Have We Ever Encountered This Before? Retrieving Out-of-Distribution Road Obstacles from Driving Scenes

Y. Shoeb[1,4] R. Chan[2,4] G. Schwalbe [1] A. Nowzad[1] F. Güney[3] H. Gottschalk[4]
[1] Continental AG [2] Bielefeld University [3] Koç University [4] Technische Universität Berlin
{youssef.shoeb, azarm.nowzad, gesina.schwalbe}@continental.com
fguney@ku.edu.tr {chan, gottschalk}@math.tu-berlin.de

Abstract

In the life cycle of highly automated systems operating in an open and dynamic environment, the ability to adjust to emerging challenges is crucial. For systems integrating data-driven AI-based components, rapid responses to deployment issues require fast access to related data for testing and reconfiguration. In the context of automated driving, this especially applies to road obstacles not included in the training data, commonly referred to as out-of-distribution (OoD) road obstacles. Given the availability of large uncurated driving scene recordings, a pragmatic approach is to query a database to retrieve similar scenarios featuring the same safety concerns due to OoD road obstacles. In this work, we extend beyond identifying OoD road obstacles in video streams and offer a comprehensive approach to extract sequences of OoD road obstacles using text queries, thereby proposing a way of curating a collection of OoD data for subsequent analysis. Our proposed method leverages the recent advances in OoD segmentation and multi-modal foundation models to identify and efficiently extract safety-relevant scenes from unlabeled videos. We present a first approach for the novel task of text-based OoD object retrieval, which addresses the question "Have we ever encountered this before?".

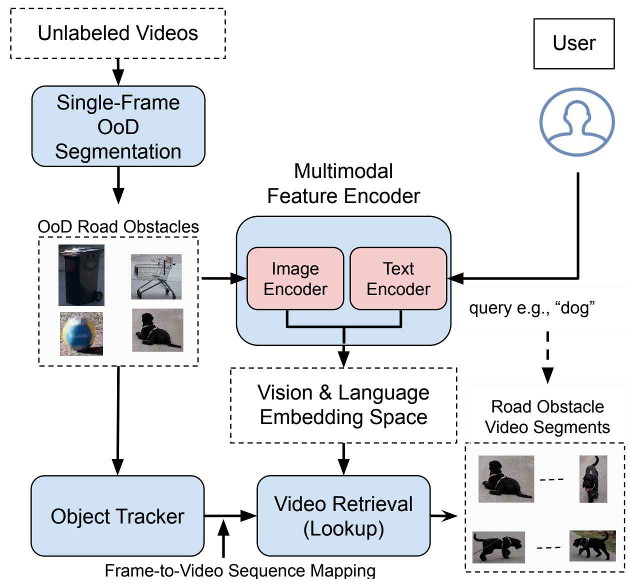

Figure 1. **Overview of Our Method**: We extract specific safety-critical driving scenes due to out-of-distribution (OoD) road obstacles from unlabeled videos based on a text query, such as "dog". The approach leverages single-frame OoD segmentation, object tracking, and multi-modal feature encoding of OoD images to enable *text-to-video retrieval* of *OoD road obstacles*.

1. Introduction

Imagine a scenario where a self-driving vehicle is involved in a collision with a dog. Following the incident, an investigation team is set up to determine the root cause of the accident. For data-driven AI-based components, the investigation team would prioritize acquiring sensory data, such as videos with prior encounters with dogs, to reproduce the error and asses the perception system. Having an understanding of the situation, including the vehicle's environmental perception leading up to the incident, can help in refining its driving policy and prevent future incidents.

The previous example illustrates the need to acquire targeted video data in future life cycles of perception components in self-driving cars. Promptable video synthesis using generative models [25,55,64] could be a suitable way to acquire such data. However, questions about the coverage of the generated distribution and the extent of the domain gap would still prevail [43, 53]. An alternative that we follow here is to retrieve relevant data from real-world recordings. However, existing video retrieval approaches [40,46,54] require processing up to millions of hours of recorded data, which is highly resource-intensive and slow when applied. This is why an efficient screening and preselection of relevant scenes is key.

