Two Video Data Sets for Tracking and Retrieval of Out of Distribution Objects

Kira Maag[1], Robin Chan[2], Svenja Uhlemeyer[3], Kamil Kowol[3] and Hanno Gottschalk[3]

[1] Ruhr University Bochum, Germany kira.maag@rub.de
[2] Bielefeld University, Germany rchan@techfak.uni-bielefeld.de
[3] University of Wuppertal, IZMD, Germany
{suhlemeyer,kowol,hgottsch}@uni-wuppertal.de

Abstract. In this work we present two video test data sets for the novel computer vision (CV) task of out of distribution tracking (OOD tracking). Here, OOD objects are understood as objects with a semantic class outside the semantic space of an underlying image segmentation algorithm, or an instance within the semantic space which however looks decisively different from the instances contained in the training data. OOD objects occurring on video sequences should be detected on single frames as early as possible and tracked over their time of appearance as long as possible. During the time of appearance, they should be segmented as precisely as possible. We present the SOS data set containing 20 video sequences of street scenes and more than 1000 labeled frames with up to two OOD objects. We furthermore publish the synthetic CARLA-WildLife data set that consists of 26 video sequences containing up to four OOD objects on a single frame. We propose metrics to measure the success of OOD tracking and develop a baseline algorithm that efficiently tracks the OOD objects. As an application that benefits from OOD tracking, we retrieve OOD sequences from unlabeled videos of street scenes containing OOD objects.

Keywords: Computer vision, video, data sets, out of distribution.

1 Introduction

Semantic segmentation decomposes the pixels of an image into segments that adhere to a pre-defined set of semantic classes. In recent years, using fully convolutional deep neural networks [48] and training on publicly available data sets [14,18,28,68,29,82], this technology has undergone a remarkable learning curve. Recent networks interpret street scenes with a high degree of precision [17,78].

When semantic segmentation is used in open world scenarios, like in automated driving as area of application, objects could be present on images, which adhere to none of the semantic classes the network has been trained on and therefore force an error. Such objects from outside the network's semantic space form a specific class of out of distribution (OOD) objects. Naturally, it is desirable that the segmentation algorithm identifies such objects and abstains a decision on the semantic class for those pixels that are covered by the OOD object. At the same time, this additional requirement should not much deteriorate the performance on the primary segmentation task, if no OOD object is present. In other cases, an OOD object might be from a known class, however with an appearance that is very different from the objects of the same class in the training data, so that a stable prediction for this object is unrealistic. Also in this case, an indication as OOD object is preferable over the likely event of a misclassification. The computer vision (CV) task to mark the pixels of both kinds of objects can be subsumed under the notion of OOD segmentation. See [22,10,11,16,15,31,32,55,54] for recent contributions to this emerging field.

In many applications, images do not come as single frames, but are embedded in video sequences. If present, OOD objects occur persistently on subsequent frames. Tracking of OOD objects therefore is the logical next step past OOD segmentation. This ideally means identifying OOD objects in each frame on which they are present and give them a persistent identifier from the frame of first occurrence to the frame in which the OOD object leaves the image.

In this article we introduce the novel task of OOD tracking as a hybrid CV task inheriting from the established fields of OOD detection, OOD segmentation and object tracking. CV tasks often are dependent on suitable data sets, and OOD tracking is no exception in this regard. As our main contribution, we present two new labeled data sets of video sequences that will support the research effort in this field. The Street Obstacle Sequences (SOS) data set is a real world data set that contains more than 1,000 single frames in 20 video sequences containing one or two labeled OOD objects on streets along with further meta information, like distance or object ID. The SOS data set thus allows to evaluate the success of OOD tracking quantitatively for different kinds of OOD objects. As a second data set, we present CARLA-WildLife (CWL), a synthetic data set that consists of 26 fully annotated frames from the CARLA driving simulator in which a number of OOD objects from the Unreal Engine [27] collection of free 3D assets are introduced. Each frame in these video sequences contains in between 1 and 4 OOD instances. The meta data is consistent with SOS. In addition, the labeling policy is largely consistent with the single frame based road obstacle track in the SegmentMeIfYouCan benchmark [15]. Thereby, both data sets will also support standard OOD segmentation benchmarks. As a second contribution, we propose numerous metrics that can systematically measure the success of an OOD tracking algorithm. As a third contribution, we provide a first baseline that combines single frame OOD segmentation with tracking of segments. Using a single frame Nvidia DeepLabV3+ as a single frame segmentation network, we employ entropy maximization training for OOD segmentation with meta-classification to reduce the number of false positive OOD objects, following [16]. We then track the obtained OOD masks over the video sequences using an adjusted version of the light-weight tracking algorithm based on semantic masks introduced in [58,59,57]. We hope that this simple baseline will motivate researchers to develop their own OOD tracking algorithms and compare performance against our baseline.

It remains to show that OOD tracking is useful. Here we present an example from the context of automated driving and apply OOD tracking on the unsupervised retrieval of OOD objects. To this purpose, we combine our OOD tracking baseline with feature extractor based on DenseNet [42]. For each detected OOD object, we obtain a time series of feature vectors on which we employ a low dimensional embedding via the t-SNE algorithm [60]. Here the time series viewpoint makes it easy to clean the data and avoid false positives, e.g. by setting a filter to the minimum length. Clustering of similar objects, either on the basis of frames or on time series meta-clusters then enables the retrieval of previously unseen objects [69,77]. We apply this on the SOS and the CWL data sets as well as on self-recorded unlabeled data that contains OOD road obstacles. This provides a first method that enables the unsupervised detection of potentially critical situations or corner cases related to OOD objects from video data. The source code is publicly available at https://github.com/kmaag/OOD-Tracking and the datasets at https://zenodo.org/communities/buw-ood-tracking/.

