Cumulative Consensus Score: Label-Free and Model-Agnostic Evaluation of Object Detectors in Deployment

Avinaash Manoharan[1,2], Xiangyu Yin[3], Domenik Helms[1], and Chih-Hong Cheng[2,3]

***Abstract*— Evaluating object detection models in deployment is challenging because ground-truth annotations are rarely available. We introduce the Cumulative Consensus Score (CCS), a label-free monitoring signal for continuous evaluation and comparison of detectors in real-world settings. CCS applies test-time data augmentation to each image and measures the spatial consistency of predicted bounding boxes across augmented views using Intersection over Union. The resulting consensus score serves as a proxy for reliability without requiring bounding box annotations. In controlled experiments on Open Images and KITTI, CCS achieved over 90% congruence with F1-score, Probabilistic Detection Quality, and Optimal Correction Cost, with qualitative consistency further confirmed on COCO and BDD100K across model pairs. The method is model-agnostic, working across single-stage and two-stage detectors, and operates at the case level to highlight under-performing scenarios. We also provide a simplified theoretical link between expected CCS and detection correctness. Altogether, CCS provides a robust foundation for DevOps-style monitoring of object detectors.**

I. Introduction

Deep learning has enabled major advances in computer vision, particularly in safety-critical domains such as autonomous driving. However, the reliability of perception modules, including object detection, remains a concern. Object detectors can make errors under distribution shifts and inherently suffer from epistemic uncertainty, stemming from incomplete training data and limited coverage of real-world conditions [1]–[3]. This uncertainty makes it difficult for engineers and end users to judge whether a newly trained object detector is more trustworthy than an existing, well-established one. Ground-truth annotations are typically required for supervised metrics (e.g., mAP, F1-score), yet such labels are rarely available at deployment time. This creates a gap between controlled lab evaluation and the operational domain, where continuous monitoring and safe upgrades are most needed.

To address this challenge, we introduce the *Cumulative Consensus Score (CCS)*, a label-free method for evaluating object detectors in deployment. CCS builds on the idea of functional monitoring by enabling direct comparison of a new detector against an existing baseline without requiring annotated data. Its central principle is to assess the *spatial consistency of predictions*: detectors that generalize better are expected to exhibit more stable outputs when the input is subjected to benign transformations. By turning this stability into a measurable signal, CCS provides a practical proxy for reliability. In addition, we provide a simplified theoretical analysis that links CCS to detection correctness under an idealized setting, offering intuition for why spatial consensus reflects reliability.

A key difficulty is that many uncertainty estimation techniques require architectural changes or large ensembles, making comparisons costly. CCS instead uses Test-Time Data Augmentation (TTDA) and requires no additional training. In practice, several photometric variations of an image are processed to produce bounding boxes, and their overlap is quantified using Intersection over Union (IoU). CCS aggregates these overlaps into a single image-level consensus score, enabling comparison between two detectors.

We validate CCS in controlled experiments by comparing its outputs against ground-truth-based metrics including F1 score, Probabilistic Detection Quality (pPDQ), and Optimal Correction Cost (OC-cost). Experiments across datasets such as Open Images [4] and KITTI [5], and across architectures including Faster R-CNN [6], RetinaNet [7] and SSD [8] show that CCS achieves over 90% congruence with these established measures, with qualitative consistency further confirmed on COCO [9] and BDD100K [10]. Importantly, CCS provides results at the image level, enabling the identification of under-performing cases where predictions become unstable across benign transformations allowing engineers to pinpoint problematic inputs and guide targeted improvements.

II. Related Work

Several label-free indicators have been explored. Schmidt et al. [11] introduced a consensus score in an ensemble-based active learning setting, relying on Bayesian variants [12], [13] that increase deployment cost. Yang et al. [14] proposed the Box Stability Score (BoS) and Yu et al. [15] proposed the Detection Adaptation Score (DAS); both correlate stability-related signals with supervised metrics but require internal feature access and mainly target training-time model selection. Yu et al. [16] extended consistency ideas to active learning (CALD), reducing labeling but still relying on labeled retraining. In contrast, CCS operates directly on detector outputs under TTDA, requires neither ensembles nor feature access, and provides a per-image monitoring signal based on augmentation-induced spatial consistency. We further provide a simplified theoretical link between CCS and detection correctness under an idealized setting.

Test-time data augmentation is commonly used to improve robustness and accuracy [17], [18]; CCS instead uses TTDA

[1] DLR Institute of Systems Engineering for Future Mobility, Germany
[2] Carl von Ossietzky University of Oldenburg, Germany
[3] Chalmers University of Technology, Sweden
Contact: avinaash.manoharan@dlr.de, chih-hong.cheng@uol.de

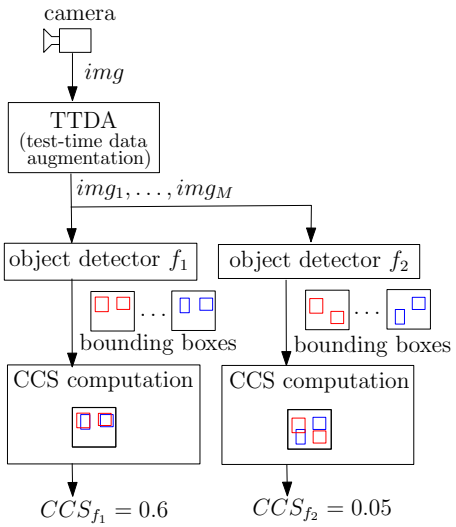


Fig. 1: Workflow of comparing the result of two object detectors f_1, f_2 using the corresponding CCS values.

to quantify prediction stability as a monitoring signal. Probabilistic object detection explicitly models localization and classification uncertainty [19]–[21], but typically requires architectural modifications or ensembles and is detector-specific. Ground-truth-based metrics such as pPDQ [2] and OC-cost [22] are not applicable without labels. CCS complements these directions by providing a lightweight, model-agnostic proxy for spatial uncertainty that aligns with supervised metrics when labels exist, while remaining directly usable in label-free deployment settings.