In this work, we focus on safety-critical driving scenes containing unknown road obstacles. In self-driving cars, deep neural networks (DNNs) are employed for perception tasks, and they are trained to identify and locate objects within images given a predefined set of object categories [22,39,50]. The number of these predefined classes for standard automated driving datasets ranges from 11 classes in KITTI [35] to 19 classes in BDD100K [65] and Cityscapes [15]. Those classes include common semantic categories such as pedestrian, road, or sidewalk. However, the diversity of the real world offers a boundless set of possible object categories, making DNNs particularly error-prone when processing semantically unknown objects, commonly known as out-of-distribution (OoD) objects. A particular and safety-critical OoD subset in automated driving consists of OoD road obstacles, which are unknown objects present within the drivable area of a self-driving car [29, 37, 47]. Identifying those objects is a crucial prerequisite for building an OoD database for further analysis and subsequent adjustment of the perception system [2, 18].

Our target is to enable the perception stack to efficiently retrieve safety-relevant video sequences of OoD road obstacles from prior recordings using text queries. For the related task of image retrieval, one primary challenge is aligning image and query features into a joint embedding space for fast retrieval. Additionally, in the context of OoD road obstacle retrieval, the absence of existing OoD video segmentation approaches (instead of per-frame segmentation) poses another methodological challenge of identifying the same OoD road obstacles over multiple consecutive frames.

A parallel line of research in image retrieval explores the construction of feature embeddings based on visual similarities by utilizing DenseNet feature encodings [27] to cluster OoD objects [41, 45, 57, 58]. Those approaches, however, come with limitations that constrain the application of targeted retrieval of objects: (1) the process of clustering in the embedding space is driven by visual similarities, wherefore distinct instances may be assigned into separate clusters even if they belong to the same semantic category, (2) the retrieval from already formed clusters can only be performed content-based, which requires an image query or manually assigned labels for clusters provided by human annotators and (3) all current approaches only retrieve single frames rather than complete sequences.

In this work, we propose a method for processing unlabeled video data from commonly available in-vehicle cameras and extracting driving scenes that contain OoD road obstacles. In the first step, our approach provides detailed information about the presence and trajectory of a singular OoD road obstacle within a video *sequence*, thereby extending beyond the conventional task of identifying any OoD object in a single frame to a set of consecutive frames. Next, we offer a method to retrieve sequences that contain the same or similar OoD road obstacles matching a *textual description* provided by a user. The combination of the two steps leads to a novel approach for resource-efficient and fast text-to-video retrieval of safety-critical driving scenes that leverages the most recent advances both in single-frame OoD segmentation and multi-modal feature encoding.

In particular, we perform single-frame OoD segmentation first and track identified OoD road obstacles through frames using a lightweight object tracker. The single frames are then embedded in a multi-modal embedding space. This use of semantically meaningful embedding space enables the retrieval of frames containing OoD road obstacles that match the given text query, while the tracking information allows for the retrieval of the complete sequence of frames where the OoD road obstacle was present. This is the first work to leverage the recent progress in single frame OoD segmentation [44] and the power of the recently established multi-model foundation models [48] for combined image and language understanding for OoD retrieval with application in automated driving. An overview of our method is shown in Figure 1.

We summarize our contribution as follows:

- We propose a novel modular approach for efficient text-to-video retrieval of safety-critical driving scenes containing OoD road obstacles. Our framework implements and validates the key ideas: (1) leveraging a multi-modal embedding space for text-to-image retrieval, (2) utilizing temporal information and object persistency by making use of tracking to extend single-frame retrieval to video data, and (3) using meta classification for segment-wise false positive removal to refine the OoD segmentations for better accuracy.
- By using CLIP's multi-modal embedding space [48], we generate clusters of OoD road obstacle sequences in low-dimensional feature space that enable proper text-to-video retrieval based on semantic similarities rather than visual similarities.
- Through extensive experiments, we investigate the interaction of each component within our framework and their impact on the overall retrieval performance. Our findings underline the significance as well as the apparent positive effect of each module on the OoD road obstacle retrieval performance, which are object-level processing, object tracking, prediction of region on interest, and meta classification.

