

Robustness report

REPORT METADATA	
PROJECT	GENERATED BY
helmet-100	Knotest
EVALUATION REPORT NUMBER	DATE OF ISSUE
evaluation-f208bab2-699b-4d81-b870-8acb680146a5	2026-09-20

TEST METADATA		
EXECUTION	EVIDENCE	OVERALL ASSESSMENT
Completed	Complete	Criteria not met
TASK	SAMPLES	EVALUATED SCENARIOS
object_detection	100	1
REPORTED METRICS	MODEL BACKEND	RUNTIME ENGINE
1	YOLO	ultralytics 8.4.96
EXECUTION PROVIDER	CONFIDENCE THRESHOLD	NMS IOU THRESHOLD
CPU	0.25	0.5

Module assessment

Criteria not met

1 failed · 0 passed · 0 N/A

Key Findings

Detection mAP 50-95 Fog · Severity 3	RELATIVE DEGRADATION 66.22% degradation	MAXIMUM ALLOWED DEGRADATION (%) <= 20.00%	CRITERIA NOT MET
--	---	---	-------------------------

Baseline performance

Original inputs · 100 evaluated samples
Reference scores used to calculate robustness degradation.

Metric	Baseline score
Detection mAP 50-95	0.02

Scenario Results

Scenario	Samples	Reported metrics	Metric outcomes
Fog · Severity 3	100	1	1 fail / 0 pass / 0 N/A

Method Specification

Atmospheric fog

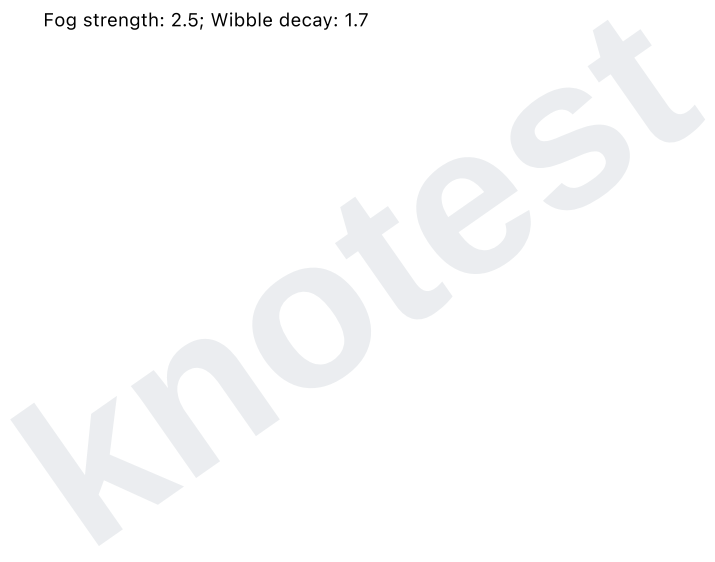
Applies ImageCorruptions fog using the resolved recipe and derived sample seed before model inference.

INPUT	Source image converted to RGB floating-point channels in the range 0 to 1 at its original dimensions.
APPLIED TO	Every selected image sample.
RANDOMNESS	The transformation uses a deterministic sample-specific seed derived from the base seed and sample identity. Seed: 0.
CHANGES	Image pixel values
UNCHANGED	Image dimensions; Ground-truth labels and annotations; Sample identities; Sample order

EXACT SETTINGS

Base seed	0
Base value used to derive a deterministic seed for each sample.	

Severity 3	Fog strength: 2.5; Wibble decay: 1.7	Samples: 100
------------	--------------------------------------	--------------



Maximum allowed degradation (%): <= 20.00%
Score above · Relative change below: + degradation, – improvement, 0 no change.

Metric	Baseline	Severity 3
		Samples: 100
Detection mAP 50-95	0.02	0.01
		+66.22%

