Technische Hochschule mannheim

CeMOS Intelligente Systeme

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Efficient Post-hoc Calibration in Object Detection without Held-Out Data

Nikolas Ebert, Didier Stricker, Oliver Wasenmüller

17 August 2026 | International Conference on Pattern Recognition 2026

Overconfident Detectors Are a Safety Risk!

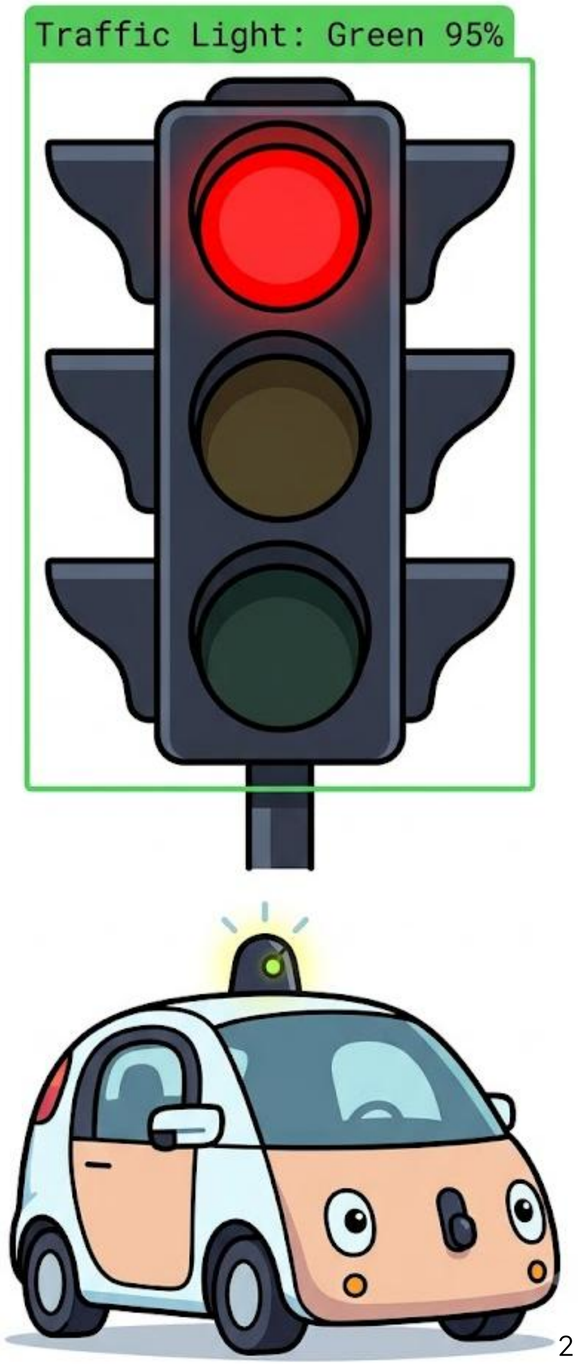
- **Object detectors** are used in **safety-critical systems**

  ➔ **Accuracy alone is not enough** - models must know when they might be wrong.

  ➔ Deep neural networks are frequently **over- or underconfident**

- A well-calibrated model's **confidence** score should match its real-world **accuracy** ➔ Model Confidence ≈ Accuracy

- **Goal of Calibration:** align predicted confidence with actual precision.



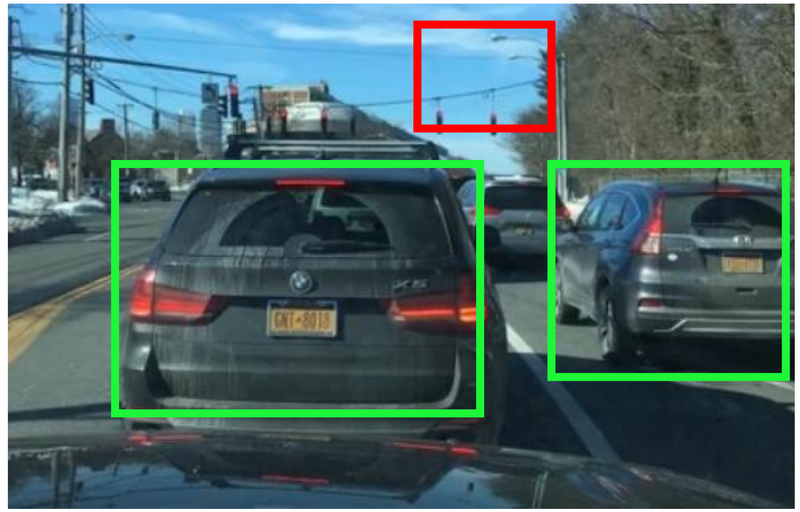
Background
Measuring Calibration

Technische Hochschule mannheim

CeMOS Intelligente Systeme

D-ECE: average gap across all detections $\widehat{D}$ between predicted confidence $\bar{z}$ and actual precision $\rho(j)$, across J confidence bins [1].

