COPA

Paper Presentation

Robust Vision-Based Runway Detection through Conformal Prediction and Conformal mAP

Alya Zouzou, Léo Andéol, Mélanie Ducoffe and Ryma Boumazouza

10th September 2025

ANITI DEEL SNCF AIRBUS

Motivation: Visual Landing System (VLS)

An ambition : to create an AI-based autonomous system that ensures a safe landing.

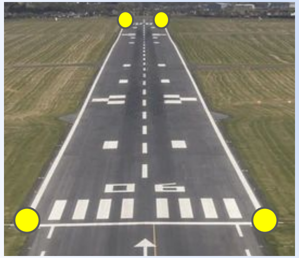
Pilot support

Input Image

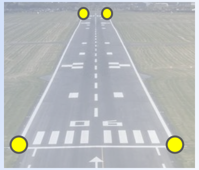
1. Runway Detection

Crop

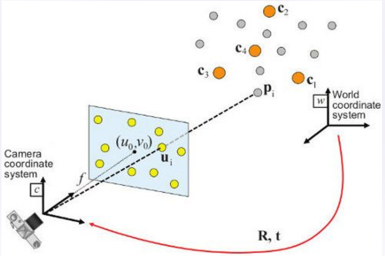
2. 2D Keypoints extraction



Keypoints Features Extraction



3. Pose estimation



⇒ This research industrial use case ***is being studied by a diverse range of institutes and companies***, including Airbus (ATTOL, A³), Boeing, Daedalean, and various Chinese research organizations.

Our Contributions

Airbus AI Research

1. **Application of Conformal Prediction (CP)** ⇒ allows **uncertainty quantification** for Runway Detection using **PUNCC library**.

   DEEL PUNCC A Conformal Prediction Library

2. **Conformal Mean Average Precision (C-mAP):**
   ⇒ introduce **C-mAP**, a novel metric combining accuracy & robustness evaluation.

3. **Open-Source Contribution:**
   ⇒ **codebase** & **trained models** on GitHub, to promote **reproducibility** and encourage further research in this area. (e.g *conformal_runway_detection* on git)

 Our Repository : https://github.com/alyasltd/conformal_runway_detection

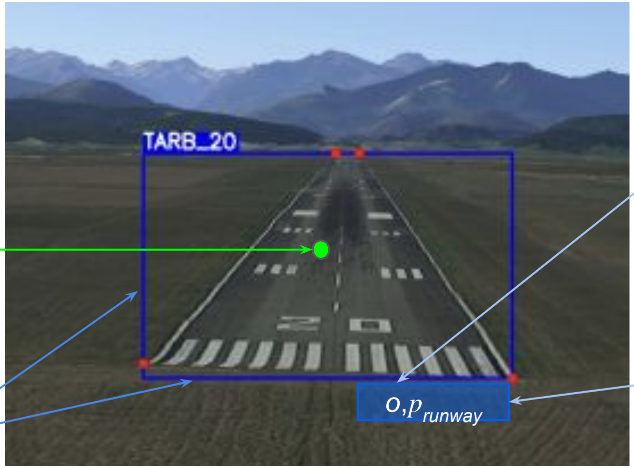
Quick Object Detection Overview

What's YOLO ?

- **Identify and localize objects** within an image
- **Classify** them into predefined categories
- **Single-stage detector :** directly predict bounding boxes and class probabilities for each object in a single pass through the network

⇒ ***low run-time & accuracy***

Relative center (x,y)



Width and Height (w,h)

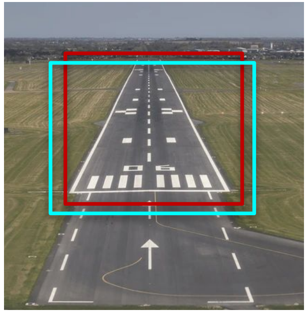
O = Objectness Score ***"Is there something here?" score*** - Used to quickly filter out background areas.

p_{runway} = Class Probability ***"What is it?" score*** that represents the probability of the object belonging to a particular class.

