This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Interpretable Model-Agnostic Plausibility Verification for 2D Object Detectors Using Domain-Invariant Concept Bottleneck Models

Mert Keser*[1,2], Gesina Schwalbe*[2], Azarm Nowzad[2], and Alois Knoll[1]

[1]Technical University of Munich, Germany, mert.keser@tum.de, knoll@mytum.de
[2]Continental AG, Germany, first.last@continental-corporation.com

Abstract

Despite the unchallenged performance, deep neural network (DNN) based object detectors (OD) for computer vision have inherent, hard-to-verify limitations like brittleness, opacity, and unknown behavior on corner cases. Therefore, operation-time safety measures like monitors will be inevitable—even mandatory—for use in safety-critical applications like automated driving (AD). This paper presents an approach for plausibilization of OD detections using a small model-agnostic, robust, interpretable, *and* domain-invariant *image classification model. The safety requirements of interpretability and robustness are achieved by using a small concept bottleneck model (CBM), a DNN intercepted by interpretable intermediate outputs. The domain-invariance is necessary for robustness against common domain shifts, and for cheap adaptation to diverse AD settings. While vanilla CBMs are here shown to fail in case of domain shifts like natural perturbations, we substantially improve the CBM via combination with trainable color-invariance filters developed for domain adaptation. Furthermore, the monitor that utilizes CBMs with trainable color-invarince filters is successfully applied in an AD OD setting for detection of hallucinated objects with zero-shot domain adaptation, and to false positive detection with few-shot adaptation, proving this to be a promising approach for error monitoring.*

1. Introduction

Recent advancements in Deep Neural Networks (DNNs) have made DNN-based object detectors increasingly prevalent in AD assistance systems [33]. The accurate detection and classification of objects in the surrounding environment is essential for AD assistance systems, as it facilitates safe navigation of vehicles on the road [15]. However,

*Equal Contribution

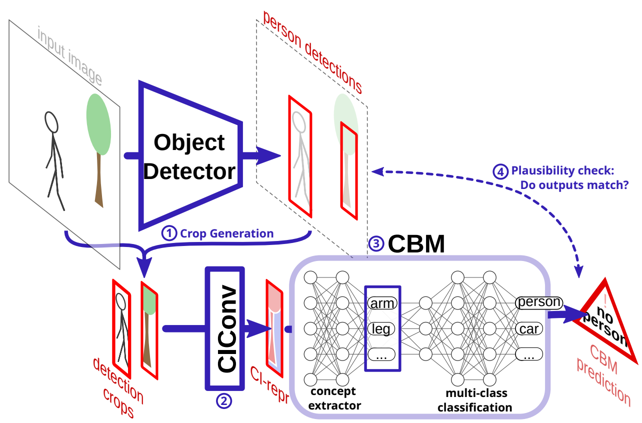


Figure 1. Proposed interpretable, model-agnostic monitoring approach for identification of false positive person detections. It uses an interpretable concept bottleneck model (CBM) as independent classifier on the color-invariant representations (CI-repr) of the object detections (here: for class person). For details see Section 3.

DNNs used for object detection are susceptible to safety-relevant errors in many scenarios [29, 37, 39], in particular when the driving scenes are very different from the training sets [19, 36] or there is an adversarial attack [55]. Examples of safety-relevant errors are misclassification of traffic participants, and false proposal of detections (false positives) like "hallucinated objects". False positives, if untreated, may not only cause uncomfortable driving experience due to sudden jerks, but also hazards like rear crashes. Hence, for real world deployment of DNNs in OD for AD, operation-time system-level measures for error identification and treatment are inevitable, as reflected in upcoming AD safety standards like ISO/TR 4804 [21].

A DNN monitor, also called network observer [21] or runtime monitor [16], provides a score or decision about trustworthiness of a DNN output, based on inputs, outputs, and/or internal processing information of the DNN [44]. Alarms and low trustworthiness scores can be subsequently used to, *e.g.*, discard, correct, or reevaluate the OD prediction, or to propagate the low trustworthiness to later processing stages. While model-specific monitors can be optimized for errors of a specific DNN, they need to be adapted on each update of that model, which may be costly and error-prone. The same holds for monitors based on trustworthiness estimations trained into the DNN outputs, like uncertainty estimates [20]. Hence, it is desirable to complement with model-agnostic monitors that only use the DNN inputs and/or bounding box (bbox) outputs. Model-agnostic approaches share that they check against plausibility constraints, like temporal consistency [47] or semantic relations [12]. However, such constraints must be available, appropriate, and sufficient. A well-known straightforward one is that the OD predictions shall coincide with ones of an independent model [22]. In practice, this faces the challenges of providing a (mostly) *independent* real-time prediction model, which does *not add unnecessary complexity* to the safety assessment, but still is *domain-invariant*, *i.e*. works in diverse AD settings given only cheap adaptations. The idea of this work is to combine methods from explainable artificial intelligence [43] (for independence and safety assessability) and domain adaptation [50] (for domain-invariance) into a monitoring architecture to achieve this.