LaECE: average gap across all detections $\widehat{D}$ between predicted confidence $\bar{z}$ and a joint score combining classification accuracy and localisation quality (IoU), across J confidence bins [2].



$$\text{D-ECE} = \sum_{j=1}^{J} \frac{|\hat{D}_j|}{|\hat{D}|} \left| \bar{z}_j - \rho(j) \right|$$

True Positive (TP):

- predicted class is correct
- IoU with ground truth >= threshold τ.

$$\text{LaECE} = \frac{1}{C} \sum_{c=1}^{C} \sum_{j=1}^{J} \frac{|\hat{D}_j^c|}{|\hat{D}^c|} \left| \bar{z}_j^c - \rho(c,j) \cdot \overline{\text{IoU}}(c,j) \right|$$

(1) Kuzucu et al. *On calibration of object detectors: Pitfalls, evaluation and baselines*. ECCV, 2024
(2) Kuppers et al. *Multivariate confidence calibration for object detection.* CVPRW, 2020.

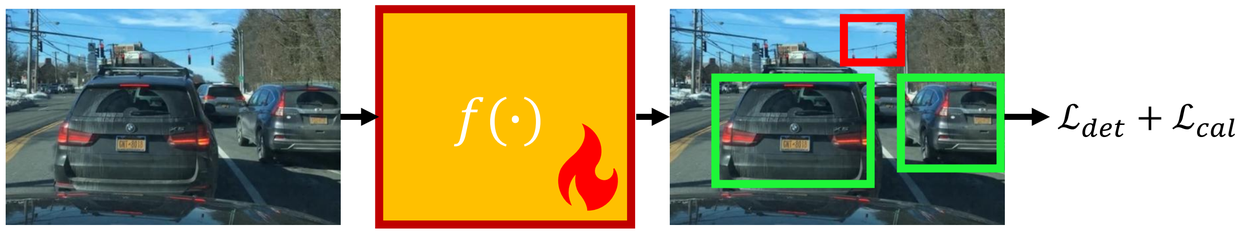
Related Work
Train-time vs. Post-hoc Calibration

Technische Hochschule mannheim

CeMOS Intelligente Systeme

Train-time Calibration [1,2]

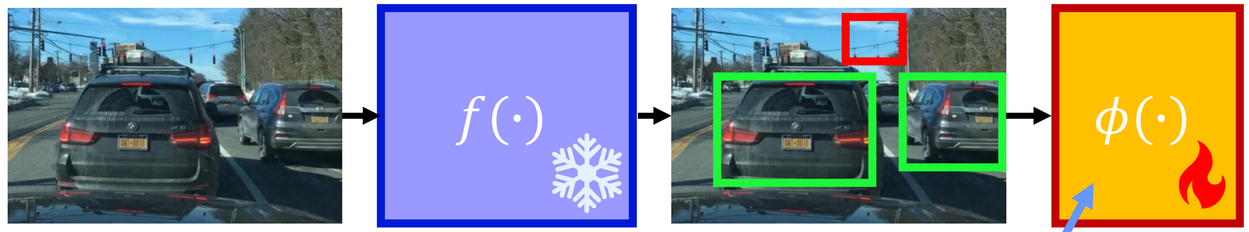
Training Set



- \+ No post-processing needed
- \+ No holdout set required
- – Inflexible after deployment
- – Conflicts with accuracy objective
- – Fails under distribution shift

Post-hoc Calibration [3,4]

Calibration Set



- \+ Flexible after deployment
- \+ Model-agnostic approach
- \+ Simple to apply
- – Requires holdout set
- – Separate calibration step

Typically [3,4]:
- Platt Scaling (PS)
- Temperature Scaling (TS)
- Isotonic Regression (IR)