III. Inside Cumulative Consensus Score

This section describes the methodology for computing the Cumulative Consensus Score (CCS) at the image level. Fig. 1 illustrates the workflow and the mindset for how CCS is designed, as well as comparing the result of two object detectors f_1 and f_2. Given an image img, TTDA is conducted to build M variations using TTDA techniques i.e. photometric augmentations such as shifting weather conditions. As there is no shearing or cropping of images in our augmentation pipeline, for all augmented images, any object (e.g., Car) should be located at the same location in the image plane. In this regard, a *necessary condition* of a high-performing object detector is to generate bounding boxes among images that induce a larger overlap. In Fig. 1, object detector f_1 performs better, as the bounding boxes being produced across different images have higher overlap. As shown in subsequent paragraphs, CCS utilizes such a concept and characterizes the degree of overlap by computing an aggregated value utilizing the standard IoU computation across augmented images.

A. Defining CCS

1) Basic setting: Let M denote the number of augmentations applied to each input image. For a given image, let N_i and N_j denote the number of bounding boxes predicted by the detector on the i^{th} and j^{th} augmentations, respectively.

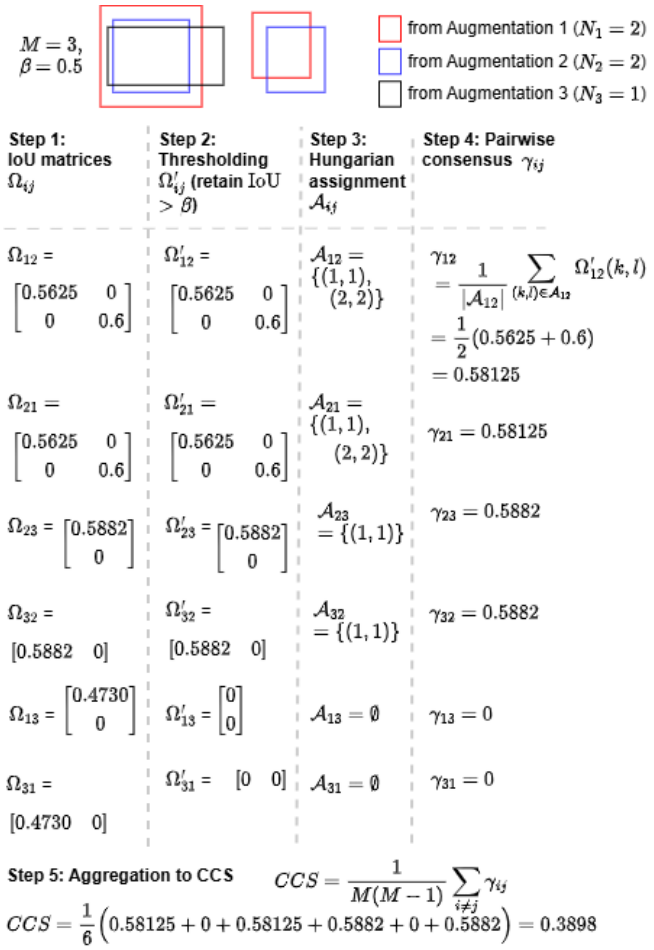


Fig. 2: Predictions from three augmentations shown in different coloured boxes and the associated CCS computation.

For each ordered pair (i, j) with $i \neq j$, we compute the Intersection over Union (IoU) matrix

$$\Omega_{ij} \in \mathbb{R}^{N_i \times N_j},$$

where each entry $\Omega_{ij}(k, l)$ represents the IoU between the k^{th} predicted bounding box from augmentation i and the l^{th} predicted bounding box from augmentation j. Since N_i and N_j may differ, Ω_{ij} is not necessarily a square matrix.

As an illustrative example, consider the three augmentations shown in Fig. 2, where the red, blue, and black bounding boxes correspond to augmentations 1, 2, and 3, respectively. For the pair $(1, 2)$, the IoU matrix is

$$\Omega_{12} = \begin{bmatrix} 0.5625 & 0 \\ 0 & 0.6 \end{bmatrix},$$

indicating that the left red box overlaps strongly with the left blue box (IoU 0.5625), and the right red box overlaps with the right blue box (IoU 0.6), while cross-overlaps are zero. Similarly, for the pair $(2, 3)$,

$$\Omega_{23} = \begin{bmatrix} 0.5882 \\ 0 \end{bmatrix},$$

since augmentation 2 produces two detections while augmentation 3 produces only one.

2) Apply thresholding to the IoU matrix: To suppress weak overlaps, we apply a threshold β to obtain a filtered IoU matrix Ω'_{ij}:

$$\Omega'_{ij}(k,l) \stackrel{\text{def}}{:=} \begin{cases} \Omega_{ij}(k,l), & \text{if } \Omega_{ij}(k,l) > \beta, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Only IoU values strictly greater than β are retained. This step removes minor overlaps and emphasizes meaningful spatial agreement. In our example with $\beta = 0.5$, the entries in Ω_{12} remain unchanged since both non-zero values exceed the threshold, whereas in Ω_{13} the value 0.4730 is suppressed to zero, resulting in Ω'_{13} being identically zero. The choice of β follows the convention established in [23], where an IoU of at least 0.5 is considered the minimum requirement for associating a detection with a ground-truth object.

3) Assignment-based consensus computation: After thresholding the IoU matrix Ω_{ij} to obtain Ω'_{ij}, we determine correspondences between detections from the i^{th} and j^{th} augmentations using a one-to-one assignment formulation. This step resolves ambiguities that arise when multiple objects are present and when the numbers of detections differ across augmentations.

Let $\Omega'_{ij} \in \mathbb{R}^{N_i \times N_j}$ denote the filtered IoU matrix. To identify consistent detection pairs, we solve a linear assignment problem on Ω'_{ij} using the Hungarian algorithm. The assignment maximizes the total retained IoU while ensuring that each detection participates in at most one correspondence. Let $\mathcal{A}_{ij} \subseteq \{1,\dots,N_i\} \times \{1,\dots,N_j\}$ denote the set of matched index pairs returned by the Hungarian algorithm. The number of matched pairs is denoted by $|\mathcal{A}_{ij}|$.