2. Related Work

OoD Segmentation: Semantic segmentation models group pixels in an image into segments that adhere to specific predefined semantic classes. Image parts that don't belong to any of the predefined set of classes are referred to as out-of-distribution (OoD). Typically, semantic segmentation models struggle to detect OoD segments [10].

One approach to overcome this limitation is to leverage sampling-based uncertainty estimation approaches, such as Monte-Carlo dropout [19], ensembles [33], or variational inference [20]. Those quantify the predictive uncertainty of the model and use it to identify OoD objects as unknown. However, these methods are computationally expensive and suffer from numerous false positives in boundary regions between objects, as these exhibit natural uncertainties [32]. A more effective approach is to include auxiliary training data as a proxy for unknown objects to either maximize the softmax entropy [11] or minimize the maximum logits score [44] of unknown objects. Other methods have achieved promising results by aggregating pixel-level uncertainty information into mask-level predictions [23,44,51]. These approaches use segmentation models that perform mask-level segmentation to make predictions about unknown objects. More recent techniques [61] detect road obstacles by explicitly learning a feature embedding space that models the multi-modal appearance of road surfaces.

The approach we consider most promising is the one proposed by [44], as indicated by their results on the SegmentMeIfYouCan road obstacle segmentation benchmark [10]. Their approach succeeds due to the use of mask classification to preserve objectness and a scoring function that eliminates irrelevant sources of uncertainty. We follow the same method for segmenting OoD road obstacles; however, we enhance the frame-based detection module of [44] by incorporating segment tracking on videos to eliminate some of the false positive detections.

Multiple Object Tracking (MOT) is the task of determining the spatial and temporal location of multiple objects in a sequence of images. Two possible approaches for MOT tracking are: (a) Converting existing detectors into trackers and combining both tasks in the same framework. These methods either use 3D convolutions on consecutive frames to incorporate temporal information [30, 31, 60] or propagate frame-level information to subsequent frames [3,5,67]. However, combining tracking and detection into one model sacrifices the modularity of the tasks, which is desirable for reuse and inspectability in safety-relevant applications [28, 5.4.2 c)]. (b) Tracking by detection methods, which first utilize a pre-trained object detector to detect objects and then track them through a sequence of frames, for example, via data association [8, 34], visual cues [21, 63], or motion patterns [6, 9]. [38] proposed an approach for tracking developed explicitly for use in open-world conditions. Their method uses optical flow and an appearance-based similarity score to detect and track moving objects in an open-world setting. In this work, we use the lightweight tracking by detection approach proposed in [42] to track OoD objects in a sequence of images. This tracking method is a post-processing method based on the overlap of detections between consecutive frames.

Retrieval Methods are generally designed to identify and recover samples from a large database corresponding to a given query. For image retrieval, methods can be classified into two categories: content-based image retrieval and text-based image retrieval [12,66].

Content-based image retrieval methods are based on a query image. These methods aim to select images from a database representing a similar content as the query image. Content-based retrieval techniques analyze visual features of images, including color, texture, or shape, to establish similarity between the images in the database and the given query image [13, 17]. Text-based image retrieval methods focus on selecting images that exhibit the highest level of relevance to a given text query. These systems utilize textual information, such as keywords or natural language descriptions, to retrieve images from a database that best aligns with the provided text [1, 24, 26].

For the task of text-video retrieval, a rich line of research has evolved from the global matching of features via video-sentence alignment [40] to more fine-grained matching via frame-word alignment [62]. These studies have demonstrated remarkable performance and significantly outperformed previous models on the task of text-video retrieval. This is mainly due to the powerful pre-aligned visual and textual representation offered by open-source models like CLIP [48]. In [40], the authors utilize a temporal transformer on top of CLIP to fuse sequential features into a single high-dimensional representation and directly retrieve video segments. However, for the automotive use case, hardware constraints have to be fulfilled. Therefore, in this work, we use a lightweight tracking module on top of CLIP to perform text-video retrieval.