This paper is organized as follows: section 2 relates our work with existing OOD data sets as well as approaches in OOD segmentation, object tracking and object retrieval. The following section 3 introduces our data sets for OOD tracking in street scenes and details on our labeling policy. In section 4, we introduce a set of metrics to measure the success of OOD segmentation, tracking and clustering, respectively. The experiments are presented in section 5 consisting of the method description, i.e., details of our OOD segmentation backbone, the tracking algorithm for OOD objects as well as OOD retrieval, and numerical results for the SOS as well as the CWL data set. Our findings are summarized in section 6, where we also shortly comment on future research directions.

2 Related Work

OOD Data Sets OOD detection in the field of CV is commonly tested by separating entire images that originate from different data sources. This includes e.g. separating MNIST [49] from FashionMNIST [80], NotMNIST [12], or Omniglot [46], and, as more complex task, separating CIFAR-10 [45] from SVHN [30] or LSUN [83]. Other data sets specifically designed to OOD detection in semantic segmentation are for instance Fishyscapes [10] and CAOS [38]. These two data sets either rely on synthetic data or generate OOD examples by excluding certain classes during model training. To overcome the latter limitations, data sets such as LostAndFound [71], RoadAnomaly [55], and also RoadObstacle21 [15] include images containing real OOD objects appearing in real world scenes. To this end, the established labeling policy of the semantic segmentation data set Cityscapes [18] serves as basis to decide whether an object is considered as OOD or not. However, all the outlined OOD data sets are based on single frames only. Although CAOS [38], LostAndFound [71], and RoadObstacle21 [15] include several images in the same scenes, they do not provide video sequences with (annotated) consecutive frames. In particular, mainly due to the labeling effort, none of the real world data sets provides a sufficient density of consecutive frames such that tracking of OOD objects could be applied and evaluated properly. One such but synthetic data set is StreetHazards [38]. This latter data set, however, mostly contains street scenes with OOD objects appearing in safety-irrelevant locations such as the background of the scene or in non-driveable areas.

In this work, we provide two novel video data sets with OOD objects on the road as region of interest. Therefore, our data sets can be understood to tackle the safety-relevant problem of obstacle segmentation [15]. While one of these two data sets consists of real-world images only, the other consists of synthetic ones. Both data sets include multiple sequences with pixel level annotations of consecutive frames, which for the first time enable tracking of OOD objects.

OOD Segmentation OOD detection on image data was first tackled in the context of image classification. Methods such as [39,50,51,37,63] have proven to successfully identify entire OOD images by lowering model confidence scores. These methods can be easily extended to semantic segmentation by treating each pixel individually, forming common baselines for OOD detection in semantic segmentation [1,9], i.e., OOD segmentation. In particular, many of these OOD detection approaches are intuitively based on quantifying prediction uncertainty. This can also be accomplished e.g. via Monte-Carlo dropout [26] or an ensemble of neural networks [47,34], which has been extended to semantic segmentation in [3,44,65]. Another popular approach is training for OOD detection [21,40,63], which includes several current state-of-the-art works on OOD segmentation such as [8,16,7,22,32]. This type of approach relies on incorporating some kind of auxiliary training data, not necessarily real-world data, but disjoint from the original training data. In this regard, the most promising methods are based on OOD training samples generated by generative models as extensively examined in [19,66,55,79,54].

All existing methods are developed to operate on single frames. In this present work, we aim at investigating how such OOD segmentation methods could be extended to operate on video sequences with OOD objects appearing in multiple consecutive frames.

Object Tracking In applications such as automated driving, tracking multiple objects in image sequences is an important computer vision task [64]. In instance segmentation, the detection, segmentation and tracking tasks are often performed simultaneously in terms of extending the Mask R-CNN network by an additional branch [6,81] or by building a variational autoencoder architecture on top [52]. In contrast, the tracking-by-detection methods first perform segmentation and then tracking using for example a temporal aggregation network [43] or the MOTSNet [72]. In addition, a more light-weight approach is presented in [13] based on the optical flow and the Hungarian algorithm. The tracking method introduced in [59] serves as a post-processing step, i.e., is independent of the instance segmentation network, and is light-weight solely based on the overlap of instances in consecutive frames. A modified version of this algorithm is used for semantic segmentation in [58].

Despite all the outlined works on object tracking, none of them were developed for OOD objects. In this present work, we therefore extend the post-processing method for tracking entire segments in semantic segmentation, that has originally been proposed in [58], to the unprecedented task of tracking OOD objects in image sequences.

Object Retrieval Retrieval methods in general tackle the task of seeking related samples from a large database corresponding to a given query. Early works in this context aim to retrieve images that match best a query text or vice versa [41,2,33,62]. Another sub task deals with content-based image retrieval, which can be sub-categorized into instance- and category level retrieval. This is, given a query image depicting an object or scene, retrieving images representing the same object/scene or objects/scenes of the same category, respectively. To this end, these images must satisfy some similarity criteria based on some abstract description. In a first approach called QBIC [25], images are retrieved based on (global) low level features such as color, texture or shape. More advanced approaches utilize local level features [4,56], still they cannot fully address the problem of semantic gap [75], which describes the disparity between different representation systems [36]. Recent methods such as [67,61] apply machine/deep learning to learn visual features directly from the images instead of using hand-crafted features.

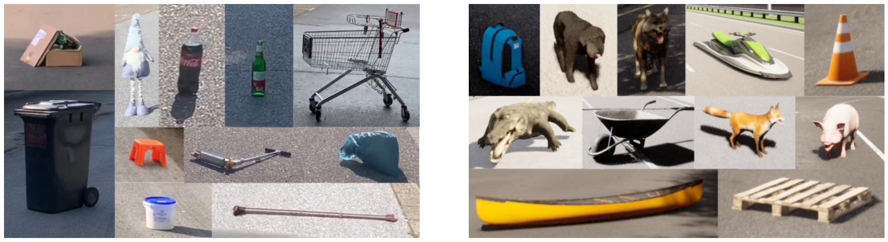
In this work, we do not directly retrieve images for some given query image, but instead we cluster all objects/images that are contained in our database based on their visual similarity, as it has been proposed in [69]. This particularly includes OOD objects. We extend this described single frame based approach to video sequences, i.e., we enhance the effectiveness by incorporating tracking information over multiple frames.