We define the assignment matrix $\Pi_{ij} \in \mathbb{R}^{N_i \times N_j}$ as

$$\Pi_{ij}(k,l) = \begin{cases} \frac{1}{|\mathcal{A}_{ij}|}, & \text{if } (k,l) \in \mathcal{A}_{ij}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

By construction,

$$\sum_{k=1}^{N_i} \sum_{l=1}^{N_j} \Pi_{ij}(k,l) = 1. \tag{3}$$

Using Π_{ij}, the pair-wise consensus score γ_{ij} is defined as

$$\gamma_{ij} \stackrel{\text{def}}{:=} \sum_{k=1}^{N_i} \sum_{l=1}^{N_j} \Pi_{ij}(k,l)\, \Omega'_{ij}(k,l). \tag{4}$$

Since Π_{ij} distributes uniform mass over the matched pairs, the above expression simplifies to

$$\gamma_{ij} = \frac{1}{|\mathcal{A}_{ij}|} \sum_{(k,l)\in\mathcal{A}_{ij}} \Omega'_{ij}(k,l). \tag{5}$$

The case $|\mathcal{A}_{ij}| = 0$ arises when no detection pairs satisfy the IoU threshold, or when one of the augmentations yields no detections. In this situation, the set $\mathcal{A}_{ij}$ is empty, so the sum $\sum_{(k,l)\in\mathcal{A}_{ij}} \Omega'_{ij}(k,l)$ contains no terms and the normalization by $|\mathcal{A}_{ij}|$ would be undefined. This includes the case where both augmentations yield no detections, in which Ω'_{ij} does not exist, as well as the case where Ω'_{ij} is identically zero after thresholding; in both situations we define $\gamma_{ij} = 0$.

4) Calculate CCS via aggregation: After computing the pairwise consensus score γ_{ij} for each ordered augmentation pair (i,j) with $i \neq j$, we obtain the Cumulative Consensus Score (CCS) by averaging across all such pairs:

$$CCS \stackrel{\text{def}}{:=} \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} \gamma_{ij}. \tag{6}$$

The normalization factor $M(M-1)$ accounts for all ordered augmentation pairs and ensures that CCS remains independent of the number of augmentations applied. By aggregating consensus across all pairs, CCS provides a normalized measure of localization consistency for the image that is independent of the number of augmentations applied.

B. Minimal Link Between CCS and Detection Correctness

In a highly simplified setting, we provide the following result to build intuition for why the TTDA agreement reflected by CCS can correlate with detection correctness.

a) Setting: Assume a single object exists in the image. We apply k augmentations. For augmentation $t \in \{1,\dots,k\}$, let $X_t \in \{0,1\}$ indicate whether the detector is *correct*. If $X_t = 1$, the detector outputs a perfect bounding box; otherwise no box is produced. We consider two augmented views to be consistent if they yield the same outcome.

b) CCS under i incorrect augmentations: Let i be the number of incorrect augmentations, hence $k-i$ are correct. Among the $\binom{k}{2}$ unordered augmentation pairs, the number of consistent pairs equals $\binom{k-i}{2} + \binom{i}{2}$. Therefore, the CCS value conditioned on i incorrect augmentations is

$$CCS_k(i) = \frac{1}{\binom{k}{2}} \left(\binom{k-i}{2} + \binom{i}{2} \right). \tag{7}$$

c) Expected CCS with correctness p: Assuming i.i.d. trials $X_t \sim \text{Bernoulli}(p)$, where p denotes the detector's augmentation-wise correctness, the number of incorrect augmentations i follows a binomial distribution. Thus, the expected CCS is

$$\mathbb{E}[CCS_k] = \sum_{i=0}^{k} \binom{k}{i} p^{k-i} (1-p)^i \left[\frac{1}{\binom{k}{2}} \left(\binom{k-i}{2} + \binom{i}{2} \right) \right]. \tag{8}$$

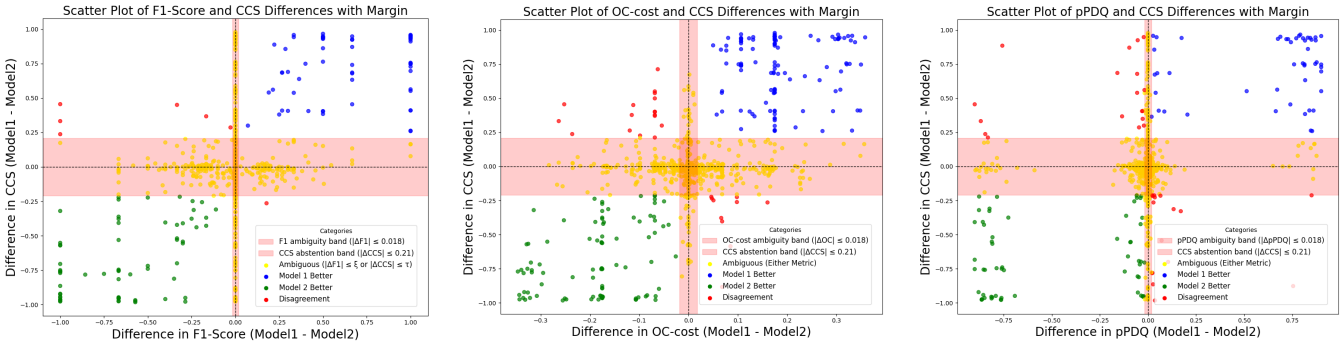
Lemma 1: Consider two object detectors f_1 and f_2 with augmentation-wise correctness p_1 and p_2, respectively, under the idealized single-object setting in Sec. III-B. Let CCS_k be defined as in Eq. (7)–Eq. (8). For any margin $\Omega \geq 0$, we have

