

TAHAI Workshop KI Sicherheit

HTW Berlin



Rudolf Hoffmann / 13.03.2025



**Hochschule für Technik
und Wirtschaft Berlin**

University of Applied Sciences



Robustheitsanalyse für Vision-Modelle

HTW Berlin

Rudolf Hoffmann / 13.03.2025



Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Agenda

- 1 Projektüberblick
- 2 Robustheit von Vision Modellen
- 3 Deepbench: Vertrauen durch Robustheit
- 4 Use Cases und Corruptions
- 5 Domäne-Spezifische Analyse
- 6 Evaluation

Projektüberblick



imagen-3.0-generate-002

Arbeitspaket A7

- Erstellung eines Frameworks zur Risiko- und Robustheitsanalyse von KI-Modellen

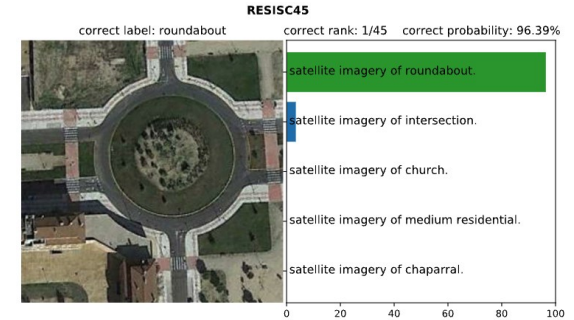
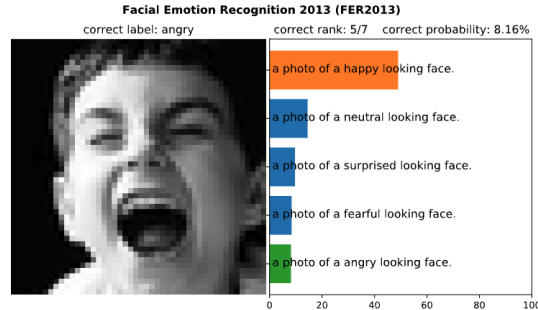
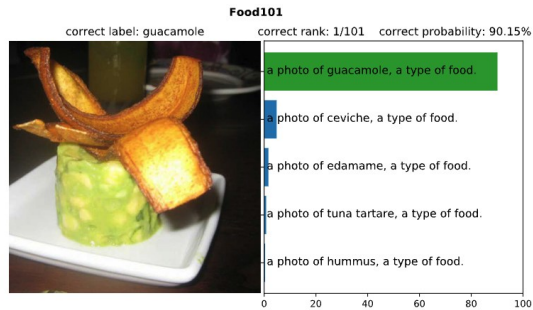
Ziel des Frameworks

- Nutzerzentrierte Evaluierung existierender KI-Modelle zur Identifikation des optimalen Ansatzes für spezifische Use-Cases