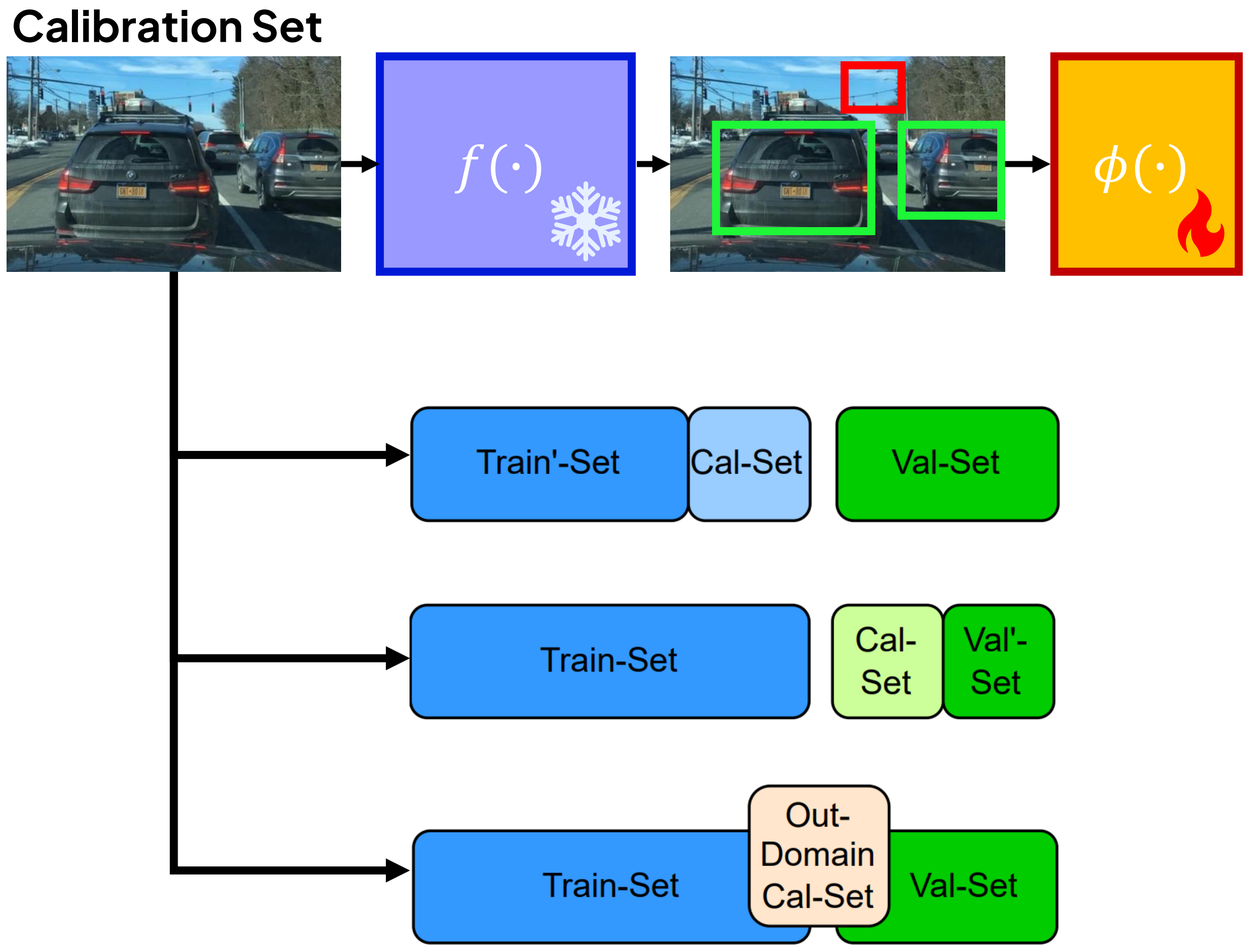
(1) Munir et al. *Bridging precision and confidence: A train-time loss for calibrating object detection*. CVPR, 2023.
(2) Munir et al. *Cal-detr: calibrated detection transformer*. NeurIPS, 2023.
(3) Kuzucu et al. *On calibration of object detectors: Pitfalls, evaluation and baselines*. ECCV, 2024
(4) Kuppers et al. *Multivariate confidence calibration for object detection*. CVPRW, 2020.

Related Work
The Held-Out Data Bottleneck

- Post-hoc calibration methods need a **held-out calibration set**.

- The calibration set comes from:
  a. Training date
  b. Validation data [2,3]
  c. Out-of-Domain Dataset [1]

➔ *Costly in data-scarce domains like medical imaging*



(1) Munir et al. *Cal-detr: calibrated detection transformer.* NeurIPS, 2023.
(2) Kuzucu et al. *On calibration of object detectors: Pitfalls, evaluation and baselines.* ECCV, 2024
(3) Kuppers et al. *Multivariate confidence calibration for object detection.* CVPRW, 2020.