$$\mathbb{E}[CCS_k^{f_1}] - \mathbb{E}[CCS_k^{f_2}] > \Omega \tag{9}$$

if and only if

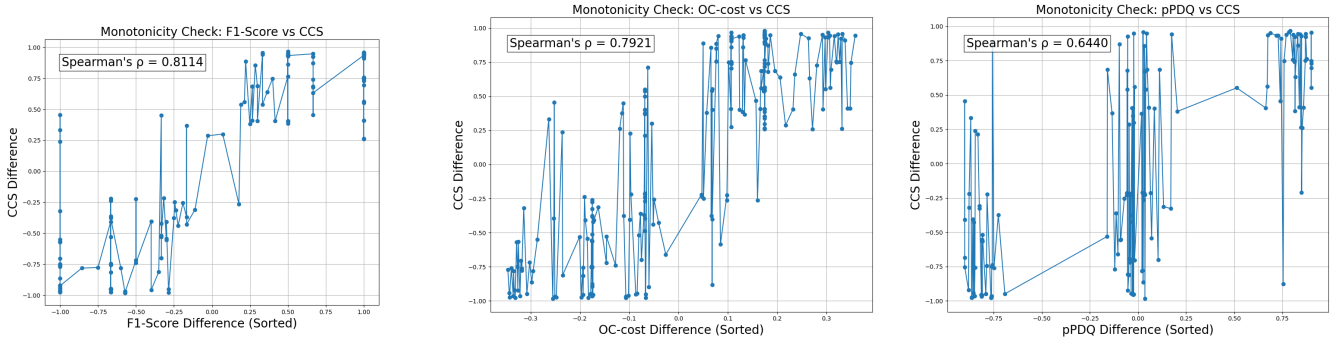
$$\sum_{i=0}^{k} \tbinom{k}{i} \left(p_1^{k-i}(1-p_1)^i - p_2^{k-i}(1-p_2)^i \right) \left[\frac{1}{\binom{k}{2}} \left(\binom{k-i}{2} + \binom{i}{2} \right) \right] > \Omega. \tag{10}$$

Proof: The creation of Eq. (10) immediately follows by integrating Eq. (8) into Eq. (9). ■



(a) Comparison of F1-Score and CCS (b) Comparison of OC-cost and CCS (c) Comparison of pPDQ and CCS

Fig. 3: Scatter plots comparing CCS with established metrics: F1-Score, pPDQ, and OC-cost.



(a) Sorted trend of F1-Score and CCS (b) Sorted trend of OC-cost and CCS (c) Sorted trend of pPDQ and CCS

Fig. 4: Sorted trend analysis showing the alignment of CCS with F1-Score, pPDQ, and OC-cost.

The following result shows that, under simplifying theoretical assumptions, a more performant detector also has a higher CCS value.

Lemma 2 (Monotonicity intuition): For any $k \geq 2$ and $1 > p_1 > p_2 > 0.5$, we have $\mathbb{E}[CCS_k^{f_1}] > \mathbb{E}[CCS_k^{f_2}]$.

Proof: Consider $k \geq 2$ i.i.d. Bernoulli trials $X_1, \ldots, X_k$ with $\Pr(X = 1) = p$. Let i be the number of failures, so $i \sim \mathrm{Bin}(k, 1-p)$. The quantity $CCS_k^f \stackrel{\text{def}}{:=} \frac{\binom{k-i}{2}+\binom{i}{2}}{\binom{k}{2}}$ is actually the fraction of unordered pairs (a, b) with $a < b$ that have the same outcome (both successes or both failures). Therefore, by linearity of expectation, its expectation equals the probability that a single pair matches:

$$\begin{aligned}\mathbb{E}\Big[CCS_k^f\Big] &= \Pr(X_a = X_b) \\ &= \Pr(X_a = 1, X_b = 1) + \Pr(X_a = 0, X_b = 0) \\ &= p^2 + (1-p)^2. \end{aligned} \tag{11}$$

Hence

$$\begin{aligned}&\mathbb{E}\Big[CCS_k^{f_1}\Big]\Big|_{p=p_1} - \mathbb{E}\Big[CCS_k^{f_2}\Big]\Big|_{p=p_2} \\ &= \Big(p_1^2 + (1-p_1)^2\Big) - \Big(p_2^2 + (1-p_2)^2\Big) \\ &= 2(p_1 - p_2)(p_1 + p_2 - 1). \end{aligned} \tag{12}$$

If $1 > p_1 > p_2 > \frac{1}{2}$, then $p_1 - p_2 > 0$ and $p_1 + p_2 - 1 > 0$, so

$$\mathbb{E}\Big[CCS_k^{f_1}\Big]\Big|_{p=p_1} > \mathbb{E}\Big[CCS_k^{f_2}\Big]\Big|_{p=p_2}.$$

■

Note that our theoretical result is derived in an idealized setting and does not model localization noise, multiple objects, or class confusion. It therefore serves primarily to justify the monotonic intuition. While the proof is presented in this simplified regime for analytical clarity, our experiments in later sections show that the same ordering trend holds in realistic settings with noisy localization and multi-object assignment. In practice, localization noise makes IoU < 1 and multi-object scenes require assignment. Nevertheless, this relationship motivates CCS as a deployment signal. Our implementation adopts a conservative variant by setting $\gamma_{ij} = 0$ when both augmentations yield no detections, so Lemma 2 serves as intuition rather than an exact characterization of our implementation.

IV. Experiments and Results

A. Experimental Setup

1) Dataset: We train object detectors on Open Images Dataset (OID) [4], [24] and evaluate on KITTI [5], [25].

2) Evaluation metrics: We compare CCS against F1-score, OC-cost, and pPDQ.

F1-score: Following [3], we apply one-to-one matching at IoU threshold α_{iou}. A prediction is a TP if its class matches the ground truth and IoU $\geq \alpha_{\text{iou}}$, otherwise FP; unmatched ground-truth boxes are FN. The F1-score is

$$\mathrm{F1} = 2\,\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{13}$$

TABLE I: Comparison of old and new object detectors trained on 2000 Car and 2000 Truck samples from OID [24].

