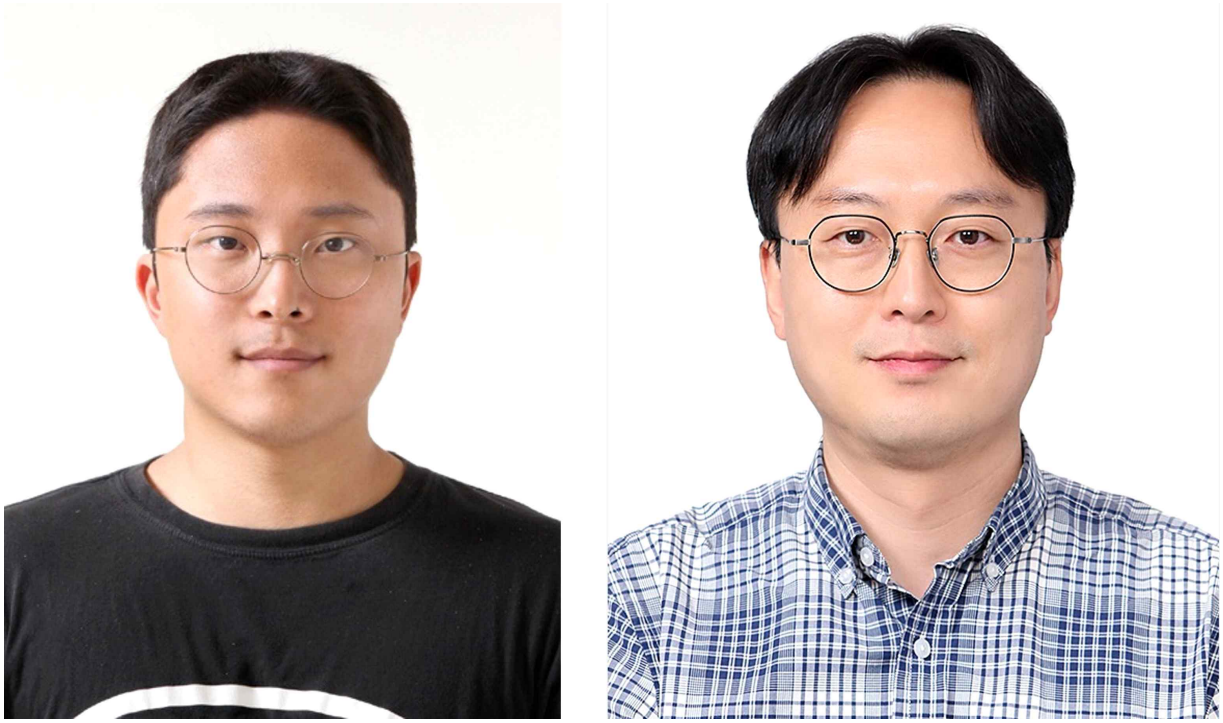
"Pedestrians as Signs? We've Discovered a Blind Spot in Self-Driving Cars" GIST developed a new algorithm that exposes vulnerabilities in self-driving car recognition systems

- Professor SeungJun Kim's research team from the Department of AI Convergence presents a new attack technique that reflects the hierarchical structure between objects such as pedestrians, lanes, and traffic lights... The recognition rate plummeted from 95.3% to 3.23%, demonstrating a 3.4x higher attack success rate

- Focusing on the vulnerability of deep learning-based visual recognition models to subtle data modifications (adversarial attacks)... The research is expected to contribute to enhancing the safety of various AI fields, including intelligent transportation systems (ITS), smart city security, robotics, and national defense.

- Scheduled for publication in 《IEEE Robotics and Automation Letters》 and presentation at ICRA 2026



▲ (From left) GIST Department of AI Convergence doctoral student Gwangbin Kim and Professor SeungJun Kim

The Gwangju Institute of Science and Technology (GIST, President Kichul Lim) announced that a research team led by Professor SeungJun Kim of the Department of AI Convergence has developed a new "adversarial attack*" algorithm capable of diagnosing security vulnerabilities in the visual recognition systems used in autonomous vehicles.

Autonomous vehicles must accurately recognize various objects on the road, such as pedestrians, vehicles, lanes, and traffic lights, in real time to operate safely. The core technology that makes this possible is the semantic segmentation model*.