While standard opaque DNNs excel in computer vision tasks, assessability according to current standards and legislation demands for interpretability [13, 23]. Otherwise, it becomes challenging for developers, users and, in particular, safety assessors to comprehend the reasoning behind predictions [2, 52]. In contrast, human understanding relies on semantic, *i.e*., natural language, concepts and their relations [41, 54]. For instance, from a human perspective, the validity of labeling an object as a car can be verified through the presence of high-level concepts, such as license plates, wheels, and windows. This makes Concept Bottleneck Models (CBMs) [27] a promising candidate to unite assessability and performance: A CBM is a classification DNN that is intercepted by a layer of intermediate outputs, which are trained to correspond to interpretable concepts.

However, as shown in this work, CBMs suffer from insufficient domain-invariance. This restricts applicability to diverse AD settings, and also means that the costly labels for CBM concept training cannot be reused for similar target domains. Hence, we suggest to leverage Color-invariant Convolution (CIConv) filters [30], a method from domain adaptation, to remove irrelevant features from the CBM input that can cause domain-shift issues. As a bonus, biases based on color, *e.g*., skin-color, are ruled out by design.

Our approach for model-agnostic plausibilization is as follows (see Fig. 1): Given an OD class prediction for an image region (*e.g*., a bbox), we check whether the class coincides with the prediction of a small CIConv-CBM-classifier. If they differ, *e.g*., the OD proposes a person, but the CIConv-CBM rejects this, an alarm is raised. Our main contributions are:

(i) We introduce a novel method for model-agnostic, robust, flexible, and human assessible operation-time plausibilization of OD detections.

(ii) We show that the used novel combination of CBMs with CIConvs yields interpretable classifiers that achieve competitive task-performance, and substantially improved robustness of concept representations against domain-shifts, like natural corruptions [35]. As a bonus, concept data sets can, thus, be cheaply reused when training for different target domains.

(iii) The approach is evaluated on different AD OD settings (two datasets, detectors, and object classes), with domain shift from monitor training data. Results prove effectivity in identifying hallucinated objects, and—after few-shot fine-tuning—general OD false positives.

2. Related Work

Model-agnostic False Positive Identification for OD Various works have been proposed to verify the existence of detected objects by object detectors, either based on DNN trustworthiness outputs or plausibilization against given constraints. A common trustworthiness score provided by DNNs are uncertainty estimates. For example, Gaussian YOLOv3 [5] identifies false positives by calculating bbox localization errors. For a thorough overview, the reader is referred to [6]. Unfortunately, uncertainty estimation either relies on modifications of the DNN (*e.g*., specialized architectures [7, 24], output calibration [14]) or is model-agnostic but expensive (*e.g*., ensembling). Autoencoders have also been employed to reveal false positives [46]. The reconstruction error constraint does not rely on a trustworthiness output, however, falls short on desired interpretability. More simple constraints like temporal consistency as in [47] alleviate this, but are less powerful. Many other methods focus on sensor fusion-based verification, such as using the Dempster-Shafer Theory to estimate the existence probability of objects in the environment [1] and verifying detection plausibility with roadside sensors [10]. Khesbak *et al*. [25] utilize a sequential process of checks to ensure that both detections are in agreement before concluding the object's existence. Additionally, Vivekanandan *et al*. [48] apply energy-based optimization methods to analyze the consistency of object detections in multiple sensor streams and identify false positives. In contrast to sensor fusion-based approaches, our method does not rely on additional data sources. To the best of our knowledge, this represents the first model-agnostic method capable of operation-time OD plausibilization that provides human-understandable explanations.

A different approach is pursued by the vast field of out-of-distribution (OOD) detection methods [18]: The idea is that a DNN is more prone to errors on samples or objects scarcely represented in the training data. However, this, by design, is only a proxy target for false positive detection, and, in particular, does not cover in-distribution errors.

Concept Bottleneck Models In 2020, Koh *et al*. proposed Concept Bottleneck Models [27]. Unlike end-to-end DNN training for image classification, CBMs first learn a set of human-interpretable labels, and then use them for prediction. Concept labels can be binary [27] or semantic segmentations [32]. Besides interpretability, intercepting human-interpretable concepts offers the ability to intervene with prediction generation by adjusting the concept outputs [27], *e.g*., for inspection purposes [27].

To train CBMs, it is necessary to have semantically rich annotated data sets [42], such as CUB [49] or Broden(+) [3, 53], that provide labels for all desired concepts. Several studies have proposed solutions to address the costly labeling for CBMs by reducing the number of required samples, such as weakly supervised multi-task learning with concepts [4], concept distillation using an attention-based distillation model [45], and combining supervised and unsupervised concepts through adversarial learning [40]. In contrast to these methods, we aim to reuse results from high-quality, non-scarce data for new target domains. Post-Hoc CBMs [56] is a recent approach that uses concept activation vectors (CAV) [26] or the multi-modal CLIP model [38] to automatically create a concept dataset. However, Post-Hoc CBMs does not fully address the problems of the CBMs since CAV still requires densely annotated concept data and it can only be applied with CLIP image encoder.

It has been demonstrated that CBMs can achieve prediction performance competitive with end-to-end methods [27, 32], and excel in terms of confidence calibration [32] and robustness to background shifts [27]. In addition, our method is capable of extracting concepts that remain effective under various realistic image distortions and weather conditions, a crucial trait for practical AD scenarios.