Metric	**Model 1:** FRCNN (50 ep.)	**Model 2:** RetinaNet (100 ep.)
COCO mAP (baseline)	0.384	0.409
mAP (after training)	0.4307	0.4979
Validation Loss	0.1872	0.1746

OC-cost: OC-cost [22] evaluates detectors via minimum-cost bipartite matching between predictions and ground truths. Unmatched boxes incur a constant penalty β_{dummy} (set to 0.6, which aligns with human preference as recommended in [22]). The per pair matching cost is

$$C_{\text{total}} = \lambda C_{\text{loc}} + (1-\lambda) C_{\text{cls}}, \tag{14}$$

where $C_{\text{loc}} = 1 - \text{IoU}$. In our setup, we set $\lambda = 1$ to focus exclusively on localization.

pPDQ: pPDQ [2] combines spatial and semantic uncertainty via the geometric mean between a ground truth $\mathcal{G}_i^{\text{img}}$ and a detection $\mathcal{D}_j^{\text{img}}$:

$$\text{pPDQ}(\mathcal{G}_i^{\text{img}}, \mathcal{D}_j^{\text{img}}) = \sqrt{Q_S(\mathcal{G}_i^{\text{img}}, \mathcal{D}_j^{\text{img}}) \cdot Q_L(\mathcal{G}_i^{\text{img}}, \mathcal{D}_j^{\text{img}})}. \tag{15}$$

Here Q_S measures how well the spatial probability distribution of a detection aligns with the ground truth, while Q_L is the probability the detector assigns to the correct class label. Unlike a simple L_2 norm, Q_S is formally defined as

$$Q_S(\mathcal{G}_i^{\text{img}}, \mathcal{D}_j^{\text{img}}) = \exp\Big(- \big[L_{\text{FG}}(\mathcal{G}_i^{\text{img}}, \mathcal{D}_j^{\text{img}}) + L_{\text{BG}}(\mathcal{G}_i^{\text{img}}, \mathcal{D}_j^{\text{img}}) \big] \Big). \tag{16}$$

where L_{FG} is the average negative log-probability assigned to foreground pixels and L_{BG} penalizes probability mass assigned outside the ground-truth bounding box.

Implementation note. Although PDQ was originally defined for pixel-accurate masks [2], we adapt it to bounding boxes by assigning probability $1-\epsilon$ inside the predicted box and ϵ outside, with $\epsilon = 0.1$ to enforce strict localization penalties. As the evaluated detectors here (Faster R-CNN, RetinaNet, SSD) are deterministic and do not output explicit spatial uncertainty distributions, this adaptation (with fixed ϵ) should be interpreted as a standardized approximation for consistent supervised comparison across deterministic architectures. By this, PDQ serves as a supervised baseline metric rather than a full probabilistic evaluation.

B. Congruence of CCS with Established Metrics

This subsection evaluates how well CCS aligns with standard supervised metrics (Sec. IV-A.2). In Fig. 1, each image is processed by two object detectors f_1 and f_2 (Tab. I). For each supervised metric $M \in \{\text{F1, pPDQ, OC-Cost}\}$, we obtain per-image scores M_{f_1} and M_{f_2} and form the deltas

$$\Delta M = M_{f_1} - M_{f_2}. \tag{17}$$

Similarly, the difference of CCS between the two models is

$$\Delta CCS = CCS_{f_1} - CCS_{f_2}. \tag{18}$$

TABLE II: Effect of percentile p on CCS abstention and alignment for the illustrative model pair (KITTI test set).

p (%)	τ	**Keep (unamb.)**	**Abstain (amb.)**	ρ	**Congr.**
50	0.0287	0.6731	0.5008	0.7585	0.7510
75	0.2134	0.3626	0.7504	0.7933	0.9470
90	0.6652	0.2060	0.8995	0.7398	1.0000
95	0.9059	0.1154	0.9498	0.7090	1.0000

Fig. 3 plots ΔCCS against ΔM for the model pairs under study. To avoid attributing meaning to near-ties, we introduce (i) an *ambiguity band* for the supervised reference, and (ii) a *CCS abstention band* around zero.

a) F1 ambiguity band (ξ): We use F1 as the reference to determine whether supervised evaluation is image-wise conclusive. For a fixed model pair and test set, we compute per-image ΔF1 values (Eq. (17)). We then estimate the uncertainty of the *mean* ΔF1 via nonparametric bootstrap and define ξ as the half-width of the resulting 95% confidence interval (CI). Images with $|\Delta\text{F1}| \le \xi$ are treated as *F1-ambiguous*, while images with $|\Delta\text{F1}| > \xi$ are treated as *F1-unambiguous*.

b) Calibrating CCS abstention threshold (τ) from F1-ambiguous images: We calibrate τ using the F1-ambiguous subset so that CCS abstains on a controlled fraction of cases where the supervised reference is itself inconclusive. Let $\mathcal{A}_{\text{F1}} = \{\text{images} : |\Delta\text{F1}| \le \xi\}$. We set

$$\tau(p) \;=\; \text{Percentile}_p\big(\{\, |\Delta CCS| : \text{image} \in \mathcal{A}_{\text{F1}} \,\}\big), \tag{19}$$

and report a sweep over $p \in \{50, 75, 90, 95\}$. A larger p yields a more conservative τ (higher abstention), while a smaller p yields a less conservative τ (higher coverage).

c) Kept set definition: For a given (ξ, τ), we restrict analysis to images where *both* the supervised reference and CCS express a confident preference:

$$\mathcal{K}(\xi, \tau) \;=\; \{\text{images} : |\Delta\text{F1}| > \xi \;\wedge\; |\Delta CCS| > \tau\}. \tag{20}$$

Images outside $\mathcal{K}(\xi, \tau)$ are treated as abstentions, either because the supervised reference is inconclusive (F1-ambiguous) or because CCS assigns a near-zero difference.

d) Percentile selection for the illustrative model pair: Tab. II summarizes the effect of varying the percentile p used to define τ, where τ is computed from the distribution of $|\Delta CCS|$ restricted to the F1-ambiguous subset determined by ξ (i.e., $|\Delta\text{F1}| \le \xi$). Thus, the sweep is performed with respect to the ΔF1-based calibration.