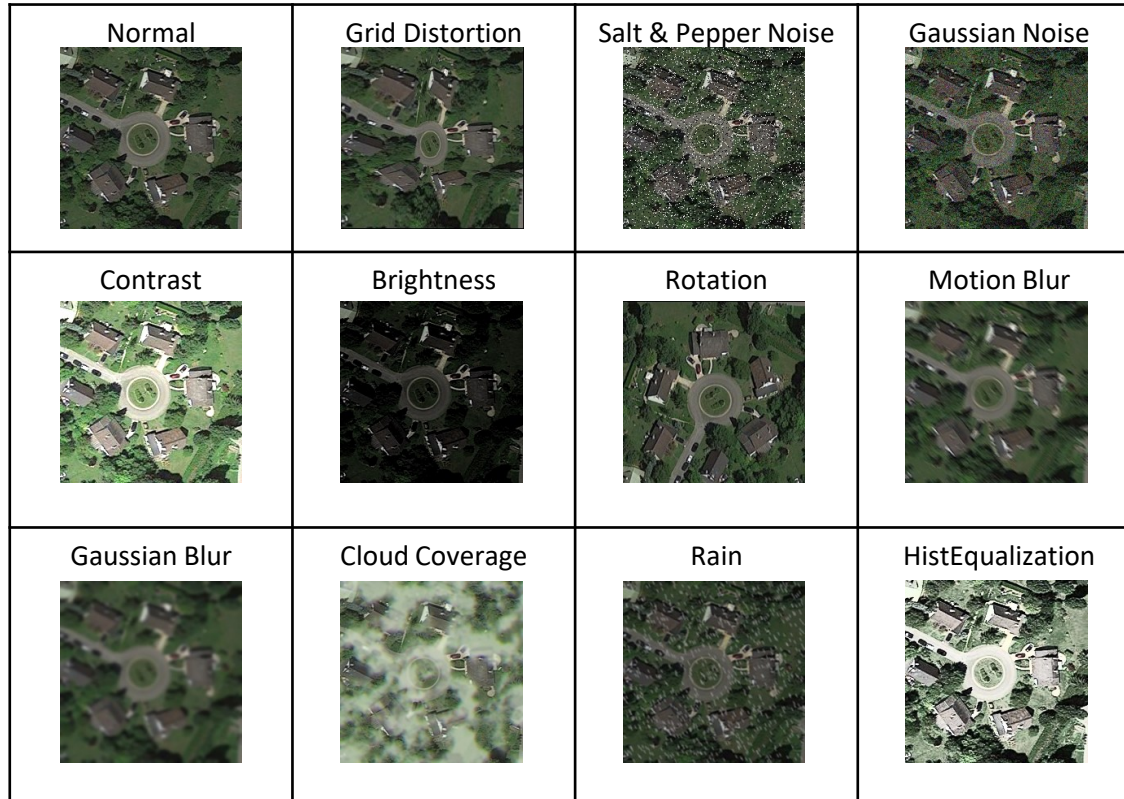
Relevanz für KI-Sicherheit

- Bewertung von KI-Modellen gegen Manipulation und Fehlfunktionen - Verlässlichkeit
- Minimierung von Sicherheitsrisiken durch domänenspezifische Evaluierung

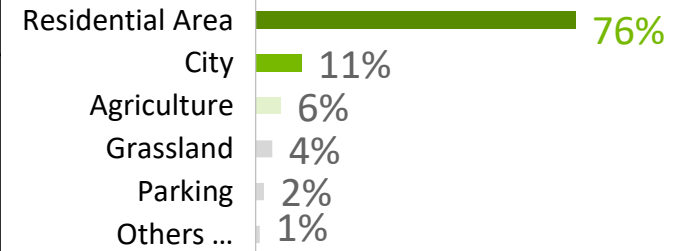
Robustheit von Vision Modellen



Robustheit von Vision Modellen

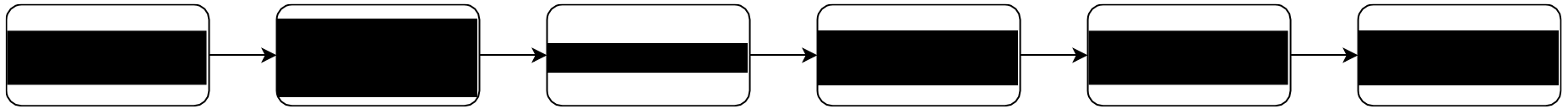


Beispielwerte

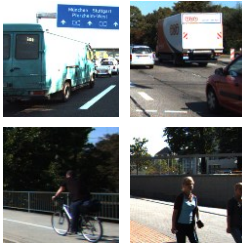


Deepbench: Vertrauen durch Robustheit

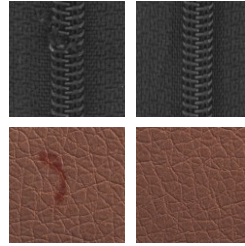
- Flexibles und skalierbares Framework zur Robustheitsanalyse von KI-Modellen
- Fokus auf Vision und Vision-Language Modelle
- Use-Case spezifische Analyse



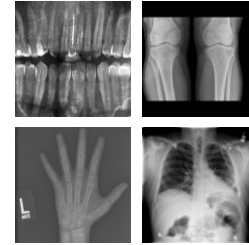
Use Cases



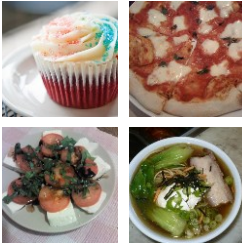
Autonomes Fahren



Manufacturing Quality



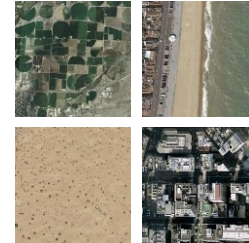
MedicalDiagnosis



Handheld Device



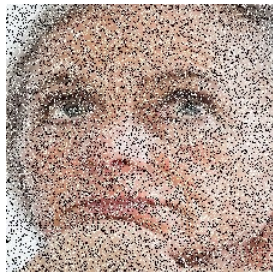
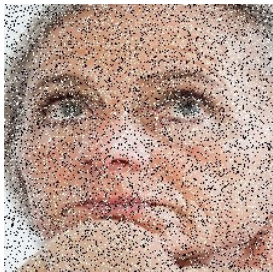
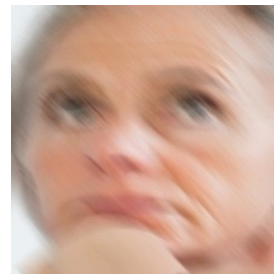
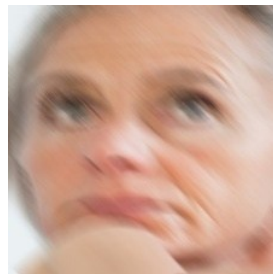
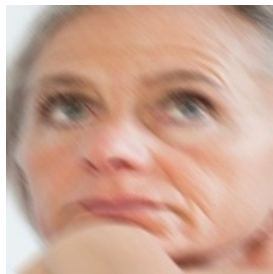
Security (Emotions)