3. Approach for OD Monitoring

Our approach aims to predict for each detection of an OD whether this is to be considered implausible, *i.e*., "spurious". The OD is treated as black box that produces predictions consisting of bbox coordinates, and an object class. Given the predictions of the OD for an input image, the subsequent plausibility checker consists of the following steps for each detection, illustrated in Fig. 1:

1) **Crop generation:** The bbox is cropped from the original image, and resized to uniform size.
2) **CIConv:** The CIConv creates a single-channel, color-invariant representation (CI-repr) of the crop.
3) **CBM:** The CI-repr is fed through a multi-class classification CBM trained to recognize the same object classes as the OD.
4) **Plausibility check:** The CBM prediction is compared to the originally predicted class, raising an alarm if they differ.

The setup for the CIConv and the CBM are explained in the following.

3.1. Concept Bottleneck Models

For our monitoring application we need to realize a multi-class classification of the detection crops with respect to those object classes that should be checked for errors. Standard DNN classifiers consist of one opaque module that receives inputs and provides the final prediction. Instead, we use a Concept Bottleneck Model architecture as introduced by Koh *et al*. [27] to modularize this into two subsequent DNNs with interpretable intermediate output (see Fig. 1):

1) *concept extractor*: multi-label binary classification of input image into pre-selected, task-related concepts; output: presence scores for each concept.
2) *multi-class classification* of concepts' presence score vector into object classes of interest.

The output layer of the concept-extractor which produces the concept presence scores is also called *concept bottleneck layer*. The overall models can be chosen comparatively small, as was already shown in [27] who used models up to the size of a ResNet-18 [17] as concept extractor, followed by a 3-layer fully connected DNN. This small size ensures small computational overhead of our method, as (1) inference of the model is neglegible compared to a state-of-the-art object detector, (2) inference only needs to be done for each considered bounding box, not the complete image, and (3) processing of predicted bounding boxes can be parallelized. For training, we rely on the joint model training scheme that was shown to perform best in [27]: The CBM ist trained end-to-end with a multi-task loss, *i.e*., a weighted sum of the classification losses for the concept and the final outputs, both using logistic regression.

3.2. Color-Invariant Representations

Geirhos *et al*. [9] discovered in 2018 that ImageNet-trained convolutional neural networks (CNNs) are significantly biased towards recognizing textures instead of shapes, unlike human behavioral evidence. Using fine-tuning with a stylized version of ImageNet they were able to remove the texture bias of the CNNs. As a result, the networks both improved accuracy and robustness against various image distortions. Inspired by this, our approach for robustifying concept representations against domain shifts also relies on (partly) removing texture features like color from the learned representations. Unlike aforementioned study, we automatically remove color and illumination features from all inputs using respective pre-filters.

Following the approach developed by [30], our CBM first layer is replaced by a Color-Invariant Convolution (CIConv) layer, which is a trainable, color-invariant edge detector. The authors of [30] utilize the invariant edge detectors from [11] that were derived from the Kubelka-Munk theory for material reflections [28]. The theory provides an approximate formula to describe the light spectrum reflected from an object into the viewing direction, depending on the original light source and the object material reflectivity. From this, different mappings (the CIConv variants) can be approximated that map an RGB image to a one-channel image representation which is invariant to one or several of: scene geometry (shadows, viewing direction, position of light source), Fresnel reflections, illumination intensity, or color. All CIConv variants come with a parameter σ (the width of the Gaussian used for edge detection in the formula) that determines the trade-off between preserved detail and noise robustness of the resulting representation. Since the CIConvs are differentiable with respect to σ, the optimal value of σ can be trained jointly with the other CBM parameters via backpropagation.

This work uses the CIConv variant W from [30] that focuses on invariance with respect to illumination and achieved best results in preliminary comparative experiments.

The effect of the CIConv-layer on the input data is visualized using the Simple Concept Database (SCDB) [34]. SCDB is a synthetic dataset and consists from the randomly placed large geometric shapes on the black background. These large shapes display random rotations, varying sizes, and a range of colors. The dataset also contais small geometric shapes in a variety of colors, shapes, locations and orientations. Two predefined classes. C1 and C2 are represented by distinct combinations of small geometric shapes within the larger shapes.

CBM and CBM with the CIConv-layer were trained on the SCDB training dataset and assessed on the test dataset. Small geometric shapes form the concepts of the bottleneck layer for both CBM and CBM with the CIConv-layer. The results indicate that the CBM with the CIConv-layer surpasses the performance of the CBM, as the concepts are exclusively based on shape-based concepts. As shown in Fig. 2, the CBM with the CIConv-layer produces robust and informative edge maps of geometric shapes, regardless of background color, shape, location, and orientation.

For implementation details the reader is referred to [30].