Lower percentiles (e.g., p=50) are permissive, retaining many unambiguous cases but abstaining on only half of the ambiguous subset, which reduces congruence and ranking consistency. Higher percentiles (p=90 and 95) enforce strict abstention, eliminating nearly all ambiguous cases and achieving perfect sign agreement, but at the cost of retaining too few samples and lowering Spearman's ρ.

The choice p=75 provides the best trade-off: it abstains on approximately 75% of ambiguous cases while retaining a sufficient number of unambiguous images for reliable correlation analysis, and yields the highest observed Spearman correlation together with high congruence. The resulting τ

TABLE III: Alignment of CCS with supervised metrics for the illustrative model pair in Tab. I. All results are computed over 1000 test images (KITTI test set). Calibration uses ΔF1 ($\xi = 0.018$, $\tau = 0.21$ at the 75th percentile).

Metric	Yel.	Kept	Grn.	Blu.	Red	Congr.(%)	ρ
ΔF1 vs ΔCCS	870	130	65	59	6	95.38	0.8114
ΔOC-cost vs ΔCCS	766	234	94	114	26	88.89	0.7921
ΔpPDQ vs ΔCCS	860	140	58	49	33	76.43	0.6440

is then fixed and used consistently when comparing CCS against ΔOC-Cost and ΔpPDQ.

C. Observations from Fig. 3

In Fig. 3, each point corresponds to one test image. Colors encode agreement between ΔM and ΔCCS, with an explicit abstention region:

- **Yellow (abstain):** at least one measure abstains, i.e., $|\Delta\text{F1}| \leq \xi$ *or* $|\Delta CCS| \leq \tau$. This includes (a) supervised near-ties (F1-ambiguous images) and (b) cases where CCS itself assigns near-zero difference and therefore does not support a confident ranking.
- **Blue (agree, f_1 better):** $\Delta M > 0$ and $\Delta CCS > 0$ on the kept set $\mathcal{K}(\xi, \tau)$.
- **Green (agree, f_2 better):** $\Delta M < 0$ and $\Delta CCS < 0$ on the kept set $\mathcal{K}(\xi, \tau)$.
- **Red (disagree):** $\text{sign}(\Delta M) \neq \text{sign}(\Delta CCS)$ on the kept set $\mathcal{K}(\xi, \tau)$.

Thus, blue/green/red points correspond to images for which both the supervised reference and CCS are confident, while yellow points are deliberately excluded from congruence and Spearman's ρ computations to avoid over-interpreting near-ties.

D. Sorted Trend Analysis for Congruence Assessment

Fig. 3 presents scatter plots of ΔCCS against ΔM for $M \in \{\text{F1, pPDQ, OC-Cost}\}$. Each point corresponds to one test image, and agreement categories are determined on the kept set $\mathcal{K}(\xi, \tau)$ defined in Eq. (20).

To assess whether CCS preserves the ordering induced by a supervised metric, we perform a sorted trend analysis (Fig. 4). For each metric M, images in $\mathcal{K}(\xi, \tau)$ are sorted according to ΔM, and the corresponding ΔCCS values are plotted in the same order. If CCS is consistent with M, the resulting curve exhibits a monotonic trend: images strongly favoring f_1 under M (large positive ΔM) tend to have positive ΔCCS, and images favoring f_2 tend to have negative ΔCCS. Agreement on $\mathcal{K}(\xi, \tau)$ is quantified using two complementary measures. First, directional agreement (congruence) is defined as $\mathbb{P}\big[\text{sign}(\Delta M) = \text{sign}(\Delta CCS)\big]$, which, in empirical form, equals $(\text{Green} + \text{Blue})/\text{Kept} \times 100\%$. Second, we compute Spearman's rank correlation coefficient $\rho(\Delta M, \Delta CCS)$, which evaluates monotonic consistency between the two quantities.

Spearman's ρ is used instead of a simple vector difference or Pearson correlation because the objective is to assess ranking consistency rather than absolute scale agreement. Different metrics may vary in magnitude or exhibit nonlinear scaling while preserving the same ordering, which Spearman's ρ captures directly.

E. Inference from Tab. III

Tab. III analyzes the illustrative model pair in Tab. I using calibration parameters ξ and τ derived from ΔF1. A substantial portion of images lies in the abstention region, reflecting the similar supervised performance of the two detectors and the prevalence of image-level near-ties captured by the ambiguity band ξ. On the kept set (Eq. (20)), CCS exhibits strong directional agreement with ΔF1 (95.38%) and high monotonic consistency ($\rho = 0.8114$). Under the same calibration, alignment with ΔOC-cost and ΔpPDQ remains substantial (88.89% and 76.43%; $\rho = 0.7921$ and 0.6440); the weaker pPDQ alignment is expected, as pPDQ jointly penalizes localization and classification confidence, a dimension CCS does not capture.

F. Inference from Tab. IV

Tab. IV extends the analysis to heterogeneous comparison scenarios. For each model pair, the ambiguity band ξ is computed via bootstrap confidence intervals on the mean ΔF1. The CCS abstention threshold τ is then derived from the distribution of $|\Delta\text{CCS}|$ within the F1-ambiguous subset. While the percentile used to compute τ (75th percentile in our experiments) controls the strictness of abstention, the calibration procedure itself remains data-driven.

Across all scenarios, CCS maintains strong directional agreement with supervised metrics, particularly for ΔF1 and ΔOC-cost, and consistently high Spearman correlations. In settings where the imbalance between green and blue counts becomes pronounced, Spearman's ρ decreases, as it measures rank consistency rather than mere sign agreement. Nevertheless, the overall congruence remains high, indicating that CCS continues to identify the correct direction of performance difference even when magnitude ordering becomes less strictly monotonic.

Taken together, the results from both the illustrative example (Tab. III) and the heterogeneous comparison scenarios (Tab. IV) show that whenever supervised evaluation yields a decisive preference, CCS preserves both the direction and the relative ordering of model performance differences. This consistency across architectures, training scales, and performance gaps supports the role of CCS as a model-agnostic, label-free monitoring signal.