Key Idea
Reuse Training Data for Calibration

Technische Hochschule mannheim

CeMOS Intelligente Systeme

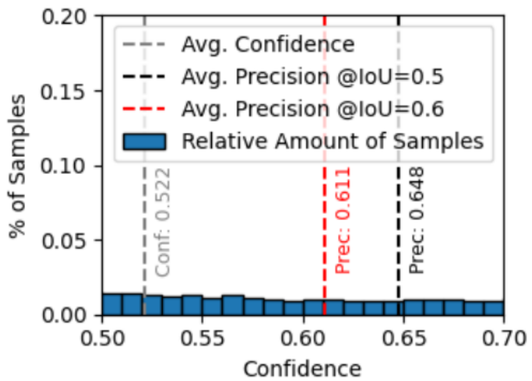
- Our method calibrates **directly on training** data

  ➔ **without excluding** it from detector training.

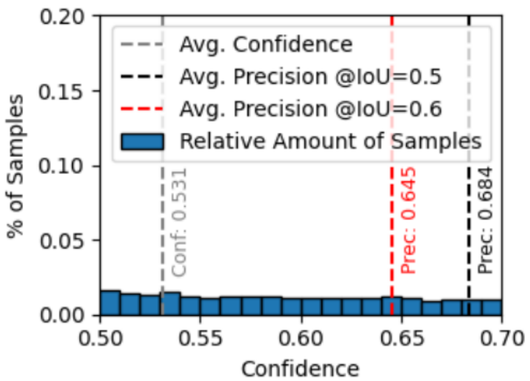
- No data is sacrificed; ideal for data-scarce domains.

Problem: Detector has already seen train images

➔ Precision is higher than on unknown data.



(a) COCO Validation



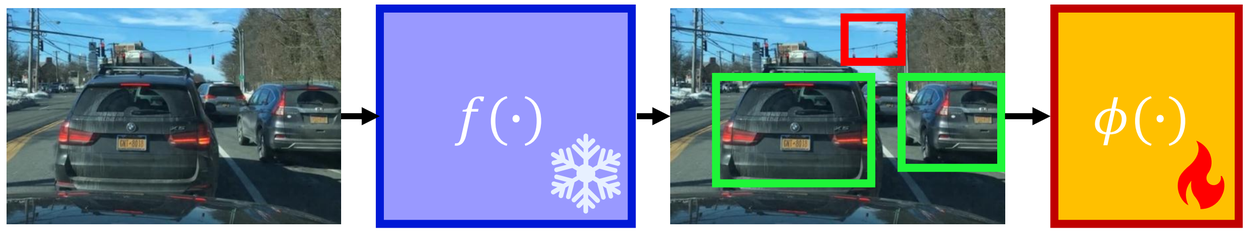
(b) COCO MiniTrain

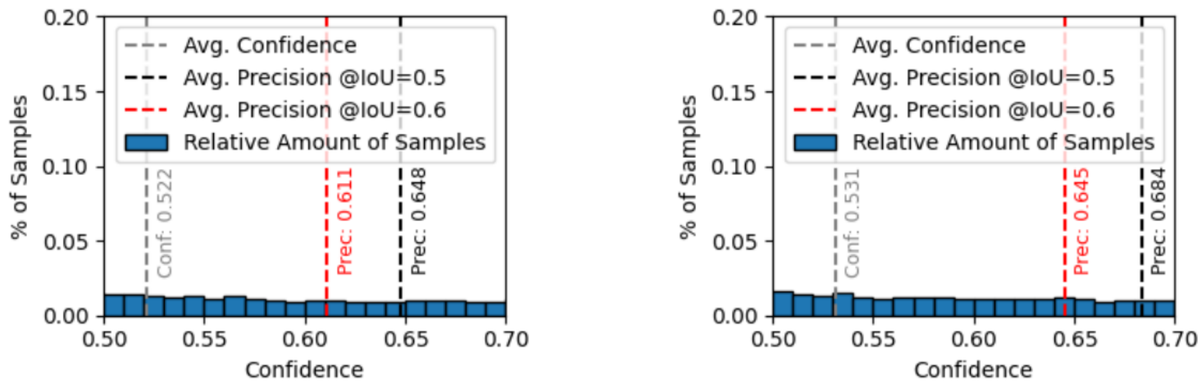
Excerpts from confidence histograms (100 bins) of Deformable DETR on the COCO MiniTrain (5k images) and validation (5k images) sets.

Our Fix
A Stricter IoU Threshold as Regularizer

- **During calibrator training only:** raise the IoU threshold τ that defines a True Positive.
- Requiring higher-quality box matches suppresses **overconfident, borderline detections**.