3 Data Sets

As already discussed in section 2, in general there is a shortage of data sets that are dedicated to OOD detection in semantic segmentation. In particular, at the time of writing, there does not exist any OOD segmentation data set containing annotated video sequences. We therefore introduce the *Street Obstacle Sequences (SOS)*, *CARLA-WildLife (CWL)* and *Wuppertal Obstacle Sequences (WOS)* data sets. Example images and more details can be found in appendix A.

3.1 Street Obstacle Sequences

The SOS data set contains 20 real-world video sequences in total. The given scenes are shown from a perspective of a vehicle that is approaching objects placed on the street, starting from a distance of 20 meters to the street obstacle. The outlined street obstacles are chosen such that they could cause hazardous street scenarios. Moreover, each object corresponds to a class that is semantically OOD according to the Cityscapes labeling policy [18]. In SOS, there are 13



(a) Street Obstacle Sequences (b) CARLA-WildLife

Fig. 1: Some exemplary OOD objects from our (a) SOS and (b) CWL data sets.

different object types, which include e.g. bags, umbrellas, balls, toys, or scooters, cf. also fig. 1(a). They represent potential causes of hazardous street scenarios, making their detection and localization particularly crucial in terms of safety.

Each sequence in SOS was recorded at a rate of 25 frames per second, of which every eighth frame is labeled. This yields a total number of 1,129 pixel-accurately labeled frames. As region of interest, we restrict the segmentation to the drivable area, i.e., the street. Consequently, SOS contains two classes, either

1) *street obstacle / OOD* , or 2) *street / not OOD* .

Note that image regions outside the drivable area are labeled as *void* and are ignored during evaluation.

Given the unique density of consecutive annotated frames, SOS allows for proper evaluation of tracking OOD objects besides their detection and pixel level localization. In this way, SOS facilitates the approach to the novel and practically relevant CV task of combining object tracking and OOD segmentation.

For a more in-depth evaluation, we further provide meta data to each obstacle in the SOS data set. This includes information such as the size of an object and their distance to the camera. In this regard, the size is approximated by the number of annotated pixels and the distance by markings on the street.

3.2 CARLA-WildLife

Since the generation of the SOS data set is time consuming and the selection of diverse real-world OOD objects is limited in practice, we additionally introduce a synthetic data set for OOD detection offering a large variety of OOD object types. The main advantage of synthetic data is that they can be produced inexpensively with accurate pixel-wise labels of full scenes, besides being able to manipulate the scenes as desired.

By adding freely available assets from Unreal Engine 4 [27] to the driving simulation software CARLA [23], we generate sequences in the same fashion as the SOS data set that we provide in the additional CWL data set. It contains 26 synthetic video sequences recorded at a rate of 10 frames per second with 18 different object types placed on the streets of CARLA. The objects include e.g.

dogs, balls, canoes, pylons, or bags, cf. also fig. 1(b). Again, the objects were chosen based on whether they could cause hazardous street scenarios. Since these objects are not included in the standard set of semantic labels provided by CARLA, each object type is added as extra class retroactively. In addition to the semantic segmentation based on the Cityscapes labeling policy (and including the OOD class), CWL further provides instance segmentation, i.e., individual OOD objects of the same class can be distinguished within each frame, and tracking information, i.e., the same object instance can be identified over the course of video frames. Moreover, we provide pixel-wise distance information for each frame of entire sequences as well as aggregated depth information per OOD object depicting the shortest distance to the ego-vehicle.

3.3 Wuppertal Obstacle Sequences

While the SOS data set considers video sequences where the camera moves towards the static OOD objects located on the street, we provide additional moving OOD objects in the *WOS* data set. It contains 44 real-world video sequences recorded from the viewpoint of a moving vehicle. The moving objects are mostly dogs, rolling or bouncing balls, skateboards or bags and were captured with either a static or a moving camera. This data set comes without labels and is used for test purposes for our OOD tracking and retrieval application.

4 Performance Metrics

In this section, we describe the performance metrics for the task of OOD tracking, i.e., OOD segmentation and object tracking, as well as clustering.

4.1 OOD Segmentation

Hereafter, we assume that the OOD segmentation model provides pixel-wise OOD scores s for a pixel discrimination between *OOD* and *not OOD*, see also section 3. As proposed in [15], the separability of these pixel-wise scores is evaluated using the area under the precision recall curve (AuPRC) where precision and recall values are varied over some score thresholds $\tau \in \mathbb{R}$ applied to s. Furthermore, we consider the false positive rate at 95% true positive rate (FPR_{95}) as safety critical evaluation metric. This metric indicates how many false positive errors have to be made to achieve the desired rate of true positive predictions.

As already implied, the final OOD segmentation is obtained by thresholding on s. In practice, it is crucial to detect and localize each single OOD object. For this reason, we evaluate the OOD segmentation quality on segment level. To this end, we consider a connected component of pixels sharing the same class label in a segmentation mask as segment. From a practitioner's point of view, it is often sufficient to only recognize a fraction of OOD objects to detect and localize them. As quality measure to decide whether one segment is considered as detected, we stick to an adjusted version of the segment-wise intersection over union ($sIoU$) as introduced in [73]. Then, given some detection threshold $\kappa \in [0,1)$, the number of true positive (TP), false negative (FN) and false positive (FP) segments can be computed. These quantities are summarized by the $F_1 = 2TP/(2TP + FN + FP)$ score, which represents a metric for the segmentation quality (for some fixed score threshold κ). As the numbers of TP, FN and FP depend on the detection threshold κ, we additionally average the F_1 score over different κ. This yields $\bar{F}_1$ as our main evaluation metric on segment level as it is less affected by the detection threshold.

For a more detailed description of the presented performance metrics for OOD segmentation, we refer to [15].