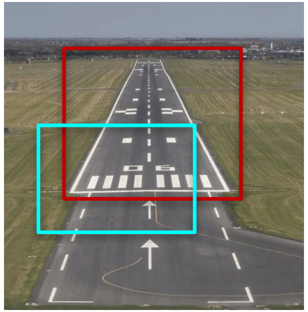
Confidence-score = $o \times p_{runway}$ ⇒ **"proxy for uncertainty estimation"**

Redmon et Al, CVPR 2016 - Li et al. 2023

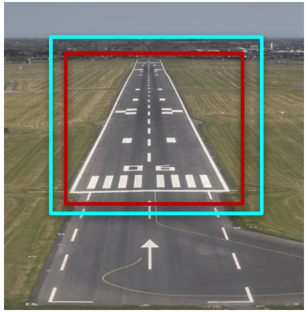
How to evaluate ONE box prediction: What's IoU & IoA ?



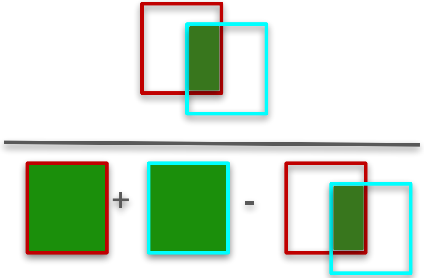
Good IoU !



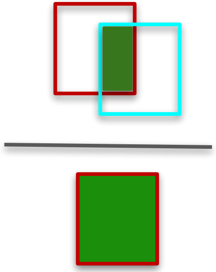
Bad IoU and IoA !



Good IoA !



⇒ **Greater the overlap**, the closer **IoU** is to **1.**
⇒ Value between **0** and **1**



⇒ **Greater the ground truth is covered** by the prediction, the closer **IoA** is to **1.**
⇒ Value between **0** and **1**

How to evaluate Object Detection accuracy ?

Mean Average Precision (mAP) - YOLO's evaluation **of average precision across all classes**
⇒ **between 0 (very bad overlap) and 1 (perfect overlap)**

$$\text{mAP} = \frac{1}{C} \sum_c \underbrace{AP}\left(\underbrace{\{\mathbf{b}_i^x\}_{\hat{c}_i^x = c}}, \underbrace{\{b_j^{gt,x}\}_{c_j^x = c}} \right),$$

Number of classes - here 1

Metric to evaluate the matching - **IoU per class**

$$\hat{c}_i^x = \arg\max_k \mathbf{p}_i^x(k).$$

Predicted Boxes with highest probabilities

GT

⇒ **mAP = AP cause we've got 1 class only**

How's AP computed ?

Airbus AI Research

1

Let's take a IoU Predefined threshold of **0.5 !**

2

b_{pred}

$\text{IoU}(b_{\text{pred}}, b_{\text{gt}}) \geq \tau$	1	1	0

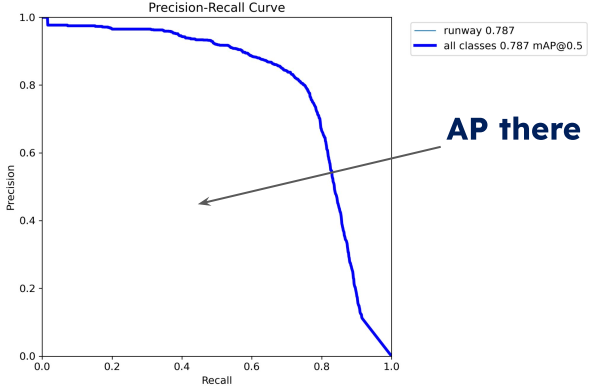
Ranking Predicted Boxes by confidence score & match GT & Predictions based on maximisation of IoU.

4

Construct interpolated precision-recall curve

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

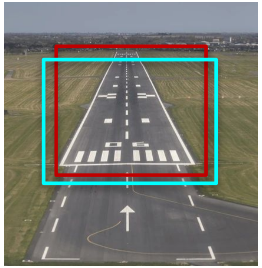


AP there !