- Effect on COCO MiniTrain: precision-confidence gap shrinks from 15.3 ($\tau = 0.50$) to 11.4 ($\tau = 0.60$).
- Evaluation itself always uses the standard $\tau = 0.50$, only calibrator training changes.



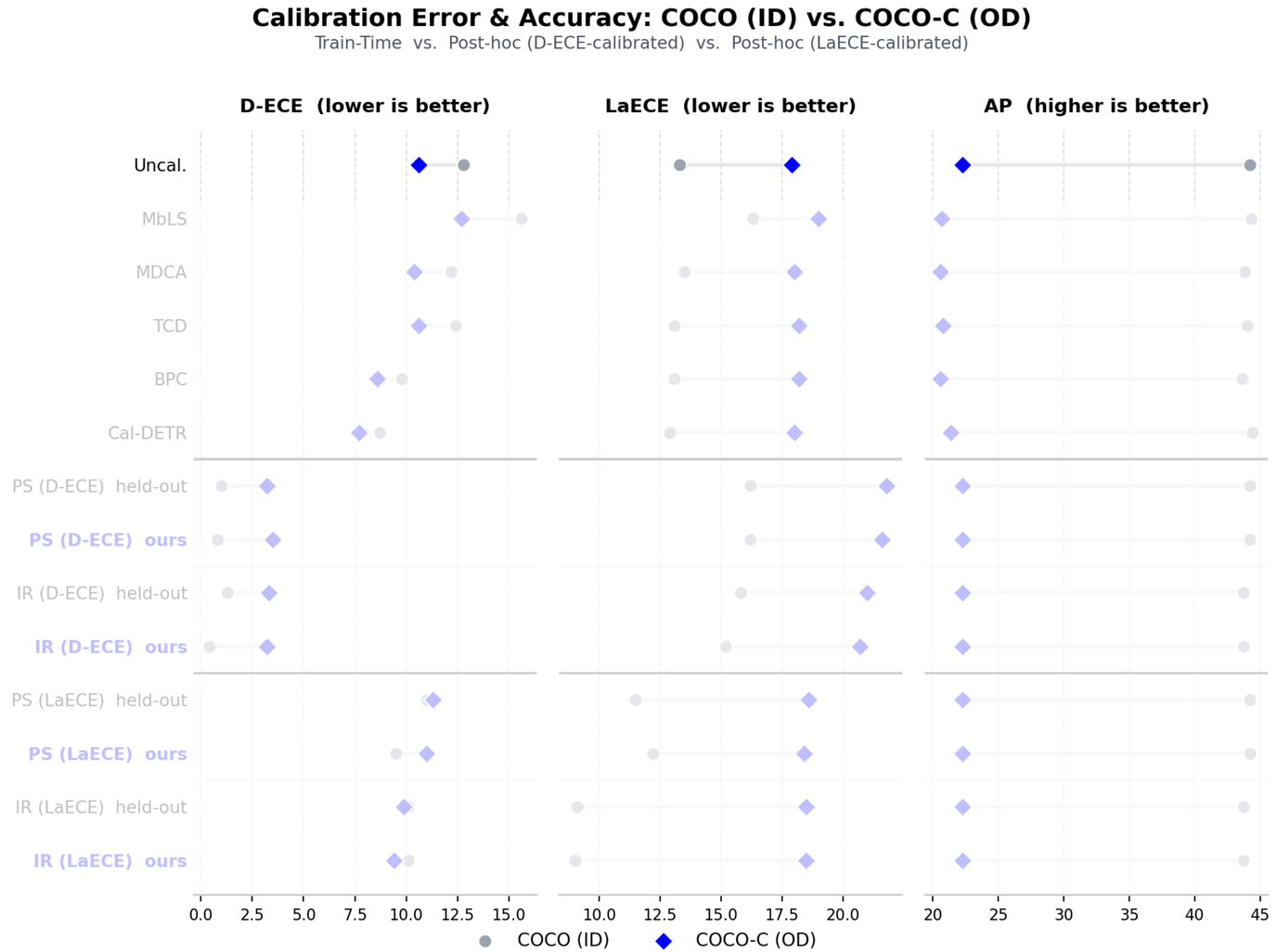


(a) COCO Validation (b) COCO MiniTrain

Excerpts from confidence histograms (100 bins) of Deformable DETR on the COCO MiniTrain (5k images) and validation (5k images) sets.