4.2 Tracking

To evaluate OOD object tracking, we use object tracking metrics such as multiple object tracking accuracy ($MOTA$) and precision ($MOTP$) as performance measures [5]. $MOTA$ is based on three error ratios: the ratio of false positives, false negatives and mismatches ($\overline{mme}$) over the total number of ground truth objects in all frames. A mismatch error is defined as the ID change of two predicted objects that are matched with the same ground truth object. $MOTP$ is the averaged distance between geometric centers of matched pairs of ground truth and predicted objects.

For the tracking measures introduced in [64], all ground truth objects of an image sequence are identified by different IDs and denoted by GT. These are divided into three cases: mostly tracked (MT) if it is tracked for at least 80% of frames (whether the object was detected or not), mostly lost (ML) if it is tracked for less than 20%, else partially tracked (PT). These common multiple object tracking metrics are created for the object detection task using bounding boxes and also applicable to instance segmentation. Thus, we can apply these measures to our detected OOD objects without any modification.

Moreover, we consider the tracking length metric l_t which counts the number of all frames where a ground truth object is tracked divided by the total number of frames where this ground truth object occurs. In comparison to the presented metrics which require ground truth information in each frame, the tracking length additionally uses non-annotated frames if present. Note that we find this case within the SOS data set where about every eighth frame is labeled. To this end, we consider frames $t, \dots, t+i$, $i > 1$, with available labels for frames t and $t+i$. If the ground truth object in frame t has a match and the corresponding tracking ID of the predicted object occurs in consecutive frames $t+1, \dots, t+i-1$, we increment the tracking length.

4.3 Clustering

The evaluation of OOD object clusters $C_i \in \{C_1, \dots, C_n\}$, which contain the two-dimensional representatives of the segments k of OOD object predictions, depends on the differentiation level of these objects. We consider an instance level and a semantic level based on object classes. Let $\mathcal{Y} = \{1, \dots, q\}$ and

$\mathcal{Y}^{\text{ID}} = \{1, \dots, p\}$ denote the set of semantic class and instance IDs, respectively. For some given OOD segment k, y_k and y_k^{ID} correspond to the ground truth class and instance ID with which k has the highest overlap. On instance level, we aspire that OOD objects which belong to the same instance in an image sequence are contained in the same cluster. This is, we compute the relative amount of OOD objects per instance in the same cluster,

$$CS_{\text{inst}} = \frac{1}{p} \sum_{i=1}^{p} \frac{\max\limits_{C \in \{C_1,\dots,C_n\}} |\{k \in C \mid y_k^{\text{ID}} = i\}|}{\sum\limits_{C \in \{C_1,\dots,C_n\}} |\{k \in C \mid y_k^{\text{ID}} = i\}|} \in [0,1] \ , \tag{1}$$

averaged over all instances. On a semantic level, we pursue two objectives. The first concerns the semantic class impurity of the clusters,

$$CS_{\text{imp}} = \frac{1}{n} \sum_{i=1}^{n} |\{y_k | k \in C_i\}| \in [1, q] \ , \tag{2}$$

averaged over all clusters $C_i \in \{C_1, \dots, C_n\}$. Secondly, we aspire a low fragmentation of classes into different clusters

$$CS_{\text{frag}} = \frac{1}{q} \sum_{i=1}^{q} |\{C \in \{C_1, \dots, C_n\} | \exists k \in C : y_k = i\}| \ , \tag{3}$$

i.e., ideally, each class constitutes exactly one cluster. Here, we average over the semantic classes in $\mathcal{Y}$.

5 Experiments

In this section, we first introduce the methods which we use for OOD segmentation, tracking as well as retrieval and second, we show the numerical and qualitative results on our two main data sets, SOS and CWL. Qualitative results on OOD object retrieval from the WOS data set are given in the appendix.

5.1 Method

Our method consists of the CV tasks OOD segmentation and object tracking. For OOD segmentation, we consider the predicted region of interest and the entropy heatmap obtained by a semantic segmentation network. Via entropy thresholding, the OOD objects are created and the prediction quality is assessed by meta classification in order to discard false positive OOD predictions. In the next step, the OOD objects are tracked in an image sequence to generate tracking IDs. Furthermore, we study the retrieval of detected OOD objects using tracking information. An overview of our method is shown in fig. 2.

OOD Object Segmentation For the segmentation of OOD objects, we use the publicly available segmentation method that has been introduced in [16]. In the latter work, a DeepLabV3+ model [84], initially trained on Cityscapes [18],

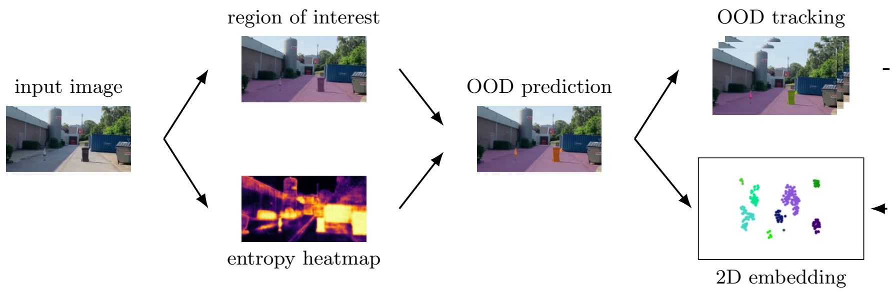


Fig. 2: Overview of our method. The input image is fed into a semantic segmentation network to extract the region of interest (here road) and the entropy heatmap. The resulting OOD prediction is used to produce the tracking IDs and a 2D embedding.

has been extended to OOD segmentation by including auxiliary OOD samples extracted from the COCO data set [53]. To this end, the model has been trained for high softmax entropy responses on the induced known unknowns (provided by COCO), which showed generalization capabilities with respect to truly unknown objects available in data sets such as LostAndFound [71] and RoadObstacle21 [15]. This outlined method is applied to single frames and utilizes the pixel-wise softmax entropy as OOD score.