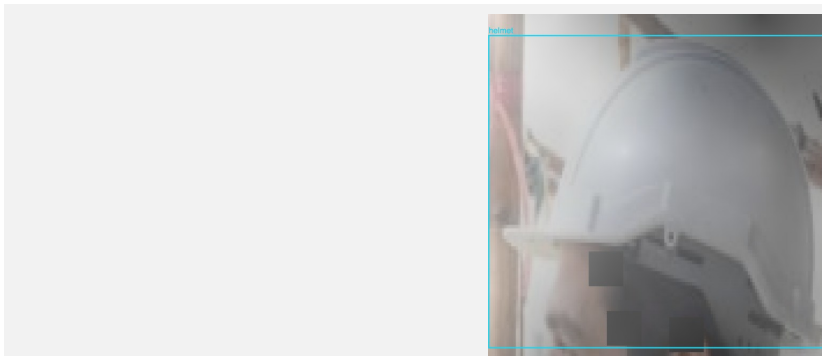
knotest

Fog · Severity 3

0_35497086_jpg.rf.689bb90e7d...865569061b8.jpg

MEAN IOU

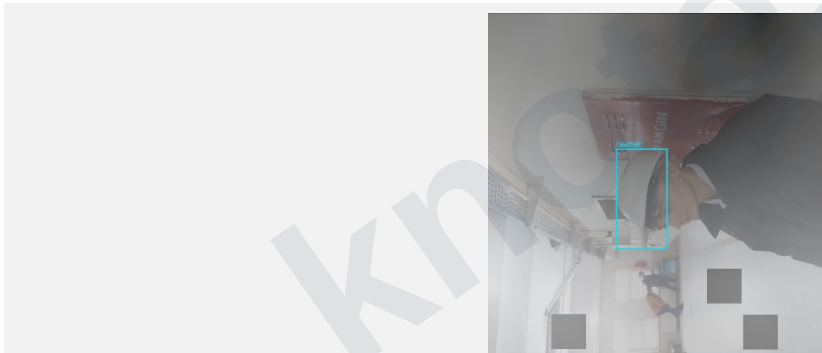
0.000



20221206_095539_jpg.rf.f2fb6...fb0e4fe0af3.jpg

MEAN IOU

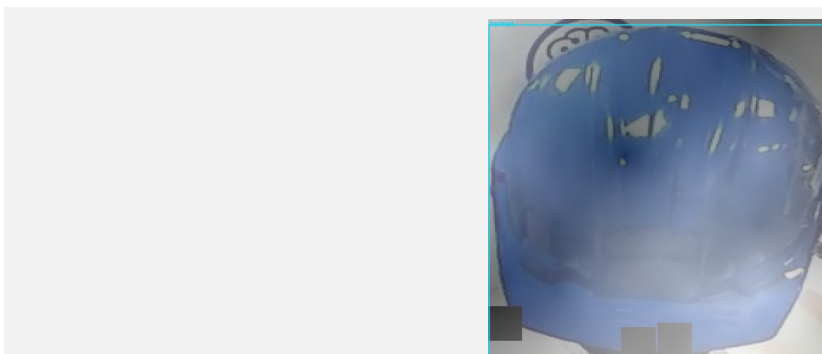
0.000



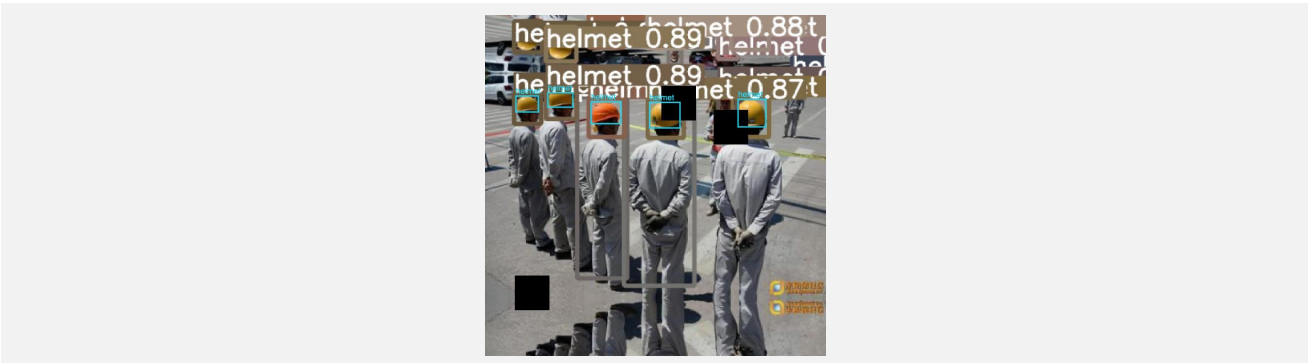
0_26614471_jpg.rf.5e0622816e...ee9edadcdc4.jpg

MEAN IOU

0.000



Robustness Baseline

hard_hat_workers3932_png_jpg...a54520b9165.jpg	MEAN IOU 0.000
	
hard_hat_workers3489_png_jpg...f584b9d3f75.jpg	MEAN IOU 0.000
	
hard_hat_workers2363_png_jpg...7c45e0752e0.jpg	MEAN IOU 0.000
	

TERM	DEFINITION
Baseline	Performance measured on the original evaluation images without any attack, corruption, or augmentation applied.
Scenario	A defined stress condition, such as a corruption, augmentation, or attack setting, evaluated against the same model and dataset.
Corruption transform	A non-adversarial image degradation, such as fog, blur, or noise, applied to test sensitivity to environmental quality loss.
Augmentation transform	A geometric or photometric image modification, such as scale, perspective, or shear, applied to test stability under view changes.
Adversarial attack	A perturbation optimized against the model, such as FGSM or PGD, to intentionally reduce model performance under a controlled constraint.
mIoU	Mean intersection over union between the predicted mask and the reference annotation mask across the evaluated samples.
Relative degradation	The percentage by which a transformed metric worsens relative to its Baseline measurement, adjusted for whether higher or lower metric values are better. Positive values indicate degradation; negative values indicate improvement.
Effective samples	The number of samples included in the reported scenario result after eligibility checks and scenario-specific filtering.

The assessment results are based on the defined model/backend evaluation scope, dataset, test configuration, and risk-based threshold approach agreed for this review. The report provides technical evidence on explainability, robustness, and related risk considerations. It should not be interpreted as full final deployment acceptance of the complete robot system unless additional deployment-level validation, including applicable hardware, sensor, and operational-environment testing, has been completed.

knotest