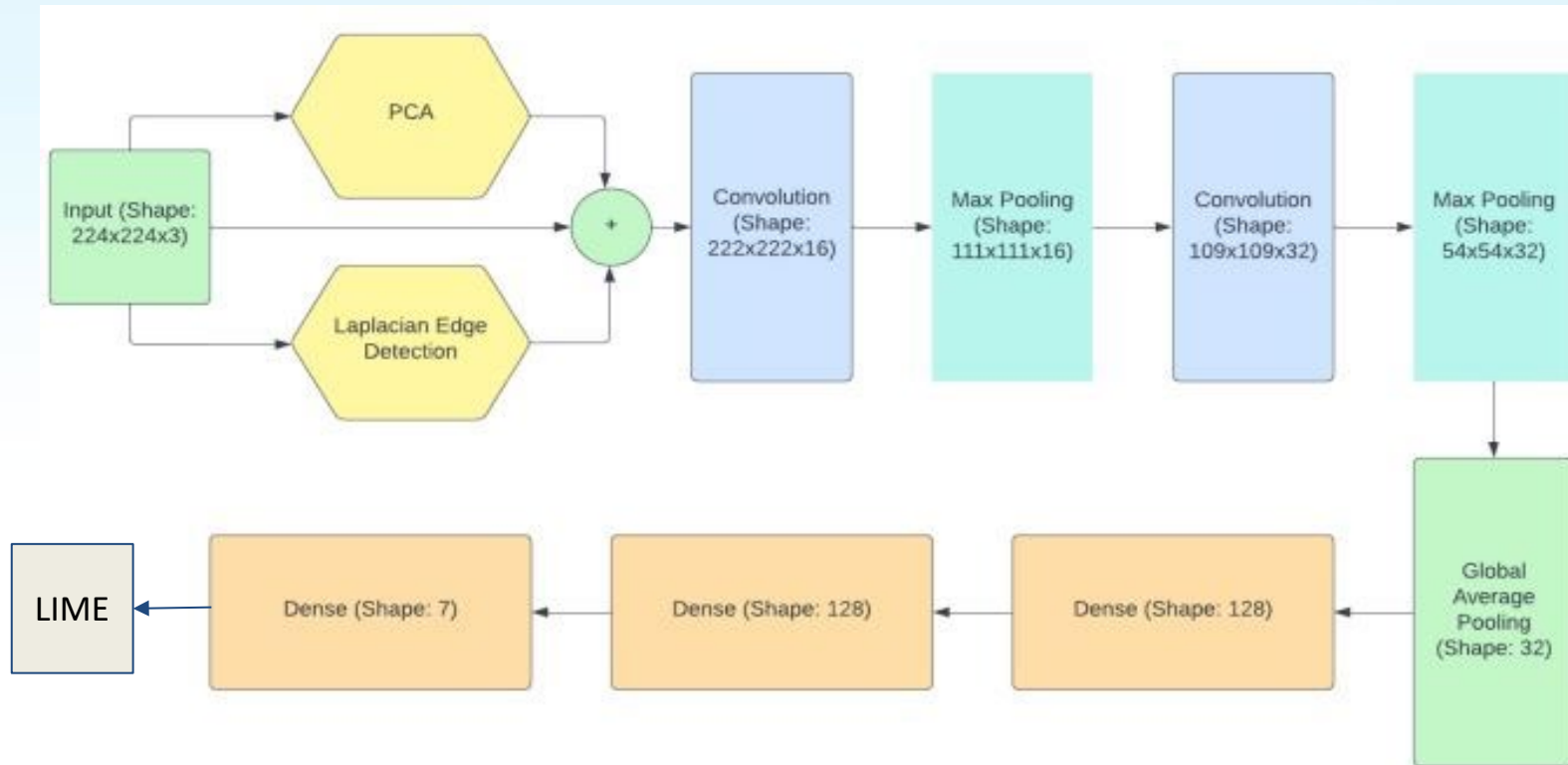


Figure 3. The proposed PCA-Laplacian-CNN drone AI/AA system architecture

- Feature Extraction: Principal Components Analysis (PCA)
- Edge Detection: Laplacian masks
- Target detection: CNN classifier
- Model Interpretability: LIME

Data Collection and Preprocessing

- 7 classes, 42000 images
 - Drones
 - Fountains
 - Grass
 - Hostages
 - Houses
 - Trailers
 - Trees
- 70:30 train-test split
 - 56% for training
 - 14% for validation
 - 30% for testing