Further, we apply meta classification [73,74] to OOD object predictions for the purpose of reducing false positive OOD indications. These false positives are identified by means of hand-crafted metrics, which are in turn based on dispersion measures like entropy as well as geometry and location information, see also [16]. These hand-crafted metrics form a structured data set where the rows correspond to predicted segments and the columns to features. Given this meta data set, we employ logistic regression with L^1-penalty on the weights (LASSO [76]) as post-processing (meta) model to remove false positive OOD object predictions, without requiring ground truth information at run time.

For more details on the construction of the structured data set, we refer the reader to [73,74]. An illustration of the single steps of the OOD object segmentation method can be found in fig. 3.

OOD Object Tracking In this section, we present the light-weight tracking approach that we use to track predicted OOD objects. This method has originally been introduced for semantic segmentation in [58] and does not require any training as it is an heuristic solely based on the overlap of OOD objects in consecutive frames. We assume that an OOD object segmentation is available for each frame x, as e.g. described in section 5.1. The idea of employing this tracking method is to match segments based on their overlap (measured by the segment-wise intersection over union, shorthand IoU) and proximity of their geometric centers in consecutive frames.

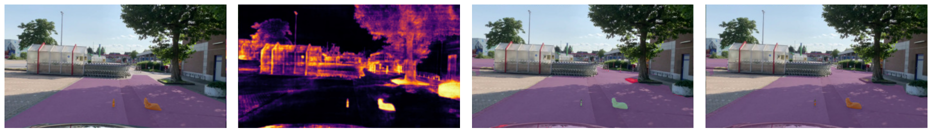


Fig. 3: Segmentation of OOD objects (orange in ground truth) on the street via entropy thresholding & prediction quality rating via meta classification (green corresponds to a high confidence of being a correct OOD object prediction, red to a low one), resulting in final prediction mask.

We apply the tracking approach sequentially to each frame $x \in \{x_t\}_{t=1}^{T}$ of an image sequence of length T. In more detail, the segments in the first frame, i.e., $t = 1$, are assigned with random IDs. Then, for each of the remaining frames $t, t > 1$, the segments are matched with the segment IDs of its respective previous frame $t - 1$. To this end, we use a tracking procedure consisting of five steps, which we will briefly describe in what follows. For a detailed description, we refer the reader to [58]. In step 1, OOD segments that are predicted in the same frame are aggregated by means of their distance. In steps 2 and 3, segments are matched if their geometric centers are close together or if their overlap is sufficiently large in consecutive frames, respectively. In step 4, linear regression is used to account for "flashing" segments (over a series of consecutive frames) or temporarily occluded as well as non-detected ones, i.e., false negatives. As final step 5, segments are assigned new IDs in case they have not received any in the steps 1-4 of the matching process.

OOD Object Retrieval On top of the segmentation and tracking of OOD objects, we perform a method similar to content-based image retrieval in order to form clusters of the OOD objects that constitute novel semantic concepts. To this end we adapt an existing approach [69,77] to video sequences by incorporating the tracking information which we obtain e.g. as described in section 5.1. This is, we require the tracking information to be available for each frame x and apply OOD object retrieval as a post-processing step which does not depend on the underlying semantic segmentation network nor on the OOD segmentation method but on given OOD segmentation masks.

For each frame x and OOD segment $k \in \hat{K}(x)$, let $\hat{y}_k^{ID}$ denote the predicted tracking ID. To diminish the number of the false positives, we only cluster predicted segments that are tracked over multiple frames of an image sequence $\{x_t\}_{t=1}^{T}$, based on some length parameter $\ell \in \mathbb{N}$ Further, each frame x is tailored to boxes around the remaining OOD segments k, which are vertically bounded by the pixel locations $\min_{(z_v,z_h)\in k} z_v$ and $\max_{(z_v,z_h)\in k} z_v$, horizontally by $\min_{(z_v,z_h)\in k} z_h$ and $\max_{(z_v,z_h)\in k} z_h$. Image clustering usually takes place in a lower dimensional latent space due to the curse of dimensionality. To this end, the image patches are fed into an image classification ResNet152 [35] (without

Table 1: OOD object segmentation, tracking and clustering results for the SOS and the CWL data set.

data set	AuPRC ↑	FPR_{95} ↓	$\bar{F}_1$ ↑	$MOTA$ ↑	$\overline{mme}$ ↓	$MOTP$ ↓	GT	MT	PT	ML	l_t ↑
SOS	85.56	1.26	35.84	−0.0826	0.0632	12.3041	26	9	14	3	0.5510
CWL	79.54	1.38	45.46	0.4043	0.0282	16.4965	62	24	30	8	0.5389

	without tracking ($\ell = 0$)			with tracking ($\ell = 10$)		
data set	CS_{inst} ↑	CS_{imp} ↓	CS_{frag} ↓	CS_{inst} ↑	CS_{imp} ↓	CS_{frag} ↓
SOS	0.8652	2.5217	2.8182	0.8955	1.7917	1.9091
CWL	0.8637	2.8181	2.2500	0.8977	2.1739	1.8000

its final classification layer) trained on ImageNet [20], which produces feature vectors of equal size regardless of the input dimension. These features are projected into a low-dimensional space by successively applying two dimensionality reduction techniques, namely principal component analysis (PCA [70]) and t-distributed stochastic neighbor embedding (t-SNE [60]). As final step, the retrieved OOD object predictions are clustered in the low-dimensional space, e.g., via the DBSCAN clustering algorithm [24].

5.2 Numerical Results

In this section, we present the numerical results on the novel task of OOD tracking. To this end, we apply simple baseline methods introduced in section 5.1 on two labeled data sets of video sequences (SOS and CWL) and motivate the usefulness of OOD tracking using an unsupervised retrieval of OOD objects in the context of automated driving.