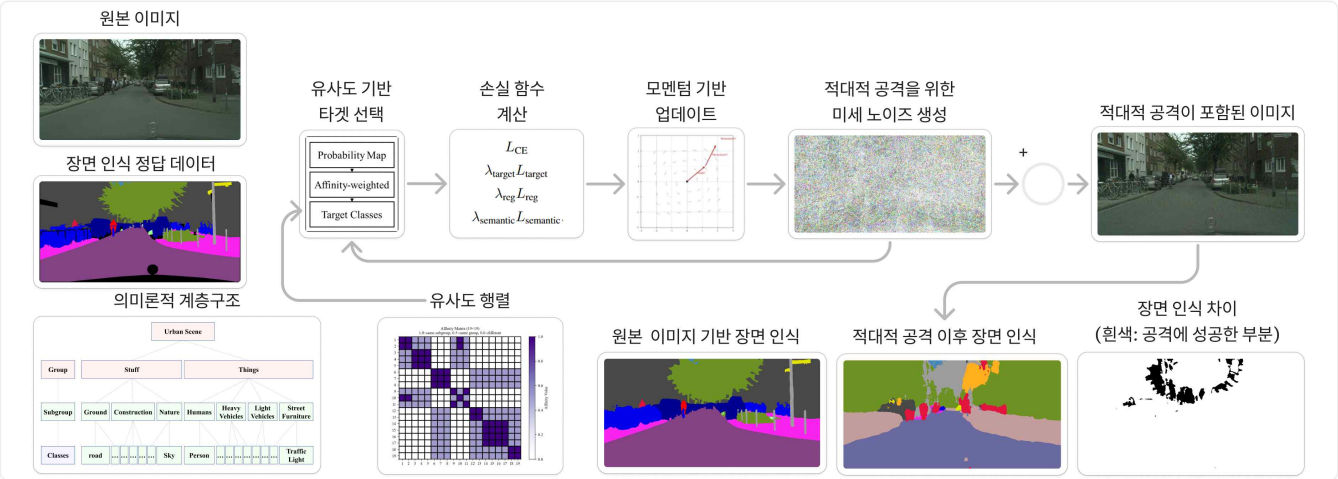
This research is expected to contribute to the development of safer and more reliable autonomous driving systems by proactively identifying potential attacks on the semantic segmentation model.

* adversarial attack: A technique that deliberately manipulates input data to cause an AI model, especially a deep learning-based model, to make incorrect judgments. For example, even subtle modifications to the pixel values of an image, imperceptible to humans, can lead the model to misrecognize an entirely different object, posing a significant threat to security and reliability.

* semantic segmentation model: An AI technology that segments every pixel in an image into specific semantic units, such as objects or backgrounds. For example, when autonomous vehicles recognize road conditions, they accurately distinguish individual objects such as lanes, pedestrians, vehicles, and traffic lights at the pixel level, enabling safe driving.

Recently, deep learning-based models have been found to be vulnerable to sophisticated data manipulation known as adversarial attacks. Even subtle changes, barely discernible to the human eye, can cause deep learning-based models to misidentify pedestrians as road signs or ignore traffic lights, posing a safety risk.

In response, the research team focused on the structural characteristics of semantic segmentation models and developed a new attack technique that exposes vulnerabilities with much greater precision than existing attack methods.



▲ Overview of a semantic hierarchy-based adversarial attack technique. The adversarial attack is generated by selecting targets using a semantic hierarchy. While the generated image is indistinguishable from the original image to the naked eye, the AI model perceives it as a completely different scene, resulting in a high attack success rate.

While existing adversarial attacks used to diagnose vulnerabilities in autonomous vehicles simply induce random errors, this study simulates more threatening error scenarios in real-world driving situations by reflecting the hierarchical structure of different objects, such as pedestrians, lanes, and traffic lights.

Experimental results showed that the newly developed technique by the research team disrupted autonomous vehicle visual recognition models by a factor of 3.4 compared to existing methods, demonstrating a significantly higher attack success rate in both day and night driving environments.

Notably, in tests simulating autonomous driving situations, the recognition rate for key objects, such as traffic lights, pedestrians, and lanes, plummeted from 95.3% to 3.23%, demonstrating the vulnerability of the semantic segmentation model.

This study is significant not only for presenting a powerful attack technique but also for identifying the semantic unit hierarchy where the model is particularly vulnerable. This can serve as a valuable foundation for designing safer and more reliable autonomous driving recognition systems in the future.

The research team expects this achievement to contribute to enhancing the safety of diverse artificial intelligence (AI) applications beyond autonomous driving, including intelligent transportation systems (ITS), smart city security, robotics, and national defense.

Professor SeungJun Kim stated, "This study provides a tool for systematically analyzing the fundamental vulnerabilities of autonomous vehicle visual recognition systems. By preemptively analyzing and diagnosing the most vulnerable scenarios, we can develop more robust defense strategies and ultimately significantly improve the safety and reliability of autonomous vehicles."

This research, supervised by Professor SeungJun Kim of the Department of AI Convergence at GIST and led by doctoral student Gwangbin Kim as first author, was supported by the GIST-MIT Joint Research Project, the National IT Industry Promotion Agency for Artificial Intelligence Graduate School Support Program, the National Research Foundation of Korea's Mid-Career Researcher Support Program, and the Korea Agency for Infrastructure Technology Advancement (KAIT).

The research results were published in the August 2025 issue of the international academic journal 《IEEE Robotics and Automation Letters》 and will be presented at the IEEE International Conference of Robotics and Automation (ICRA), the world's most prestigious robotics conference, scheduled to be held in Vienna, Austria in June next year.

Meanwhile, GIST stated that this research achievement considered both academic significance and industrial applicability, and that technology transfer inquiries can be made through the Technology Commercialization Center (hgmoon@gist.ac.kr).

GIST
Since 1993