Satellite

Corruptions

Blur



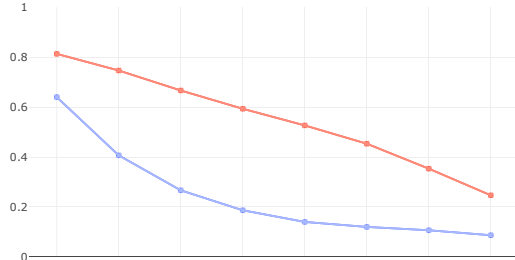
Noise

Domäne-Spezifische Analyse

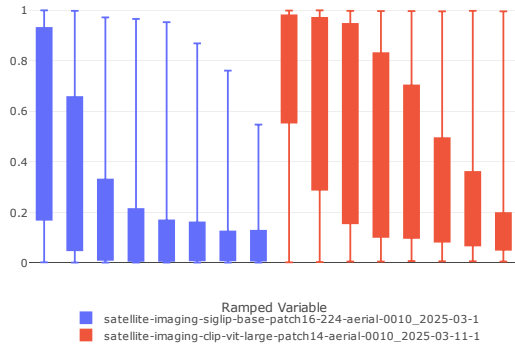
	Medical Diagno- sis	Auto- nomous Vehicles	Manu- factur- ing	People Recog- nitio	Aerial Imaging	Hand- held
Brightness	✓	✓	✓	✓	✓	✓
Cloud Generator					✓	
Contrast	✓	✓	✓	✓	✓	✓
Gaussian Blur	✓		✓			
Gaussian Noise	✓	✓	✓	✓	✓	✓
Colour Shift		✓	✓			✓
Grid Distortion						
Grid Deformation	✓		✓	✓	✓	
Hist Equalization	✓			✓		
Horizontal Flip	✓		✓	✓	✓	✓
Vertical Flip					✓	
Rotation	✓	✓	✓	✓	✓	✓
Motion Blur		✓				✓
Perspective Transformation		✓	✓	✓	✓	✓
Rain		✓				
Salt and Pepper		✓		✓		✓
Shadow		✓	✓	✓	✓	✓

Evaluation: Beispiel Gaussian Noise

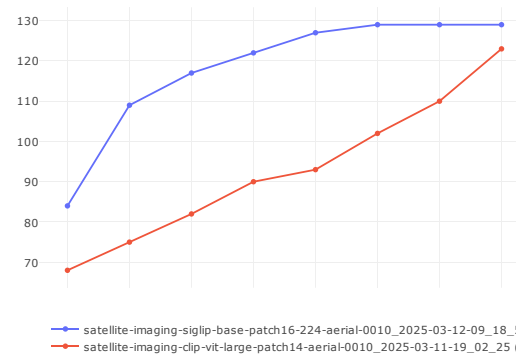
['Accuracy', 'Balanced Accuracy'] for GaussianNoise



Boxplot of Ground Truth Score for GaussianNoise



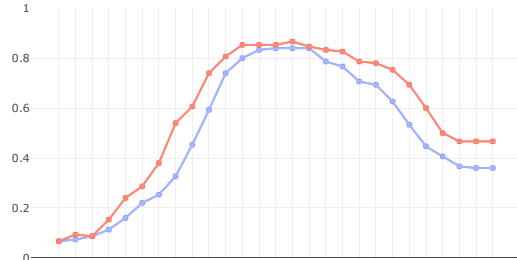
Label Flips for GaussianNoise



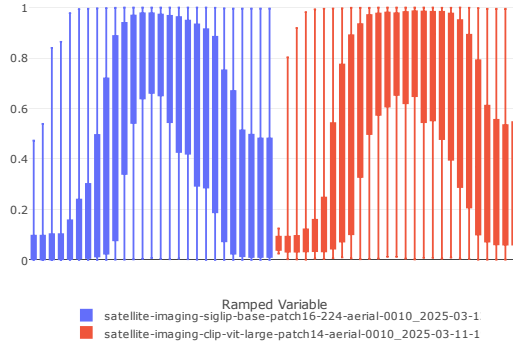
Gaussian Noise

Evaluation: Beispiel Brightness

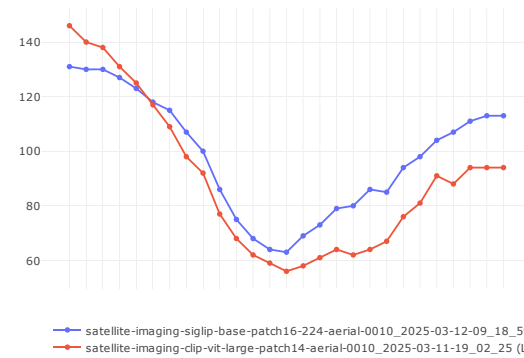
['Accuracy', 'Balanced Accuracy'] for Brightness



Boxplot of Ground Truth Score for Brightness



Label Flips for Brightness



Brightness

Abschluss

www.htw-berlin.de



**Hochschule für Technik
und Wirtschaft Berlin**

University of Applied Sciences