G. Comparison with Deployment-Time Output Signals

To assess whether simpler label-free heuristics could substitute CCS, we compare CCS against scalar signals derived directly from detector outputs without ground-truth annotations. We evaluate three commonly used proxy indicators: (i) *mean detection confidence* [26], [27]; (ii) *detection count stability*, defined as the standard deviation of the number of detections across augmentations for each image [28]; and (iii) *naïve IoU consistency*, measuring average spatial overlap agreement across augmentations without structured object association [29].

TABLE IV: Alignment of CCS with supervised metrics across diverse model comparison scenarios. All results are computed over 1000 test images from KITTI. For each model pair, $\boldsymbol{\xi}$ (bootstrap CI half-width of mean ΔF1) and $\boldsymbol{\tau}$ (75th percentile of $|\Delta\mathrm{CCS}|$ over F1-ambiguous images) are shown in the model column. Congruence and Spearman's ρ are computed on the kept set.

Training Setup	Model Pair	ΔM	Yel.	Kept	Grn.	Blu.	Red	Congr.	ρ
Cross-Architecture Images/class: 6000 Labels: Car, Truck	Model1: Faster R-CNN ResNet-101 2 epochs, mAP: 0.3308 Model2: SSD300 VGG16 250 epochs, mAP: 0.5053 $\boxed{\xi = 0.016,\ \tau = 0.119}$	ΔF1	852	148	81	54	13	91.22	0.8085
		ΔOC-cost	742	258	140	90	28	89.15	0.7482
		ΔpPDQ	825	175	71	71	33	81.14	0.7250
Same Architecture Old: 2000 imgs/class New: 6000 imgs/class	Model1: RetinaNet ResNet-50 50 epochs, mAP: 0.4837 Model2: RetinaNet ResNet-50 150 epochs, mAP: 0.5194 $\boxed{\xi = 0.016,\ \tau = 0.111}$	ΔF1	872	128	59	55	14	89.06	0.8264
		ΔOC-cost	780	220	115	86	19	91.36	0.7676
		ΔpPDQ	855	145	52	51	42	71.03	0.4984
Same Architecture Weak vs Strong Old: 1 epoch, 2000/class New: 50 epochs, 6000/class	Model1: Faster R-CNN ResNet-50 1 epoch, mAP: 0.4472 Model2: Faster R-CNN ResNet-50 50 epochs, mAP: 0.5098 $\boxed{\xi = 0.017,\ \tau = 0.062}$	ΔF1	783	217	113	46	58	73.27	0.7031
		ΔOC-cost	740	260	140	60	60	76.92	0.6662
		ΔpPDQ	772	228	112	30	86	62.28	0.4049

TABLE V: Alignment of deployment-time heuristic signals with supervised ranking (ΔF1) for the illustrative model pair (1000 KITTI test images). Calibration uses $\xi = 0.0181$. For each heuristic H, a signal-specific threshold τ_H is computed.

Signal	ρ	Congr.(%)	Kept	Inconcl.(%)	τ_H
Mean Confidence	-0.09	55.41	74	92.61	1.00
Detection Count Stability	0.07	46.32	95	90.51	1.50
Naïve IoU Consistency	0.01	56.00	100	90.01	0.52

For a model pair (f_1, f_2), we compute image-level differences $\Delta H = H_{f_1} - H_{f_2}$ and compare them against ΔF1 using the same ranking protocol as CCS. Since these signals have different numeric scales, we compute a signal-specific abstention threshold τ_H for each metric, defined as the 75th percentile of $|\Delta H|$ over the F1-ambiguous subset ($|\Delta\mathrm{F1}| \leq \xi$). Agreement is evaluated on the kept set given by ($|\Delta\mathrm{F1}| > \xi \ \wedge\ |\Delta H| > \tau_H$).

Table V reports results for the illustrative model pair of Tab. I. Despite scale-aware calibration, all heuristics exhibit negligible monotonic alignment with supervised ranking ($|\rho| < 0.1$) and directional agreement close to random consistency. Over 90% of images are classified as inconclusive under this protocol, resulting in small kept subsets.

In contrast, CCS (Tab. III) achieves substantially stronger monotonic and directional alignment under identical calibration. This pattern holds across the additional model-comparison scenarios in Tab. IV, where heuristic signals consistently exhibit weaker correlation relative to CCS.

H. Robustness to Augmentation Seeds, Architectures, and Datasets

Since CCS relies on test-time data augmentation (TTDA), we verify that its alignment with supervised metrics is not sensitive to the choice of augmentation seed. Tab. VI reports representative results for five seeds under the illustrative detector setup of Tab. I (15 seeds were evaluated in total). Spearman's ρ values remain tightly clustered: for ΔF1 vs. ΔCCS we obtain $\rho = 0.8205 \pm 0.0063$, for ΔOC-cost vs. ΔCCS, $\rho = 0.7869 \pm 0.0046$, and for ΔpPDQ vs. ΔCCS, $\rho = 0.6325 \pm 0.0080$. The small standard deviations indicate stable ranking consistency across seeds. Robustness is further supported by consistent behavior across architectures and training regimes (Tab. IV).

a) TTDA configuration: We employ nine mild, non-geometric photometric transformations per image (*brightness, contrast, blur, noise, sharpen, color shift*), with parameters sampled from narrow, bounded intervals (e.g., brightness $\alpha \in [0.9, 1.1]$, $\beta \in [-10, 10]$; Gaussian noise $\sigma \in [0.005, 0.01]$; blur kernel $\in \{3, 5\}$). These augmentations correspond to common corruption families used in robustness benchmarks which emulate realistic sensor or illumination variation rather than adversarial distortions [29], [30] and preserve semantic labels as recommended in augmentation surveys [31], [32].

Geometric transformations are excluded so that CCS evaluates stability under appearance variation rather than systematic spatial displacement. As augmentation parameters are drawn from fixed, narrow ranges, changing the seed alters only the sampled value within the interval, explaining the observed stability of ρ.

b) Dataset robustness: We additionally repeated the analysis on the COCO validation split and on BDD100K using the same model comparison configurations and calibration procedure. The qualitative behavior remained consistent: CCS preserved directional agreement with supervised metrics across architectures and training regimes, and the percentile-based abstention mechanism exhibited similar coverage–reliability trade-offs. Detailed COCO and BDD100K tables are omitted for brevity but mirror the KITTI trends.