Evaluation
State-of-the-Art on COCO

- **Model:** Deformable DETR
- **Train Data:**
  - **Train:** COCO train
  - **Train (held-out cal.):** COCO cal (2.5k imgs)
  - **Train (our cal.):** COCO mini-train (5k imgs)
- **Val Data**
  - **Val (ID):** COCO mini-val (2.5k imgs)
  - **Val (OD):** COCO-C mini-val (2.5k imgs)
- **Post-hoc Calibrator:**
  - Platt Scaling (PS)
  - Isotonic Regression (IR)



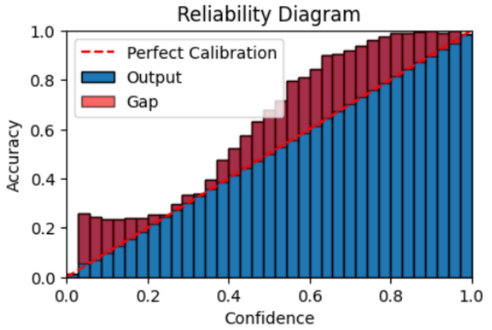
Evaluation
Reliability Diagrams

Technische Hochschule mannheim

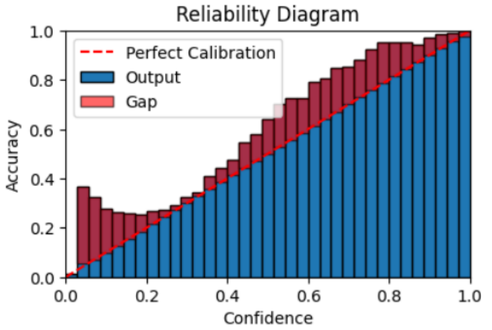
CeMOS Intelligente Systeme

- **Deformable DETR on COCO**

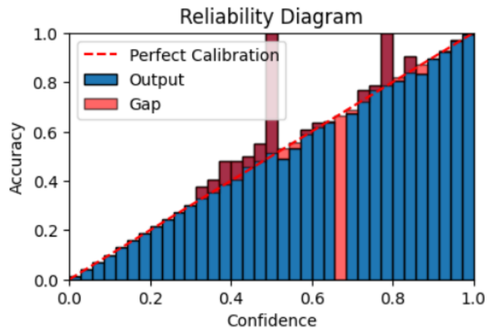
- Bars close to the diagonal = confidence matches precision.
- Uncalibrated model (a): confidence consistently exceeds actual precision.
- Our calibrated model (d): bars track the diagonal closely across confidence levels.



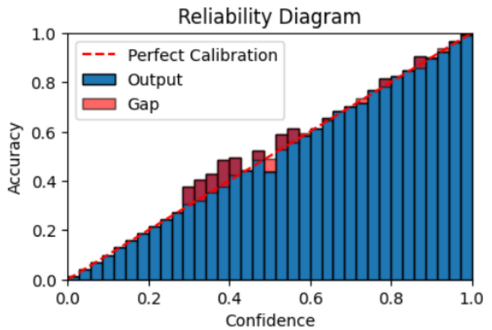
(a) Uncal. [39]



(b) Cal-DETR [23]



(c) IR [14]



(d) IR (Ours)