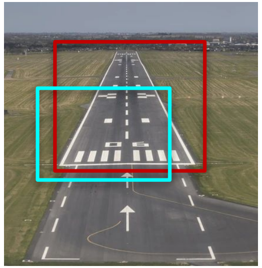
Computation of TP, FP, FN for Recall and Precision

3

True Positive
IoU = 0.96



False Positive
IoU = 0.22



False Negative
IoU = 0.00

Goal: Safety Evaluation

GOAL = Predict bounding box that **covers** the true object under a certain level of risk α

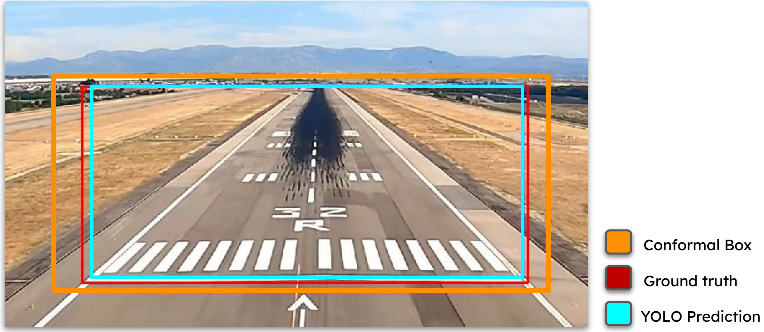
The problem …

1. Runway Detection

Crop

2. 2D Key points extractions

If the bounding box **does not fully contain** the ground truth and thus misses **at least 3 key points** ⇒ ***the VLS fails.***



But HOW ? by applying Conformal Prediction to the output of object detection models.

⇒ We need to evaluate how **conformalization** affects performance using dedicated, industrially relevant metrics.

How does CP work for Object Detection ?

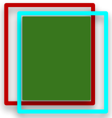
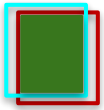
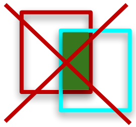
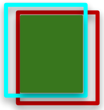
DEEL
PUNCC
A Conformal Prediction Library

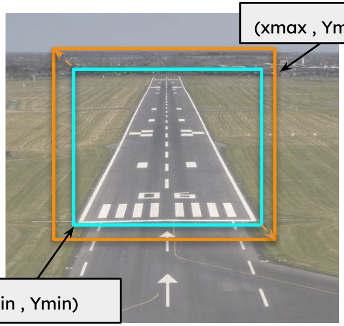
HYPOTHESIS :

- ⇒ Test data i.i.d. with respect to calibration
- **Independence:** each image must not depend on the others
- **Identical distribution:** same distribution for test and calibration

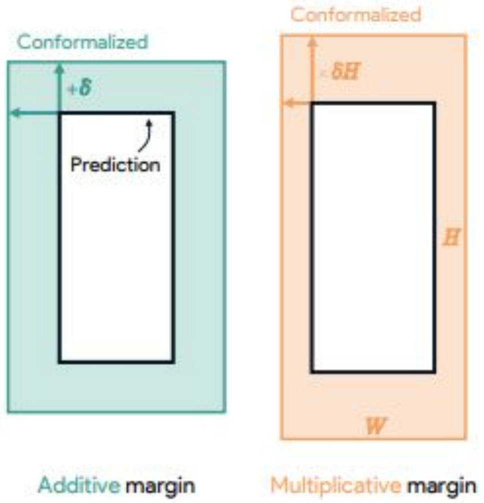
Key Steps in Conformal Object Detection with PUNCC :

1. **Train a Base Predictor** → YOLO
2. **Consider a Calibration Set** → independent from Training Set
3. **Keep only true positive** based on IoU
4. **Compute 4 Nonconformity Scores**
5. **Expand** the **predicted bounded box**.

Calibration Item				
True Positive ?				



But how to expand a box ?



Great for similar size object

Great for differents size object

<u>For example, for additive nonconformity measure:</u>

1. Computation of non-conformity scores = differences between predictions & GT - for each of the four coordinates.

$$\mathbf{r}_k^a = (\hat{x}_{\min}^k - x_{\min}^k,\ \hat{y}_{\min}^k - y_{\min}^k,\ x_{\max}^k - \hat{x}_{\max}^k,\ y_{\max}^k - \hat{y}_{\max}^k)$$

4 independent distributions … but is the risk α actually respected ?