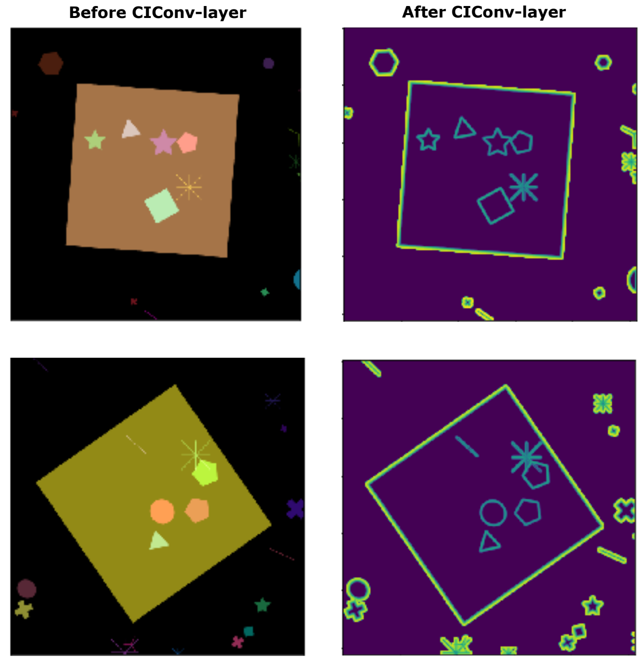


Figure 2. Example images from the SCDB test dataset: the left column displays samples from the test dataset, while the right column displays visualizations after processing through the CIConv-layer.

4. Experiments

In the following subsections we first showcase the benefits of our proposed domain-invariant CBMs with respect to robustness against domain shifts (Section 4.1). Then, this is tested in an end-to-end fashion on two real-world, AD-related data sets and state-of-the-art OD DNN models, and for two kinds of OD errors, namely object hallucinations and general false positives (Section 4.2). Settings common to both studies are detailed below.

CIConv-CBM Settings Throughout all experiments, we adopt as CIConv-CBMs architecture the baseline from [27], who use a ResNet-18 [17] as concept extractor, followed by a 3-layer fully connected DNN. This enables comparability, and yields a very small classifier that proved itself on tasks similar to AD object classification. The CIConv is realized using variant W from [30]. For training both vanilla CBMs and CIConv-CBMs, we adopted the joint training scheme suggested in [27]: All parameters, including those of the CIConv, are trained jointly against the multi-task loss of correctly predicting presence of concepts and the final class of image patches. Loss criteria are weighted as in the original work.

Concept Settings For our evaluations, we concentrated on the object classes `car` and `person` (respectively `pedes-`

Concepts	CBM	CIConv-CBM
Head	**93.6%**	93.1%
Arm	**93.4%**	93.0%
Torso	**91.6%**	91.1%
Leg	**91.4%**	90.4%
Hand	**92.6%**	92.5%
Wheel	**93.2%**	93.2%
Window	**92.7%**	92.0%
Headlight	**93.6%**	92.2%
Door	**93.6%**	93.2%
License Plate	**95.4%**	94.8%

Table 1. Comparison of concept classification accuracy between vanilla concept extractor (CBM) and color-invariant concept extractor architecture with CIConv layer (CIConv-CBM) on Broden test data. Bold numbers highlight the superior concept extractor for each concept.

Classes	CBM	CIConv-CBM
Person	**89.3%**	87.4%
Car	**91.4%**	90.3%

Table 2. Comparison of object class prediction accuracy between vanilla CBM and CBM with color-invariant concept extractor (CIConv-CBM) on Broden test data. Bold numbers indicate the highest object class prediction accuracy for each classes.

trian, ped). We defined the concepts in the bottleneck layer by selecting the five most representative object part concepts for car and person objects in the Broden dataset [3]. For car, the selected concepts were wheel, window, headlight, door, and license plate; for person respectively pedestrian (ped), they were head, arm, torso, leg, and hand. CBM model concept extractors are trained on the Broden dataset.

4.1. Learning Robust Concept Representations

Accuracy In the first phase of our study on robust concept representation learning, we compared the color-invariant concept extractor and the vanilla concept extractor with respect to both concept extractor accuracy (see Tab. 1) and overall CBM classification task accuracy (see Tab. 2) on the Broden test dataset. In terms of concept extraction and task prediction, the CIConv-CBM performs only slightly worse than the vanilla CBM, with all-in-all competitive performance.

Robustness While accuracy is competitive, we then further assessed the impact of introducing CIConvs on robustness with respect to domain shift, in particular realistic image corruption. Image corruption is a widespread problem that results from environmental factors, such as occlusions on the camera lens due to rain, mud, or frost; image blurs due to rapid camera movement; and different forms of noise corruption due to hardware and software issues. Therefore, for the CBM to be usable in autonomous driving, it should be robust and generalizable under these different conditions. For this we evaluated the task prediction accuracy of both CBM variants on test images from the Broden dataset, corrupted with a broad range of realistic image corruptions [35]. Each corruption can be adjusted for different severity from levels 1 to 5, and we selected a severity of 3, which we deemed to be both realistic and challenging. Examples of different corruptions applied to a Broden test image are shown in Fig. 3. The results are shown in Tab. 3.

	Person		**Car**	
Corruptions	**CBM**	**CIConv-CBM**	**CBM**	**CIConv-CBM**
Clean	**89.3%**	87.4%	**91.4%**	90.3%
Brightness	33.88%	**85.69%**	64.92%	**88.77%**
Contrast	33.03%	**85.60%**	52.17%	**87.46%**
Fog	34.62%	**84.74%**	69.38%	**87.30%**
Frost	34.9%	**74.84%**	69.50%	**80.12%**
Gaussian Blur	35.16%	**75.19%**	70.04%	**81.73%**
Compression	35.06%	**83.84%**	69.74%	**87.55%**
Saturate	34.91%	**85.65%**	68.89%	**89.27%**
Shot Noise	34.27%	**65.53%**	42.13%	**74.55%**
Snow	35.27%	**65.37%**	68.33%	**77.90%**

Table 3. Object class prediction accuracy comparison between vanilla CBM and CBM with CIConv layer on Broden test data with applied corruptions (severity=3). Bold numbers highlight the best prediction performance for each class and corruption type.