3. Methodology

This work focuses on retrieving OoD road obstacle sequences from unlabeled videos based on a text query. Our method consists of three key steps. First, we identify the occurrence of OoD road obstacles in single frames. Second, we track the OoD road obstacles through consecutive frames, creating sequences of frames where the same road obstacle appears. Third, we enable user interaction via text-based retrieval of sequences. Our method is set up such that after the first and second steps, a database with driving scenes containing OoD road obstacles can be established. Considering that such scenes are substantially less prevalent, this approach resolves problems with the bandwidth constraints of autonomous vehicles and potential storage limitations within cloud-based systems. Afterward, the crops of each OoD road obstacle can be embedded once in a vision-text embedding space; this embedding space allows for retrieving sequences when a user provides a text query.

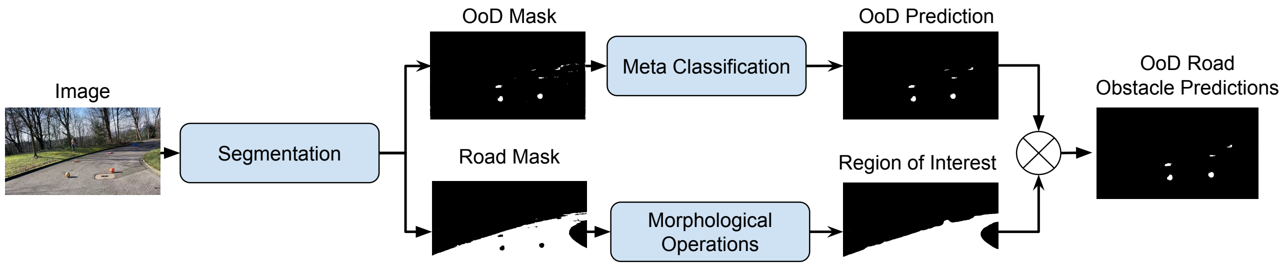

Figure 2. **OoD Road Obstacle Segmentation Overview:** Our method segments OoD objects and the road, refines OoD objects using meta classification, generates a region of interest through road mask dilation and erosion, and obtains final OoD road obstacle predictions by combining OoD predictions and the region of interest mask.

Since each crop can be associated with its respective video sequence, fast retrieval of sequences containing OoD road obstacles is enabled.

To accomplish this, our proposed method integrates various auxiliary tasks, including OoD segmentation, multi-object tracking, and text-based image retrieval. These tasks collectively constitute the overall framework. The following sections present a detailed description of each task.

3.1. OoD Road Obstacle Segmentation

A complete overview of the OoD road obstacle segmentation method is shown in Figure 2. The initial phase of the OoD road obstacle segmentation module involves using a semantic segmentation network. In our experiments, we use the Mask2Former model [14] initially trained on the Cityscapes dataset [15]. Mask2former decouples localization and classification of objects in semantic segmentation by splitting the task into two steps. Given an $H \times W$ sized image, Mask2former computes N pairs $\{(\mathbf{m}_i, \mathbf{p}_i)\}_{i=1}^N$, where $\mathbf{m}_i \in [0,1]^{H\times W}$ are mask predictions associated with some semantically related regions in the input image and $\mathbf{p}_i \in [0,1]^{K+1}$ class probabilities classifying to which semantic category the mask $\mathbf{m}_i$ belongs to. Here, the masks can be assigned to one of the K known Cityscapes classes or to one auxiliary void class. The final semantic segmentation inference is carried out by an ensemble-like approach over the pairs $\{(\mathbf{m}_i, \mathbf{p}_i)\}_{i=1}^N$ yielding pixel-wise class scores

$$\mathbf{q}[h,w,k] = \sum_{i=1}^{N} \mathbf{p}_i(k) \cdot \mathbf{m}_i[h,w] \quad \in [0,N] \qquad (1)$$