2. SOLUTION = **stricter quantiles (Bonferroni Correction)**

$$\hat{q}(j) = q_{\lceil (1-\frac{\alpha}{4})(n+1) \rceil / n}\left(\{r_k(j) \ : \ k \in \{1, \dots, n\}\}\right)$$

$$\mathbb{P}\left(\mathbf{b}_{n+1}^{gt} \subseteq \hat{\mathbf{b}}_{n+1}^{\text{conf}}\right) \geq 1 - \alpha$$

$$x_{\min} \geq \hat{x}_{\min}^{\text{conf}}, \quad y_{\min} \geq \hat{y}_{\min}^{\text{conf}}, \quad x_{\max} \leq \hat{x}_{\max}^{\text{conf}}, \quad y_{\max} \leq \hat{y}_{\max}^{\text{conf}}$$

Left Right

3. Coordinate adjustment

$$\hat{\mathbf{b}}_k^{\text{conf}} = (\hat{x}_{\min}^k - \hat{q}_1,\ \hat{y}_{\min}^k - \hat{q}_2,\ \hat{x}_{\max}^k + \hat{q}_3,\ \hat{y}_{\max}^k + \hat{q}_4)$$

Are IoU & IoA robust metrics for VLS ?

AIR
Airbus AI Research

How to guarantee a robust detection ?

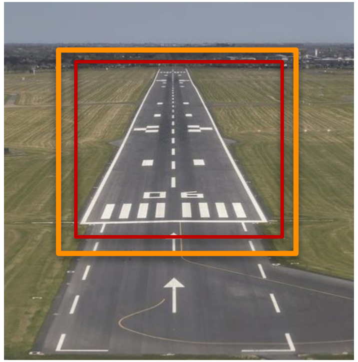
with IoA = 1 !

IoA = 1
&
IoU = **0.4**

Interpolation Error (high resolution images) ⇒ BAD Pose Estimation !

Only ?

also with a decent IoU …



IoA = 1
&
IoU = **0.9**

NO Interpolation Error ⇒ GOOD Pose Estimation !

Are IoU & IoA robust metrics for VLS ?

AIR
Airbus **AI Research**

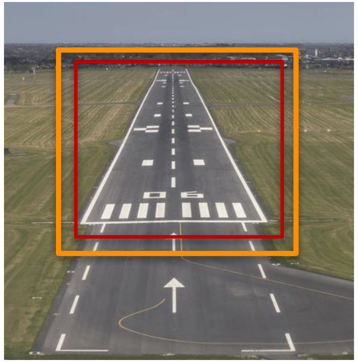
How to guarantee a robust detection ?

with IoA = 1 !

IoA = 1
&
IoU = **0.4**

Only ?

also with a decent IoU …



IoA = 1
&
IoU = **0.9**

Our **conformal predictor (CP)** only ensures one part of what we call a “Robust Detection” when the **base predictor** ensure the other.

	IoU>t	IoA=1	Robust ?
YOLO	✓	✗	✗
C-YOLO	?	✓	?

⇒ we want to evaluate the CP in term of how much the **IoA = 1** guarantee degrades the **IoU > *t*** guarantee.

Evaluation : Coverage & C-mAP

AIR
Airbus AI Research

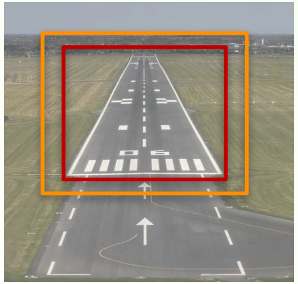
COVERAGE (eg. Andéol et al.)

$$\sum_i \mathbb{1}_{Y_i \in \mathcal{C}_{\hat{\lambda}}(X_i)}$$

1 if the **GT** Y_i for an input X_i is **contained within** the conformal bounding box $C_{\hat{\lambda}}(X_i)$

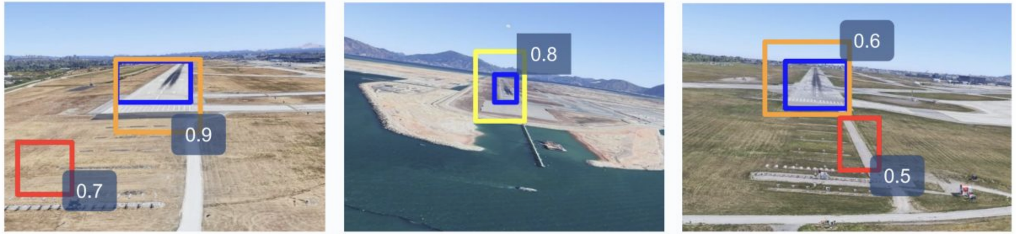
0 if the **GT** Y_i for an input X_i is **not contained** within the conformal bounding box $C_{\hat{\lambda}}(X_i)$