The object class prediction accuracy results in Tab. 3 demonstrates that CBM with CIConv layer can predict the concepts even in heavy corruptions. Conversely, the prediction accuracy derived from a vanilla CBM exhibit considerably poor performance when compared to the CBM with CIConv layer. Integrating the color-invariant CIConv layer can enhance the CBMs' ability to learn robust and generalizable representations, enabling their applicability to real-world problems, such as autonomous driving.

4.2. Plausibilization with Domain-invariant CBMs

The goal of these experiments was to evaluate the capability of our monitor setup to identify different error types of ODs in realistic setups. To evaluate this we considered the precision, recall, and F1-score of our monitor alarms. Precision here translates to the percentage of the monitor alarms that actually referred to an error of the considered type; and recall means the percentage of the considered errors that were indicated by alarms of our monitor. While low precision means more cases of unnecessary (potentially costly) error recovery actions, low recall means that many potentially safety-relevant errors remain undetected. Hence, high recall is desirable for safety-relevant applications, but also smaller recall values between 0.2 and 0.1 can mean an increase in safety.

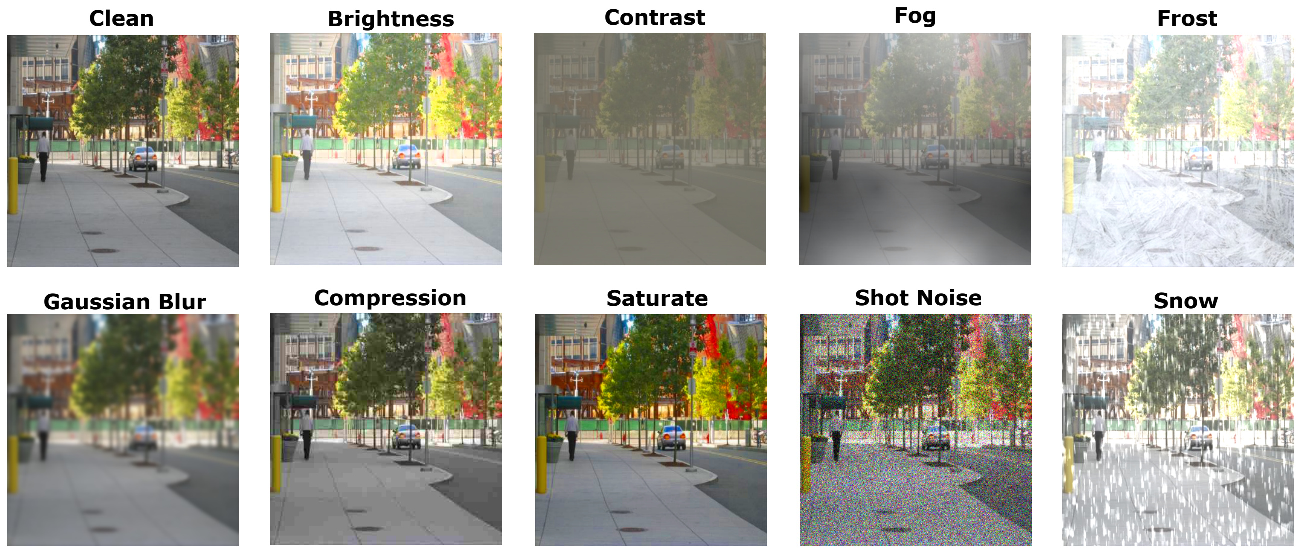


Figure 3. An image from the Broden test dataset, the original clean version, and corrupted ones by different noise model (severity=3).

For the evaluation of the monitor performance, one should note that naturally not all errors of a kind can be retrieved by one type of monitor, and an appropriate recall-precision balance is highly specific to task and system.

Considered Error Types We evaluated retrieval for two kinds of false positive errors:

- **hallucinated objects** which refers to bboxes that are assigned to an object class but have no overlap with a ground truth bbox of that class; and
- **false positives** that include hallucinated objects and localization errors (too little intersection over union of the bbox with a ground truth bbox), but no duplicates, which are automatically removed from our OD outputs by non-maximum-suppression.

Used Models and Datasets We tested the considered error types on two diverse, state-of-the-art object detectors, trained on two different AD-related, real-world object detection datasets:

- YOLOv5 [57] trained on MS COCO [31][1]
- SqueezeDet [51] trained on KITTI [8][2]

We did a random 1:1 split of the KITTI data into training and test data, resulting in each 3731 frames.

[1]Used implementation and weights: https://github.com/ultralytics/yolov5

[2]Used implementation and weights: https://github.com/QiuJueqin/SqueezeDet-PyTorch

4.2.1 Hallucinated Object Identification

Hallucinated object predictions show no overlap with any ground truth bbox of the same class. Hence, the task of the CBM is to decide, whether a bbox predicted to be of object class C does contain any features or part objects associated with C (no alarm), or not (possibly hallucinated object $\rightarrow$ alarm). Since this also was the original training objective of our Broden-trained (CIConv-)CBM, we applied it without any fine-tuning, relying solely on its domain-invariance in order to cope with the domain shift from Broden images to MS COCO crops.