OOD Segmentation For OOD segmentation, we apply the method described in section 5.1, which provides pixel-wise softmax entropy heatmaps as OOD scores (see fig. 3 (center left)). The pixel-wise evaluation results for the SOS and the CWL data sets are given in table 1 considering AuPRC and FPR_{95} as metrics (section 4.1). We achieve AuPRC scores of 85.56% and 79.54% as well as FPR_{95} scores of 1.26% and 1.38% on SOS and CWL, respectively.

To obtain the OOD segmentation given some input image, thresholding is applied to the softmax entropy values. We choose the threshold τ by means of hyperparameter optimization, yielding $\tau = 0.72$ for SOS and $\tau = 0.81$ for CWL.

As next step, meta classification is used as post-processing to reduce the number of false positive OOD segments. We train the model on one data set and evaluate on the other one, e.g. for experiments on SOS the meta classification model is trained on CWL. The corresponding $\bar{F}_1$ scores on segment level are shown in table 1. The higher $\bar{F}_1$ score of 45.46% is obtained for the CWL data set indicating that training the meta model on SOS and testing it on CWL is more effective than vice versa. In addition, we provide results for a different meta classification model which is trained and evaluated per leave-one-out cross validation on the respective data set, see appendix B. In fig. 3, an example image

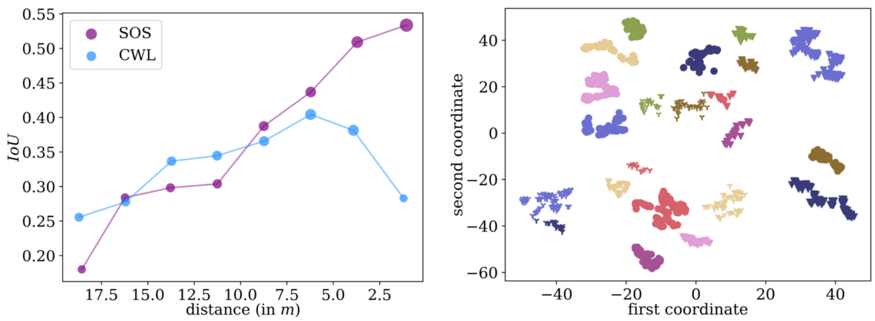


Fig. 4: *Left*: Discretized distance between ground truth objects and camera vs. mean IoU over all object types of the SOS and the CWL data set, respectively. The dot size is proportional to mean segment size. *Right*: Clustering of OOD segments predicted for the CWL data set with min. tracking frequency $\ell = 10$.

of our OOD segmentation method is presented. The final prediction mask after entropy thresholding and meta classification contains only true OOD objects. In appendix C and appendix D, more numerical results evaluated for depth binnings and on individual OOD classes are presented, respectively.

OOD Tracking Building upon the OOD segmentation masks obtained, in this subsection we report OOD tracking results. We consider several object tracking metrics (see section 4.2) shown in table 1 for the SOS and CWL data set. We observe a comparatively low $MOTA$ performance for the SOS data set. The underlying reason is a high number of false positive segments that are accounted for in this metric, as also shown in the detection metric $\bar{F}_1$.

Furthermore, most of the ground truth objects are at least partially tracked, only 3 out of 26 and 8 out of 62 ground truth objects are largely lost out for SOS and CWL, respectively. Analogously, in fig. 4, we observe that most ground truth objects are matched with predicted ones for the SOS data set. This plot shows the correlation between the IoU (of ground truth and predicted objects) and the distance of the ground truth objects to the camera as we provide meta data like depth for our data sets. We observe for both data sets that the IoU increases with decreasing distances, the only exception are very short distance objects to the ego-car for the CWL data set. Moreover, we provide video sequences[4] that visualize the final OOD segmentation and object tracking results. In appendix D, more numerical results evaluated on individual OOD classes are presented.

Retrieval of OOD Objects Finally, we evaluate the clustering of OOD segments obtained by the OOD object segmentation method introduced in section 5.1. In table 1, we report the clustering metrics CS_{inst}, CS_{imp} and CS_{frag}

[4] https://youtu.be/_DbV8XprDmc

(see section 4.3) with ($\ell = 10$) and without ($\ell = 0$) incorporating the OOD tracking information, respectively. For both, the CWL and the SOS data set, all clustering metrics improve when applying the OOD tracking as a pre-processing step. A reason for this is, that the tracking information "tidies up" the embedding space, e.g. by removing noise, which enhances the performance of the clustering algorithm. For CWL (with 18 object types), 1266/1026 OOD segments are clustered into 22/23 clusters without/with using tracking results, for SOS (with 13 object types), we obtain 23/24 clusters which contain 1437/888 OOD segments in total. For the clustering, we applied the DBSCAN algorithm with hyperparameters $\varepsilon = 4.0$ and minPts $= 15$. In fig. 4, we exemplarily visualize the clustered embedding space for the CWL data set with $\ell = 10$. The remaining visualizations as well as additional results for the second meta classification model are provided in appendix B. Furthermore, we visualize some clustering results for the WOS data set in appendix E. As WOS comes without labels, we do not report any evaluation metrics, but provide some visualizations for the 5 largest clusters.

6 Conclusion and Outlook

We created a baseline for the CV task of tracking OOD objects by (a) publishing two data sets with 20 (SOS) and 26 (CWL) annotated video sequences containing OOD objects on street scenes and (b) presenting an OOD tracking algorithm that combines frame-wise OOD object segmentation on single frames with tracking algorithms. We also proposed a set of evaluation metrics that permit to measure the OOD tracking efficiency. As an application, we retrieved new, previously unlearned objects from video data of urban street scenes.

To go beyond this baseline, several directions of research seem to be promising. First, OOD segmentation on video data could benefit from 3D CNN acting on the spatial and temporal dimension, rather than combining 2D OOD segmentation with tracking. However, at least for those OOD segmentation algorithms that involve OOD training data, new and specific video data sets would be required. Similarly, genuine video sequence based retrieval algorithms should be developed to improve our revival baseline. Such algorithms could prove useful to enhance the coverage of urban street scenes in training data sets for AI-based perception in automated driving.