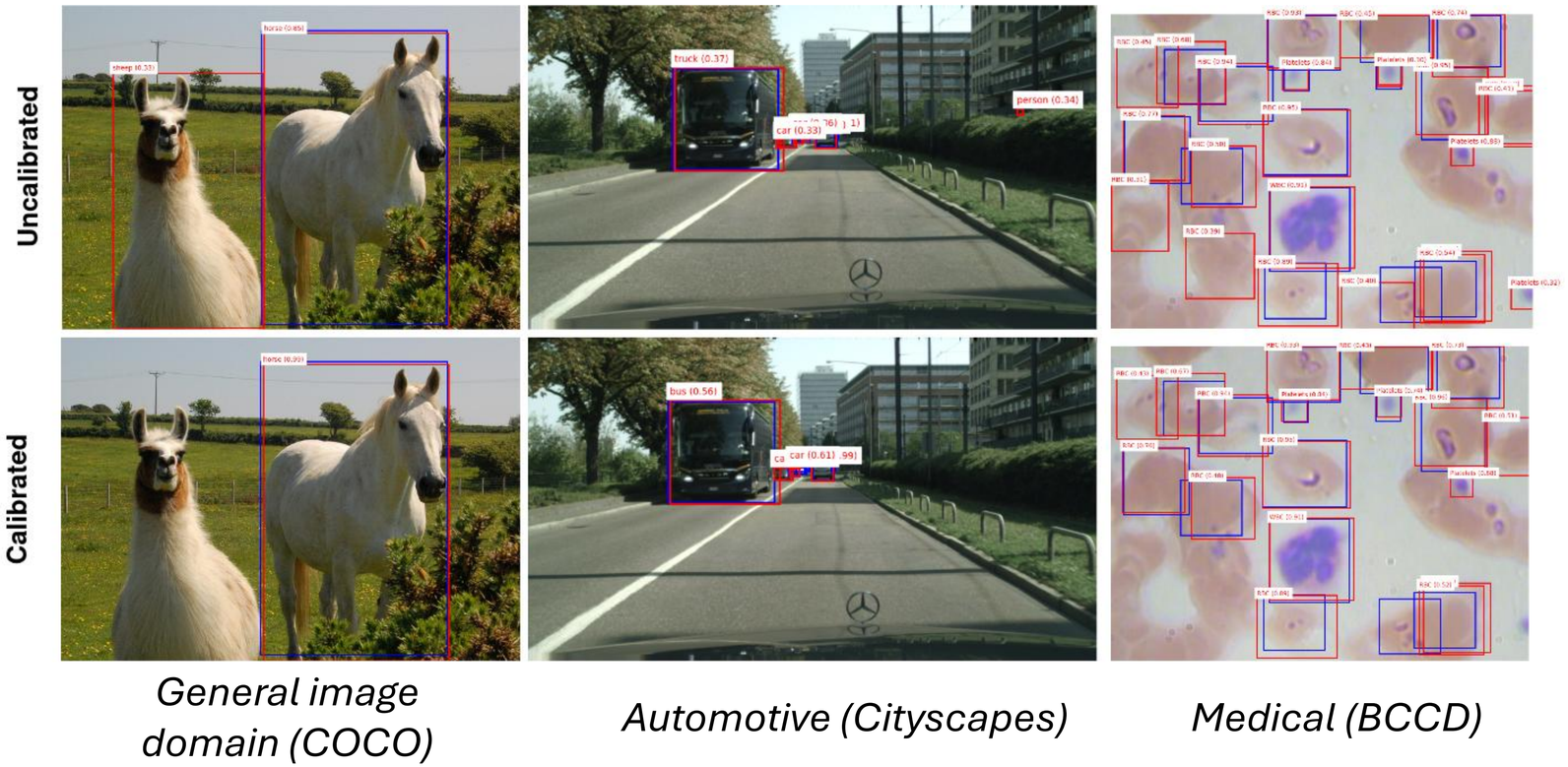
Evaluation
Generalizes Across Domains and Detectors

Technische Hochschule mannheim

CeMOS Intelligente Systeme

- **Confirmed across three Domains**:
  - General image domain (COCO)
  - Automotive (Cityscapes)
  - Medical (BCCD, AGAR)

- **Confirmed across three detectors**:
  - D-DETR
  - DINO
  - Faster R-CNN



General image domain (COCO) | *Automotive (Cityscapes)* | *Medical (BCCD)*

➔ For a detail evaluation, please refer to our paper

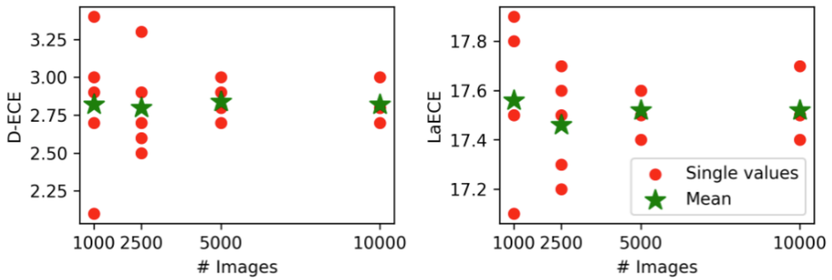
Ablation Study
Impact of Calibration Set Size

Technische Hochschule mannheim

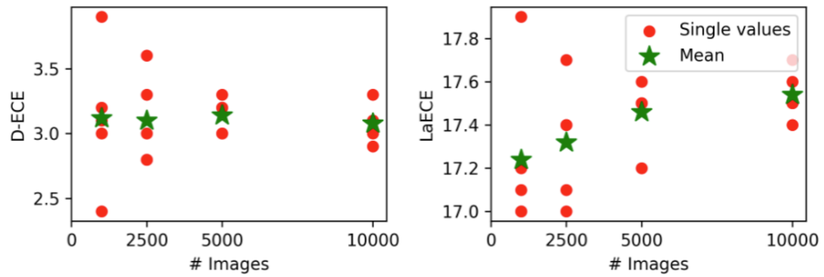
CeMOS Intelligente Systeme

- **Model:** Deformable DETR
- **Train Data:**
  - **Train:** COCO train
  - **Train (cal.):** COCO mini-train
    - with [1k, 2.5k, 5k, 10k] images
    - five subsets are randomly sampled for each train-set
- **Val Data**
  - **Val (ID):** COCO mini-val (2.5k imgs)