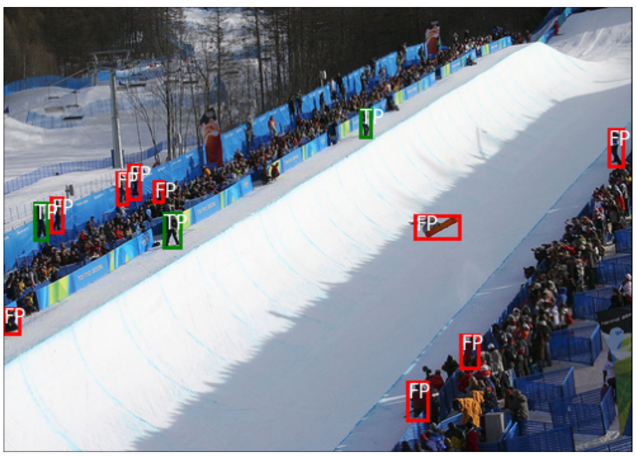
Results are shown in Tab. 4. For the person object class, only 4% of the overall ca. 3k supposedly hallucinated objects were retrieved, for car the more promising number of more than 10% and acceptable precision. To have a closer look at the problem we manually inspected more than 50 examples of supposedly hallucinated objects that were not retrieved by the monitor. This revealed that most "missed" hallucinated objects actually could be reduced to missing, inaccurate, or inconsistent labels (see Fig. 4), even for supposedly improved labels for the COCO dataset[3]. This suggests that the combination of OD DNN with our monitor might be helpful in data label quality checks.

4.2.2 False Positive Identification

In contrast to only considering hallucinated objects as errors, a prediction may also be a false positive error if does have an overlap with a ground truth bbox of the same class.

[3]https://www.sama.com/sama-coco-dataset/

Ground-Truth Annotations



Object Detections

Figure 4. An example from the COCO dataset. The upper portion displays the ground-truth annotations for the person class, and the lower portion shows the person predictions generated by YOLOv5.

Task	IoU	Precision	Recall
Ped	0.5	0.20	0.04
Ped	0.7	0.23	0.04
Car	0.5	0.26	0.12
Car	0.7	0.35	0.12

Table 4. Results for hallucinated object detection on MS COCO

General false positives are defined as cases where the intersection over union (IoU) between the predicted and ground truth box is lower than a threshold. Besides hallucinations, localization errors can cause this and can arise from shifted, too small, and too big bboxes.

We evaluated the monitoring performance for a non-fine-tuned CIConv-CBM monitor on YOLOv5 on the MS COCO dataset (also for different IoU thresholds to define the false positives), and for SqueezeDet on the KITTI

Data, Model	Task	IoU	Precision	Recall
KITTI, SDet	Car	0.7	**0.96**	0.07
KITTI, SDet	(FT) Car	0.7	0.81	**0.56**
KITTI, SDet	Ped	0.5	**0.83**	0.01
KITTI, SDet	(FT) Ped	0.5	0.72	**0.95**

Table 5. Comparison of fine-tuning (FT) and zero-shot false positive monitoring for SqueezeDet (SDet) on KITTI easy. Bold numbers highlight best-performing method for each metric and task.

dataset. In addition, we also compared the KITTI results against those of a CIConv-CBM monitor that was fine-tuned for the identification of false positives of SqueezeDet. For this, we labeled the bboxes predicted by SqueezeDet for our KITTI training data (ca. 3731 images) as false positive or not, and fine-tuned the CBM on this new task dataset of detection crops. Results are shown in Tabs. 5 and 6.

Manual inspection of our results showed that identification of localization errors, in particular shifted and too small bboxes, from only the bbox crops is a harder problem than finding hallucinated objects. The main reason are occlusions and inconsistent ground truth labeling schemes: Many datasets, including MS COCO and KITTI, box only visible parts of an object. Hence, a bbox crop containing only half of an object may be correct, because the remainder of the object is occluded or outside of the image; or it may be a localization error (shifted box, too small box). A special case is objects that are dissected by strong occlusions (*e.g.*, tree in front of car). Here we found that labeling often is inconsistent, sometimes providing two separate ground truth bboxes, and sometimes merging the far-apart object regions by one bbox.

While, by design, our crop-based approach cannot well differentiate occlusion and too short/shifted bboxes, we found that (1) still acceptable error recovery rates could be obtained without finetuning (Tabs. 5 and 6), in particular for high IoU thresholds, and (2) fine-tuning the CBM with few error samples provides good error identification capability (more than 50% of errors identified at less than 30% false alarms), as shown Tab. 5.

Mitigation measures that can be investigated in future work would be improvement of labeling consistency, or adding a small margin to the bbox, to provide the fine-tuned monitor classifier with more context, like parts of the occluding object.