Acknowledgements We thank Sidney Pacanowski for the labeling effort, Dariyoush Shiri for support in coding, Daniel Siemssen for support in the generation of CARLA data and Matthias Rottmann for interesting discussions. This work has been funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) via the research consortia Safe AI for Automated Driving (grant no. 19A19005R), AI Delta Learning (grant no. 19A19013Q), AI Data Tooling (grant no. 19A20001O) and the Ministry of Culture and Science of the German state of North Rhine-Westphalia as part of the KI-Starter research funding program.

References

1. Angus, M., Czarnecki, K., Salay, R.: Efficacy of Pixel-Level OOD Detection for Semantic Segmentation. arXiv pp. 1–13 (Nov 2019) 4
2. Arandjelović, R., Zisserman, A.: Multiple queries for large scale specific object retrieval. In: BMVC (2012) 5
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. In: Proceedings of the British Machine Vision Conference (BMVC). pp. 1–12 (Sep 2017) 4
4. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. In: ECCV (2006) 5
5. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The clear mot metrics. EURASIP Journal on Image and Video Processing (2008) 8
6. Bertasius, G., Torresani, L.: Classifying, segmenting, and tracking object instances in video with mask propagation. ArXiv **abs/1912.04573** (2019) 4
7. Besnier, V., Bursuc, A., Picard, D., Briot, A.: Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15701–15710 (October 2021) 4
8. Bevandić, P., Krešo, I., Oršić, M., Šegvić, S.: Simultaneous Semantic Segmentation and Outlier Detection in Presence of Domain Shift. In: Proceedings of the German Conference on Pattern Recognition (GCPR). pp. 33–47. Dortmund, Germany (Oct 2019) 4
9. Blum, H., Sarlin, P.E., Nieto, J., Siegwart, R., Cadena, C.: Fishyscapes: A Benchmark for Safe Semantic Segmentation in Autonomous Driving. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops. pp. 2403–2412. Seoul, Korea (Oct 2019) 4
10. Blum, H., Sarlin, P.E., Nieto, J., Siegwart, R., Cadena, C.: The fishyscapes benchmark: measuring blind spots in semantic segmentation. International Journal of Computer Vision **129**(11), 3119–3135 (2021) 2, 3
11. Brüggemann, D., Chan, R., Rottmann, M., Gottschalk, H., Bracke, S.: Detecting out of distribution objects in semantic segmentation of street scenes. In: The 30th European Safety and Reliability Conference (ESREL). vol. 2 (2020) 2
12. Bulatov, Y.: Notmnist dataset. Google (Books/OCR), Tech. Rep.[Online]. Available: http://yaroslavvb. blogspot. it/2011/09/notmnist-dataset. html **2** (2011) 3
13. Bullinger, S., Bodensteiner, C., Arens, M.: Instance flow based online multiple object tracking. pp. 785–789 (2017). https://doi.org/10.1109/ICIP.2017.8296388 4
14. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020) 1
15. Chan, R., Lis, K., Uhlemeyer, S., Blum, H., Honari, S., Siegwart, R., Fua, P., Salzmann, M., Rottmann, M.: SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. In: Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2021) 2, 3, 4, 7, 8, 10
16. Chan, R., Rottmann, M., Gottschalk, H.: Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5128–5137 (October 2021) 2, 4, 9, 10

17. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K.P., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**, 834–848 (2018) 1
18. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3213–3223 (2016) 1, 3, 5, 9
19. Creusot, C., Munawar, A.: Real-time small obstacle detection on highways using compressive rbm road reconstruction. 2015 IEEE Intelligent Vehicles Symposium (IV) pp. 162–167 (2015) 4
20. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 12
21. Devries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. ArXiv **abs/1802.04865** (2018) 4
22. Di Biase, G., Blum, H., Siegwart, R., Cadena, C.: Pixel-wise anomaly detection in complex driving scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16918–16927 (June 2021) 2, 4
23. Dosovitskiy, A., et al.: CARLA: an open urban driving simulator. In: Proc. of CoRL. pp. 1–16. Mountain View, USA (2017) 6
24. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD (1996) 12
25. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: the qbic system. Computer **28**(9), 23–32 (1995). https://doi.org/10.1109/2.410146 5
26. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (6 2016) 4
27. Games, E.: Unreal engine (2004-2022), https://www.unrealengine.com 2, 6
28. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013) 1
29. Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S., Fernandez, T., Jänicke, M., Mirashi, S.G., Savani, C., Sturm, M., Vorobiov, O., Oelker, M., Garreis, S., Schuberth, P.: A2d2: Audi autonomous driving dataset. ArXiv **abs/2004.06320** (2020) 1
30. Goodfellow, I.J., Bulatov, Y., Ibarz, J., Arnoud, S., Shet, V.: Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082 (2013) 3
31. Grcić, M., Bevandić, P., Šegvić, S.: Dense open-set recognition with synthetic outliers generated by real nvp. arXiv preprint arXiv:2011.11094 (2020) 2
32. Grcić, M., Bevandić, P., Šegvić, S.: Dense anomaly detection by robust learning on synthetic negative data. arXiv preprint arXiv:2112.12833 (2021) 2, 4
33. Guadarrama, S., Rodner, E., Saenko, K., Zhang, N., Farrell, R., Donahue, J., Darrell, T.: Open-vocabulary object retrieval. In: Robotics: Science and Systems (2014) 5