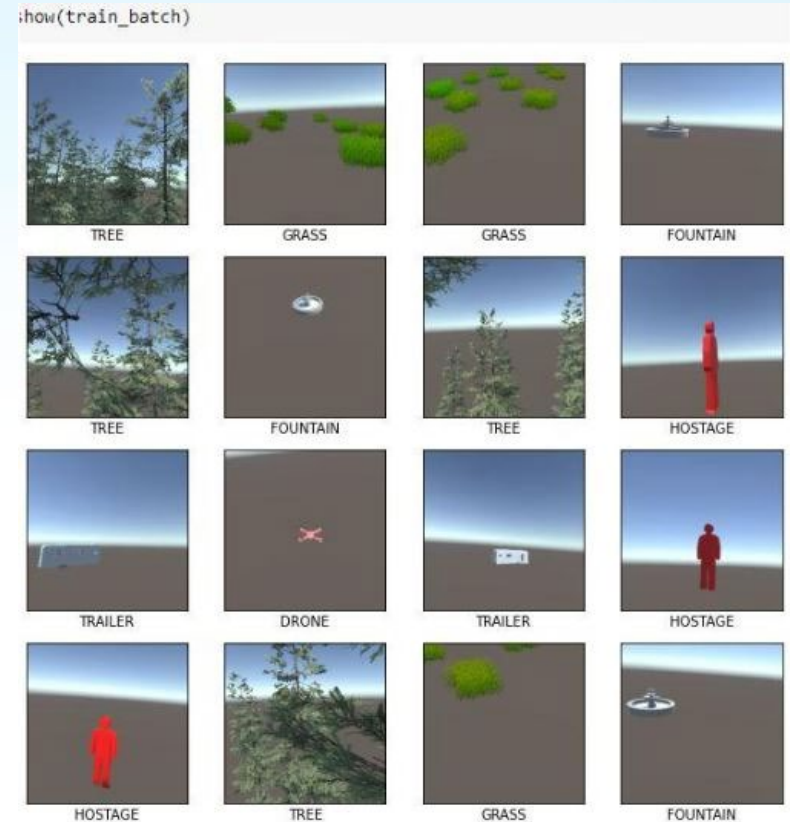


Figure 4: Samples of the images used for training and testing the model

Results and Analysis

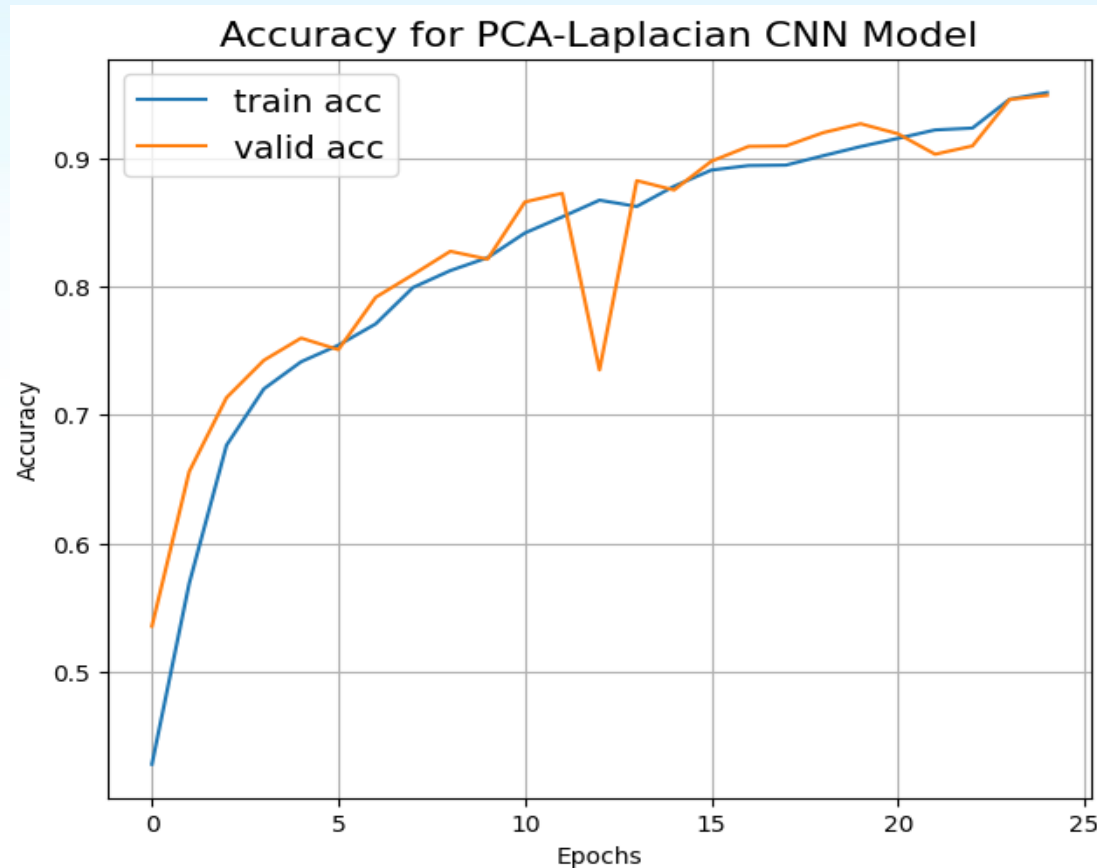


Figure 5: Training Accuracy vs Validation Accuracy

Overall Classification accuracy = 95%

Classification Results

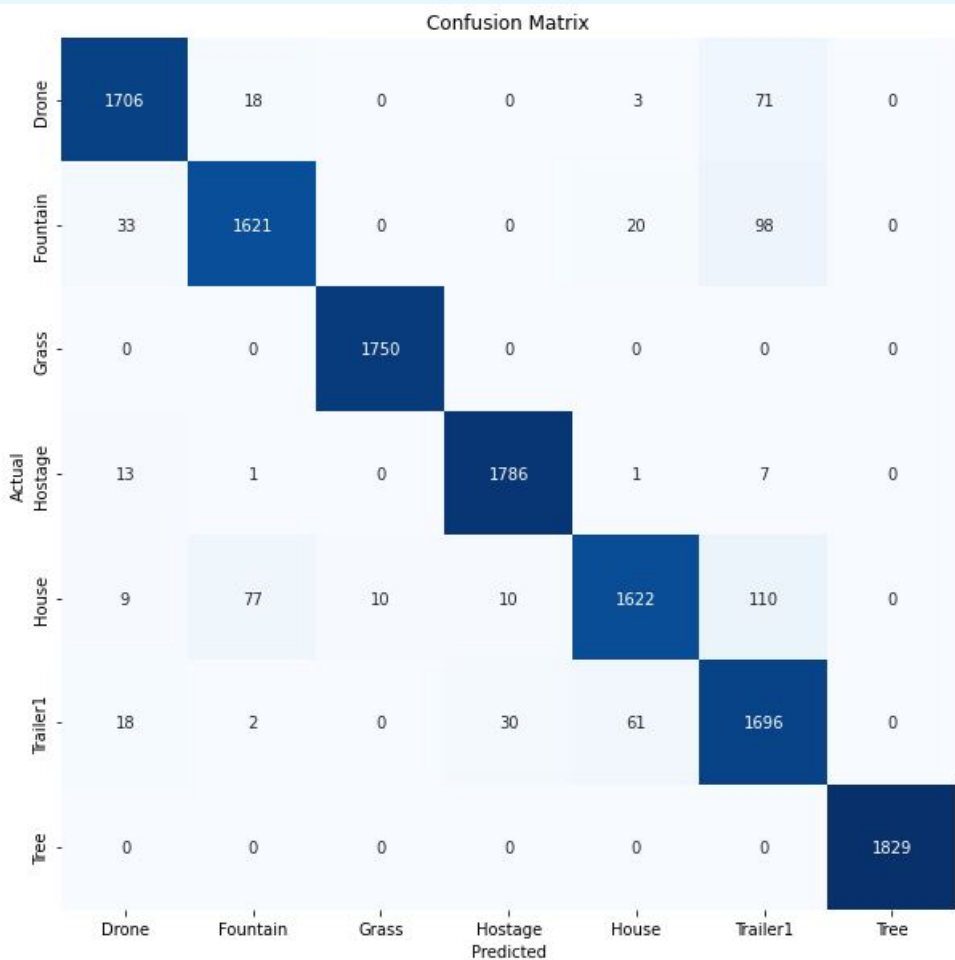


TABLE I
CLASSIFICATION PERFORMANCE OF THE PROPOSED
PCA-LAPLACIAN-CNN MODEL

Class	Precision	Recall	F1-Score
Drone	96%	95%	95%
Fountain	94%	91%	93%
Grass	99%	100%	100%
Hostage	98%	99%	98%
House	95%	88%	92%
Trailer	86%	94%	90%
Tree	100%	100%	100%
Overall Accuracy			95%
Weighted Average	95%	95%	95%

Figure 6: Confusion Matrix of Classwise Accuracies

Explainable AI (XAI): How the AI Model Learns



Figure 7: Local Interpretable Model-Agnostic Explanations (LIME) outlines regions, omits irrelevant regions, and isolates regions of interest

Table 2: For this image, LIME lists the top five guesses; the first guess is correct (hostage)

Class	Array index
0	Drone
1	Fountain
2	Grass
3	Hostage
4	House
5	Trailer
6	Tree

array([3, 5, 6, 4, 2])

Conclusion

- Overall classification accuracy: 95%
- Thus the system proved to be efficient, provided interpretability metrics, provided a means for ethical reliability validation of the AI model's decision making capabilities, and enables enhanced operator trust.
- High accuracy identification with little error means the AI system understands the objects of interests and what they look like.
- AI is a feasible option for hostage rescue scenarios and explainability models may help in building user trust
- Future Work:
 - More realistic data will be used
 - Other architectures will be explored
 - AI Policy and training recommendations
 - Human studies to understand trust improvements

Funding Acknowledgement:

Research was sponsored by DAC and was accomplished under Cooperative Agreement Number W911NF-23-2-0200. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Thank you!