5. Conclusion

We have presented a novel approach to plausibilize object detector predictions during operation using a cheap, interpretable, robust, and model-agnostic monitor. To realize this, we have substantially increased robustness of the used

Data, Model	Task	IoU	Precision	Recall
COCO, YV5	Ped	0.5	0.31	0.04
COCO, YV5	Ped	0.7	0.40	0.04
COCO, YV5	Car	0.5	0.44	0.12
COCO, YV5	Car	0.7	0.63	0.13

Table 6. Comparison of false positive monitoring results for YOLOv5 (YV5) on MS COCO for different IoU thresholds in the false positive definition.

interpretable Concept Bottleneck Models against domain-shifts, by combining them with color-invariant filter methods from the field of domain adaptation. This, for one, allows application to highly diverse AD settings, and, secondly, addresses the CBM problem of data scarcity. The benefits have been demonstrated in our experimental results. Moreover, our end-to-end monitoring tests on state-of-the-art AD settings and OD models suggest that our monitoring is a promising approach to OD error identification at operation time. As next steps we see the validation of our approach in a broader experimental setup, extension to further kinds of OD errors like false negatives, and testing and optimization of the expected real-time capabilities. Also, our qualitative evaluations suggested potential applications of our setup for label quality checks of ground truth data. Further interesting future directions could be the shift from this post-hoc monitoring approach to an ante-hoc interpretable object detectors, or combination with related monitoring techniques such as fusion-based and semantic-constraint-based ones, leveraging the interpretable intermediate outputs of the CBM.

Acknowledgement

The research leading to these results is partly funded by the German Federal Ministry for Economic Affairs and Climate Action within the project "KI Wissen". The authors would like to thank the consortium for the successful cooperation.

References

[1] Michael Aeberhard, Sascha Paul, Nico Kaempchen, and Torsten Bertram. Object existence probability fusion using dempster-shafer theory in a high-level sensor data fusion architecture. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 770–775. IEEE, 2011. 2

[2] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: a comprehensive overview and field guide for future research directions. *arXiv preprint arXiv:2112.11561*, 2021. 2

[3] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proc. 2017 IEEE Conf. Comput. Vision and Pattern Recognition*, pages 3319–3327. IEEE Computer Society, 2017. 3, 5

[4] Catarina Garcia Belém, Vladimir Balayan, Pedro dos Santos Saleiro, and Pedro Gustavo Santos Rodrigues Bizarro. Weakly supervised multi-task learning for concept-based explainability, Jan. 3 2023. US Patent 11,544,471. 3

[5] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 502–511, 2019. 2

[6] Di Feng, Ali Harakeh, Steven L. Waslander, and Klaus Dietmayer. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):9961–9980, Aug. 2022. 2

[7] Jochen Gast and Stefan Roth. Lightweight probabilistic deep networks. In *Proc. 2018 IEEE Conf. Comput. Vision and Pattern Recognition*, pages 3369–3378. IEEE Computer Society, 2018. 2

[8] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 6

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 3

[10] Florian Geissler, Alexander Unnervik, and Michael Paulitsch. A plausibility-based fault detection method for high-level fusion perception systems. *IEEE Open Journal of Intelligent Transportation Systems*, 1:176–186, 2020. 2

[11] J-M Geusebroek, Rein Van den Boomgaard, Arnold W. M. Smeulders, and Hugo Geerts. Color invariance. *IEEE Transactions on Pattern analysis and machine intelligence*, 23(12):1338–1350, 2001. 4

[12] Eleonora Giunchiglia, Mihaela Stoian, Salman Khan, Fabio Cuzzolin, and Thomas Lukasiewicz. Road-r: The autonomous driving dataset with logical requirements. In *IJCLR 2022 Workshops*, June 2022. 2

[13] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, Oct. 2017. 2

[14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proc. 2017 Int. Conf. Machine Learning*, pages 1321–1330. PMLR, July 2017. 2

[15] Abhishek Gupta, Alagan Anpalagan, Ling Guan, and Ahmed Shaharyar Khwaja. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, 10:100057, 2021. 1

[16] Vahid Hashemi, Jan Křetínskỳ, Sabine Rieder, and Jessica Schmidt. Runtime Monitoring for Out-of-Distribution Detection in Object Detection Neural Networks. In *Formal Methods*, Lecture Notes in Computer Science, pages 622–634. Springer International Publishing, 2023. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. 2016 IEEE Conf. Comput. Vision and Pattern Recognition*, pages 770–778, June 2016. 3, 4

[18] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling Out-of-Distribution Detection for Real-World Settings. In *Proceedings of the 39th International Conference on Machine Learning*, pages 8759–8773. PMLR, June 2022. 3

[19] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 1

[20] Maximilian Henne, Adrian Schwaiger, Karsten Roscher, and Gereon Weiss. Benchmarking uncertainty estimation methods for deep learning with safety-related metrics. In *Proc. Workshop Artificial Intelligence Safety*, volume 2560 of *CEUR Workshop Proceedings*, pages 83–90. CEUR-WS.org, 2020. 2

[21] ISO/TC 22 Road vehicles. *ISO/TR 4804:2020: Road Vehicles — Safety and Cybersecurity for Automated Driving Systems — Design, Verification and Validation*. International Organization for Standardization, first edition, Dec. 2020. 1

[22] ISO/TC 22/SC 32. *ISO 26262-3:2018(En): Road Vehicles — Functional Safety — Part 3: Concept Phase*, volume 3 of *ISO 26262:2018(En)*. International Organization for Standardization, second edition, Dec. 2018. 2

[23] ISO/TC 22/SC 32. *ISO 26262-6:2018(En): Road Vehicles — Functional Safety — Part 6: Product Development at the Software Level*, volume 6 of *ISO 26262:2018(En)*. International Organization for Standardization, second edition, Dec. 2018. 2