34. Gustafsson, F.K., Danelljan, M., Schön, T.B.: Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1289–1298. virtual conference (Jun 2020) 4
35. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385 11
36. Hein, A.M.: Identification and bridging of semantic gaps in the context of multi-domain engineering. In: Proceedings 2010 Forum on Philosophy, Engineering & Technology (01 2010) 5
37. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why ReLU Networks Yield High-Confidence Predictions Far Away from the Training Data and How to Mitigate the Problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 41–50. Long Beach, CA, USA (Jun 2019) 4
38. Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings (2020) 3, 4
39. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (2017) 4
40. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep Anomaly Detection With Outlier Exposure. In: Proceedings of the International Conference on Learning Representations (ICLR). pp. 1–18. New Orleans, LA, USA (May 2019) 4
41. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4555–4564 (2016) 5
42. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2261–2269 (2017) 3
43. Huang, X., Xu, J., Tai, Y.W., Tang, C.K.: Fast video object segmentation with temporal aggregation network and dynamic template matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 4
44. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NIPS (2017) 4
45. Krizhevsky, A., Nair, V., Hinton, G.: The cifar-10 dataset. online: http://www. cs. toronto. edu/kriz/cifar. html **55**(5) (2014) 3
46. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science pp. 1332–1338 (2015) 3
47. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NIPS (2017) 4
48. Lateef, F., Ruichek, Y.: Survey on semantic segmentation using deep learning techniques. Neurocomputing **338**, 321–348 (2019) 1
49. Lecun, Y.: The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/ (2010), https://ci.nii.ac.jp/naid/10027939599/en/ 3
50. Lee, K., Lee, K., Lee, H., Shin, J.: A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS/NeurIPS). pp. 7167–7177. Montréal, QC, Canada (Dec 2018) 4

51. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations (2018) 4
52. Lin, C.C., Hung, Y., Feris, R., He, L.: Video instance segmentation tracking with a modified vae architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 4
53. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Eur. Conf. Comput. Vis. pp. 740–755. Zurich, Switzerland (Sep 2014) 10
54. Lis, K., Honari, S., Fua, P., Salzmann, M.: Detecting road obstacles by erasing them. arXiv preprint arXiv:2012.13633 (2020) 2, 4
55. Lis, K., Nakka, K., Fua, P., Salzmann, M.: Detecting the unexpected via image resynthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2152–2161 (2019) 2, 3, 4
56. LoweDavid, G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (2004) 5
57. Maag, K.: False negative reduction in video instance segmentation using uncertainty estimates. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). pp. 1279–1286. IEEE (2021) 2
58. Maag, K., Rottmann, M., Gottschalk, H.: Time-dynamic estimates of the reliability of deep semantic segmentation networks. 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI) pp. 502–509 (2020) 2, 4, 5, 10, 11
59. Maag, K., Rottmann, M., Varghese, S., Hueger, F., Schlicht, P., Gottschalk, H.: Improving video instance segmentation by light-weight temporal uncertainty estimates. arXiv preprint arXiv:2012.07504 (2020) 2, 4
60. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 3, 12
61. Maji, S., Bose, S.: Cbir using features derived by deep learning. ACM/IMS Transactions on Data Science (TDS) **2**, 1 – 24 (2021) 5
62. Mao, J., Huang, J., Toshev, A., Camburu, O.M., Yuille, A.L., Murphy, K.P.: Generation and comprehension of unambiguous object descriptions. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11–20 (2016) 5
63. Meinke, A., Hein, M.: Towards Neural Networks that Provably Know When They Don't Know. In: Proceedings of the International Conference on Learning Representations (ICLR). pp. 1–18. virtual conference (Apr 2020) 4
64. Milan, A., Leal-Taixé, L., Reid, I.D., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. ArXiv **abs/1603.00831** (2016) 4, 8
65. Mukhoti, J., Gal, Y.: Evaluating bayesian deep learning methods for semantic segmentation. ArXiv **abs/1811.12709** (2018) 4
66. Munawar, A., Vinayavekhin, P., De Magistris, G.: Limiting the Reconstruction Capability of Generative Neural Network Using Negative Learning. In: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. Tokyo, Japan (Sep 2017) 4
67. Naaz, E., Kumar, T.: Enhanced content based image retrieval using machine learning techniques. 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) pp. 1–12 (2017) 5
68. Neuhold, G., Ollmann, T., Bulò, S.R., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. pp. 5000–5009 (2017) 1

69. Oberdiek, P., Rottmann, M., Fink, G.A.: Detection and retrieval of out-of-distribution objects in semantic segmentation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 1331–1340 (2020) 3, 5, 11
70. Pearson F.R.S., K.: LIII. On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 1 **2**, 559–572 (1901) 12
71. Pinggera, P., Ramos, S., Gehrig, S., Franke, U., Rother, C., Mester, R.: Lost and Found: Detecting Small Road Hazards for Self-Driving Vehicles. In: International Conference on Intelligent Robots and Systems (IROS). pp. 1099–1106. Daejeon, Korea (Oct 2016) 3, 10
72. Porzi, L., Hofinger, M., Ruiz, I., Serrat, J., Bulo, S.R., Kontschieder, P.: Learning multi-object tracking and segmentation from automatic annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 4
73. Rottmann, M., Colling, P., Hack, T.P., Hüger, F., Schlicht, P., Gottschalk, H.: Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. 2020 International Joint Conference on Neural Networks (IJCNN) pp. 1–9 (2020) 8, 10
74. Rottmann, M., Schubert, M.: Uncertainty measures and prediction quality rating for the semantic segmentation of nested multi resolution street scene images. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 1361–1369 (2019) 10
75. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.C.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1349–1380 (2000) 5
76. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B **58**, 267–288 (1996) 10
77. Uhlemeyer, S., Rottmann, M., Gottschalk, H.: Towards unsupervised open world semantic segmentation (2022) 3, 11
78. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**, 3349–3364 (2021) 1
79. Xia, Y., Zhang, Y., Liu, F., Shen, W., Yuille, A.: Synthesize Then Compare: Detecting Failures and Anomalies for Semantic Segmentation. In: Eur. Conf. Comput. Vis. pp. 145–161. virtual conference (Aug 2020) 4
80. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017) 3
81. Yang, L., Fan, Y., Xu, N.: Video instance segmentation pp. 5187–5196 (2019) 4
82. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2633–2642 (2020) 1
83. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015) 3
84. Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 8848–8857 (2019) 9