Takeaway → After ~5k images, the calibration is stable



(a) PS D-ECE-optimized

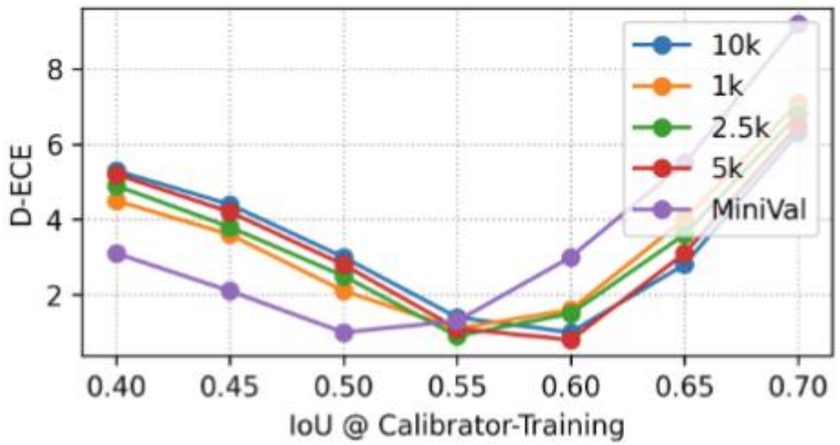


(b) IR D-ECE-optimized

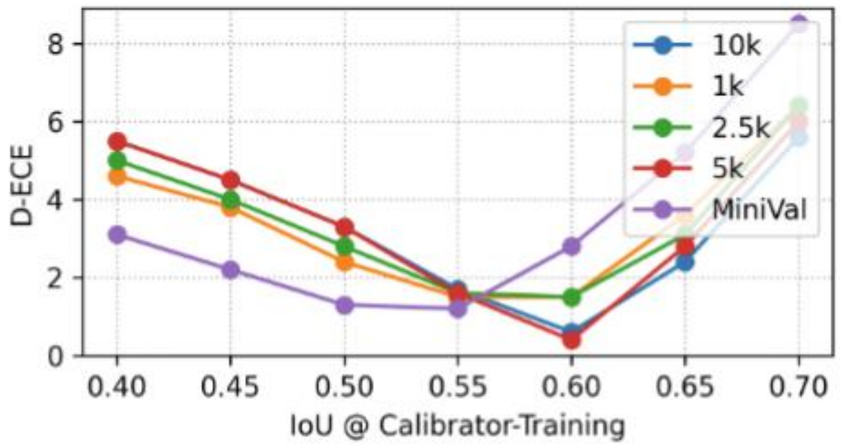
Individual results → red dots
the mean performance for each calibration → green star

Ablation Study
Choosing the IoU Threshold τ

- **Model:** Deformable DETR trained and evaluated on COCO
- Swept τ from 0.40 to 0.70 for calibrator training.
- **Held-out MiniVal baseline:** best at τ =0.50 ➔ no adjustment needed.
- **Training-data calibration:** best at τ =0.60, reaching D-ECE 0.4.



(a) D-ECE for PS



(b) D-ECE for IR

Discussion and Conclusion

Technische Hochschule mannheim

CeMOS Intelligente Systeme

Takeaways

1. **No held-out data needed:** We calibrate directly on training data ➔ no accuracy or evaluation data sacrificed.
2. **Simple regularizer:** A stricter IoU threshold during calibrator training alone prevents overconfidence bias.
3. **State-of-the-art results:** D-ECE of 0.4 on COCO ➔ beats both train-time (Cal-DETR: 8.7) and held-out post-hoc (1.3) methods.

Limitations

1. The optimal IoU threshold tau is dataset- and detector-dependent ➔ requires a small search.
2. Under some out-of-domain shifts (e.g. BDD100K), gains over baselines are smaller and mixed across metrics.
3. The IoU-threshold regularization is specific to object detection and does not directly transfer to other tasks (e.g. classification).

Technische Hochschule mannheim

CeMOS Intelligente Systeme

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Efficient Post-hoc Calibration in Object Detection without Held-Out Data

Nikolas Ebert, Didier Stricker, Oliver Wasenmüller

17 August 2026 | International Conference on Pattern Recognition 2026