[24] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30*, pages 5580–5590, 2017. 2

[25] Mohammed S Khesbak. Depth camera and laser sensors plausibility evaluation for small size obstacle detection. In *2021 18th International Multi-Conference on Systems, Signals & Devices (SSD)*, pages 625–631. IEEE, 2021. 2

[26] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 3

[27] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 2, 3, 4

[28] Paul Kubelka. Ein beitrag zur optik der farbanstriche (contribution to the optic of paint). *Zeitschrift fur technische Physik*, 12:593–601, 1931. 4

[29] Michael Truong Le, Frederik Diehl, Thomas Brunner, and Alois Knoll. Uncertainty estimation for deep neural object detectors in safety-critical applications. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3873–3878. IEEE, 2018. 1

[30] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4399–4409, 2021. 2, 4

[31] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. 13th European Conf. Computer Vision - Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer International Publishing, 2014. 6

[32] Max Losch, Mario Fritz, and Bernt Schiele. Interpretability beyond classification output: Semantic bottleneck networks. In *Proc. 3rd ACM Computer Science in Cars Symp. Extended Abstracts*, Oct. 2019. 3

[33] Jialin Lu, Shuming Tang, Jinqiao Wang, Haibing Zhu, and Yunkuan Wang. A review on object detection based on deep convolutional neural networks for autonomous driving. In *2019 Chinese Control And Decision Conference (CCDC)*, pages 5301–5308. IEEE, 2019. 1

[34] Adriano Lucieri, Muhammad Naseer Bajwa, Andreas Dengel, and Sheraz Ahmed. Explaining ai-based decision support systems using concept localization maps. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27*, pages 185–193. Springer, 2020. 4

[35] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 2, 5

[36] Julia Nitsch, Masha Itkina, Ransalu Senanayake, Juan Nieto, Max Schmidt, Roland Siegwart, Mykel J Kochenderfer, and Cesar Cadena. Out-of-distribution detection for automotive perception. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2938–2943. IEEE, 2021. 1

[37] Thomas Ponn, Thomas Kröger, and Frank Diermeyer. Identification and explanation of challenging conditions for camera-based object detection of automated vehicles. *Sensors*, 20(13):3699, 2020. 1

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[39] Manikandasriram Srinivasan Ramanagopal, Cyrus Anderson, Ram Vasudevan, and Matthew Johnson-Roberson. Failing to learn: Autonomously identifying perception failures for self-driving cars. *IEEE Robotics and Automation Letters*, 3(4):3860–3867, 2018. 1

[40] Yoshihide Sawada and Keigo Nakamura. Concept bottleneck model with additional unsupervised concepts. *IEEE Access*, 10:41758–41765, 2022. 3

[41] Gesina Schwalbe. Concept embedding analysis: A review. *arXiv preprint arXiv:2203.13909*, 2022. 2

[42] Gesina Schwalbe. Concept embedding analysis: A review. *arXiv:2203.13909 [cs, stat]*, Mar. 2022. 3

[43] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, Jan. 2023. 2

[44] Gesina Schwalbe, Bernhard Knie, Timo Sämann, Timo Dobberphul, Lydia Gauerhof, Shervin Raafatnia, and Vittorio Rocco. Structuring the safety argumentation for deep neural network based perception in automotive applications. In *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops*, pages 383–394. Springer International Publishing, 2020. 1

[45] João Bento Sousa, Ricardo Moreira, Vladimir Balayan, Pedro Saleiro, and Pedro Bizarro. Conceptdistil: Model-agnostic distillation of concept explanations. *arXiv preprint arXiv:2205.03601*, 2022. 3

[46] Noelia Vallez, Alberto Velasco-Mata, and Oscar Deniz. Deep autoencoder for false positive reduction in handgun detection. *Neural Computing and Applications*, 33(11):5885–5895, 2021. 2

[47] Serin Varghese, Yasin Bayzidi, Andreas Bar, Nikhil Kapoor, Sounak Lahiri, Jan David Schneider, Nico M. Schmidt, Peter Schlicht, Fabian Huger, and Tim Fingscheidt. Unsupervised temporal consistency metric for video segmentation in highly-automated driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 336–337, 2020. 2

[48] Abhishek Vivekanandan, Niels Maier, and J Marius Zöllner. Plausibility verification for 3d object detectors using energy-based optimization. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 602–616. Springer, 2023. 2

[49] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. Caltech-UCSD birds 200. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 3

[50] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, Oct. 2018. 2

[51] Bichen Wu, Forrest Iandola, Peter H Jin, and Kurt Keutzer. Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 129–137, 2017. 6

[52] Tianfu Wu and Xi Song. Towards interpretable object detection by unfolding latent structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6033–6043, 2019. 2

[53] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proc. 15th European Conf. Comput. Vision, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 432–448. Springer International Publishing, 2018. 3

[54] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020. 2

[55] Mingjun Yin, Shasha Li, Chengyu Song, M Salman Asif, Amit K Roy-Chowdhury, and Srikanth V Krishnamurthy. Adc: Adversarial attacks against object detection that evade context consistency checks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3278–3287, 2022. 1

[56] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[57] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios, Aug. 2021. 6