Coverage = 1



Coverage = 0

C-mAP



Confidence	0.9	0.8	0.7	0.6	0.5
b_{pred}					
$\text{IoU}(b_{pred}, b_{gt}) \geq \tau$	1	0	0	1	0
$\text{IoA}(b_{pred}, b_{gt}) = 1$	1	1	0	1	0
TP	1	1	1	2	2
FP	0	1	2	2	3
FN	2	2	2	1	1
Precision $= \frac{TP}{TP+FP}$	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{5}$
Recall $= \frac{TP}{TP+FN}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$

AIR
Airbus **AI Research**

Experiments…

Train and Calibration Data (LARD : Landing Approach Runway Detection)

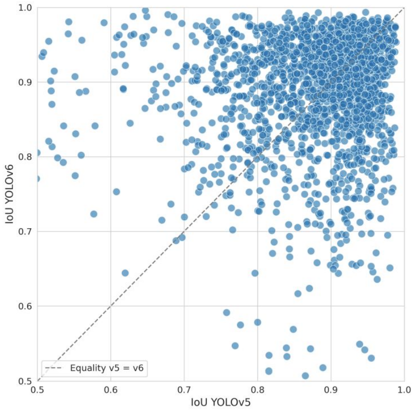
- **Public** dataset for runway detection – **single class**.
- Approach views at different distance :

- **Synthetic images** (generated with Google Earth Studio) and **real images** :

Metric	YOLOv5 (pre-trained)	YOLOv6 (pre-trained)
mAP@0.5	0.995	0.9901
mAP@0.5:0.95	0.9712	0.9413
GFLOPs	15.8	45.3

- 2 types of image data
- Data Split of the test set into 80/20 to ensure independence
  ⇒ **80%** used for **calibration** and **20%** for **evaluation**
- 2 models : **YOLOv5 & YOLOv6**

Set	Type	Images
Train	Synthetic	11,546
Validation	Synthetic	2,886
Test	Real+Synth	2,315
Test Synth	Synthetic	2,212
Test Real	Real	103



 Redmon et Al, CVPR 2016 - Li et al. 2023

Results

Unanswered questions :

Is Conformal Object Detection with IoA sufficient for VLS ?
Does Conformalised IoA imply Conformalised IoU ?

Model	mAP	C-mAP	C-mAP@50@80:100
YOLOv5	96.88	0.77	46.92
c-YOLOv5-a	92.67	**56.86**	80.73
c-YOLOv5-m	96.17	55.84	**82.18**
YOLOv6	**98.13**	1.31	51.94
c-YOLOv6-a	95.09	55.75	81.86
c-YOLOv6-m	96.71	52.71	81.93

⇒ Conformalisation: **+51 to 55 pts of C-mAP (52.7% → 56.9%)**
⇒ mAP remains high: **> 92%**

⇒ Multiplicative penalisation = **best trade-off between mAP / C-mAP**

Model	Coverage
c-YOLOv5-a	**77.06**
c-YOLOv5-m	75.88
c-YOLOv6-a	75.73
c-YOLOv6-m	73.93

⇒ Guaranteeing the risk level of 30% - all coverage above 70%
⇒ Particularly effective with **c-YOLOv5-m**

Conclusion

Airbus AI Research

- **The Good:** CP = flexible uncertainty quantification method that doesn't require retraining your model.
- **The Bad:**
  - It can be overly optimistic (Bonferroni effect).
  - High coverage alone isn't sufficient for VLS, which require precision ⇒ **CmAP more adapted metric.**

CP offers a **promising path to align with upcoming AI safety standards** (AI Act, EASA)

Limitations & Future Work

- current evaluation limited to synthetic data - only 1 runway per image.

Alya ZOUZOU
Email me : alyasltd@gmail.com

Airbus AI Research

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

DEEL
DEpendable & Explainable Learning
France-Québec

SNCF

Read our paper !

Thank you!

Léo Andéol
leoandeol@gmail.com

AIRBUS