I. Platform and Runtime

Experiments were conducted on a server with 8× NVIDIA A100 GPUs and dual-socket AMD EPYC 7742 CPUs. Detector inference runs on GPU and CCS post-processing

TABLE VI: Robustness of CCS alignment across different augmentation seeds for the illustrative model pair (Tab. I).

Aug. Seed	$\rho(\Delta$F1, ΔCCS)	$\rho(\Delta$OC-cost, ΔCCS)	$\rho(\Delta$pPDQ, ΔCCS)
15	0.8170	0.7885	0.6326
33	0.8263	0.7946	0.6447
55	0.8219	0.7832	0.6221
101	0.8107	0.7821	0.6316
150	0.8265	0.7865	0.6316

TABLE VII: CCS post-processing time per image on CPU (1000 KITTI test images; $M{=}9$ augmentations).

Host CPU	2× EPYC 7742 (256 threads)
Ordered pairs $M(M{-}1)$	72
Boxes/aug (N_i)	≤ 5
Per–pair time [ms]	median ≈ 0.05
Per–image time [ms]	0.09–32.53 (median ≈ 3.9)

on CPU. With $M{=}9$ augmentations evaluated sequentially on one A100, inference takes 26–33 ms per augmentation (median $\approx$ 265 ms per image), while CCS adds a median overhead of 3.9 ms per image (Tab. VII). For each ordered pair (i, j), CCS computes IoU, thresholds, and performs Hungarian matching. The per-image complexity is

$$\mathcal{O}\big(M(M{-}1)(N_{\max}^2 + N_{\max}^3)\big),$$

and is small in practice ($N_{\max} \leq 5$), making CCS a minor fraction of the overall pipeline.q

V. Conclusion

We presented CCS, a label-free, deployment-oriented monitoring signal that converts test-time augmentation agreement into a per-image proxy for detector behavior. By aggregating IoU overlaps across mild, non-geometric augmentations, CCS enables continuous monitoring and side-by-side comparison without ground-truth annotations. In controlled studies on Open Images and KITTI, CCS exhibits $>$ 90% directional congruence with F1-score, pPDQ, and OC-cost, alongside strong monotonic alignment and robustness across augmentation seeds and heterogeneous architectures. Qualitative consistency was further confirmed on COCO and BDD100K validation splits. Furthermore, we provided a simplified theoretical analysis that links the expected CCS to detection correctness under an idealized setting. Using a calibrated indifference margin, small discrepancies are treated as ties, emphasizing substantive disagreements and facilitating targeted inspection of unstable cases, supporting the use of CCS as a standalone monitoring signal in operational settings where labels are unavailable.

References

[1] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, "A review of uncertainty quantification in deep learning," *Inf. Fusion*, 2021.

[2] D. Hall, F. Dayoub, J. Skinner, H. Zhang, D. Miller, P. Corke, G. Carneiro, A. Angelova, and N. Sünderhauf, "Probabilistic object detection: Definition and evaluation," in *WACV*, 2020.

[3] M. T. Le, F. Diehl, T. Brunner, and A. Knoll, "Uncertainty estimation for deep neural object detectors in safety-critical applications," in *ITSC*, 2018.

[4] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4," *IJCV*, 2020.

[5] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *ECCV*, 2016.

[9] T.-Y. Lin, M. Maire, S. Belongie *et al.*, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[10] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *CVPR*, 2020.

[11] S. Schmidt, Q. Rao, J. Tatsch, and A. Knoll, "Advanced active learning strategies for object detection," in *IV*, 2020.

[12] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *NeurIPS*, 2017.

[13] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation," *Neurocomputing*, 2019.

[14] Y. Yang, W. Wang, Z. Chen, J. Dai, and L. Zheng, "Bounding box stability against feature dropout reflects detector generalization across environments," in *ICLR*, 2024, spotlight.

[15] H. Yu, J. Deng, W. Li, and L. Duan, "Towards unsupervised model selection for domain adaptive object detection," in *NeurIPS*, 2024.

[16] W. Yu, S. Zhu, T. Yang, C. Chen, and M. Liu, "Consistency-based active learning for object detection," in *CVPRW*, 2022.

[17] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, "Better aggregation in test-time augmentation," in *ICCV*, 2021.

[18] Y. Kim, E. Jang, J. Yang, and S. J. Kim, "Learning loss for test-time augmentation," in *NeurIPS*, 2020.

[19] S. Gasperini, J. Haug, M.-A. N. Mahani, A. Marcos-Ramiro, N. Navab, B. Busam, and F. Tombari, "Certainnet: Sampling-free uncertainty estimation for object detection," *IEEE RA-L*, 2021.

[20] D. Feng, A. Harakeh, S. L. Waslander, and K. Dietmayer, "A review and comparative study on probabilistic object detection in autonomous driving," *IEEE T-ITS*, 2022.

[21] M. R. Nallapareddy, K. Sirohi, P. L. J. Drews, W. Burgard, C.-H. Cheng, and A. Valada, "Evcenternet: Uncertainty estimation for object detection using evidential learning," in *IROS*, 2023.

[22] M. Otani, R. Togashi, Y. Nakashima, E. Rahtu, J. Heikkilä, and S. Satoh, "Optimal correction cost for object detection evaluation," in *CVPR*, 2022.

[23] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *ICRA*, 2018.

[24] I. Krasin, T. Duerig, N. Alldrin *et al.*, "Openimages: A public dataset for large-scale multi-label and multi-class image classification and object detection," in *ICCVW*, 2017.

[25] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, 2013.

[26] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *ICML*, 2017.

[27] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *ICLR*, 2017.

[28] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM CSUR*, 2014.

[29] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations*, 2019.

[30] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," in *NeurIPS (WS)*, 2019.

[31] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, 2019.

[32] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, 2022.