for image pixel locations $h = 1,\dots,H, w = 1,\dots,W$ and classes $k = 1,\dots,K$. Then, OoD detection is performed via the anomaly score defined by

$$\mathbf{RbA}[h,w] = -\sum_{k=1}^{K} \phi(\mathbf{q}[h,w,k]) \quad \in [0,K] \qquad (2)$$

with ϕ being the $\tanh$ activation function. Intuitively, $\mathbf{RbA}$ in Equation (2) is a measure of whether a pixel cannot be associated to any known class, and thus "Rejected by All" (RbA), of the K known classes. This scoring function has been introduced in [44]. In the same work, the authors additionally fine-tune Mask2Former for OoD detection by training for low-class scores of the known classes on OoD instances from COCO [36], which has shown to enhance OoD segmentation performance further. This fine-tuned Mask2Former serves as our method for OoD road obstacle segmentation in this work.

Post-processing OoD Predictions: To reduce false positive predictions, meta-classification [11, 51, 52] is used to obtain quality ratings for the OoD predictions. Meta-classification uses hand-crafted metrics like entropy, geometry, and location information of predicted instances to learn the features of false positive predictions on the training set. During run-time, the meta-classification model, in our case a logistic regression, can remove false positives without any ground truth information. We refer the reader to [11] for a detailed description of the approach.

Post-processing Road Segmentation: By definition, road obstacles are objects on the road. Consequently, we can restrict our predictions exclusively to objects on the road by establishing a region of interest (RoI) mask that encompasses the road area. The RoI mask can be obtained by extracting the road predictions from the Mask2Former semantic segmentation model combined with morphological closing [56] to fill gaps where potential road obstacles might be present. The final OoD road obstacle predictions are obtained by simply masking the OoD predictions with the region of interest.

3.2. OoD Object Tracking

Given predictions of OoD road obstacles in each frame, as described in section 3.1, we match subsequent predictions through consecutive frames by measuring the segment-wise intersection over union (IoU) and the geometric centers between consecutive detections.

The first step of the tracking approach assigns random identifiers to all the predicted segments in the first frame.

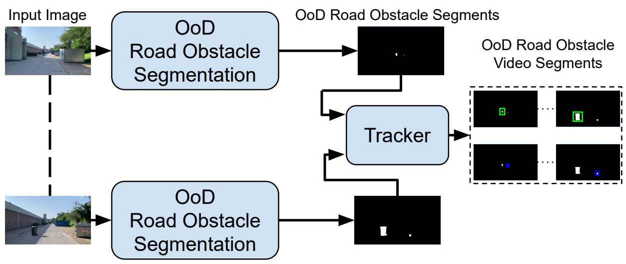

Figure 3. **OoD Tracking Overview**: Given subsequent frames of OoD obstacles, we use a lightweight tracker that assigns visually and spatially similar segments to the same tracking ID.

For the subsequent frames, each segment is matched with segments in the previous frames if their overlap is sufficiently large and their geometric centers are close enough. Over consecutive frames, linear regression is applied to account for misdetections and temporal occlusions. Segments that do not match with previous detections are assigned new identifiers, and then the process is iterated. We note that this lightweight tracker does not apply any motion models to anticipate the shifted center points of the detections. Hence, the assumption is that the differences between consecutive frames are minimal, leading to a substantial Intersection over Union (IoU) across frames.

To reduce the number of false positive detections, tracked segments in sequences of frames with a length of less than ten frames are filtered out. The assumption is that in the context of automated driving, informative OoD road obstacles persist in the field of view of the vehicle for a couple of frames. The final output of the tracking module is a sequence of cropped segments that belong to a single instance of an OoD road obstacle. An overview of the OoD tracking module is shown in Figure 3.

3.3. Retrieval of Road Obstacle Video Sequences

OoD road obstacle segmentation and tracking allows for creating a database of video sequences, with each sequence consisting of consecutive crops of an OoD road obstacle from a video recording. Given a textual query, the goal of OoD road obstacle retrieval now is to find those video sequences that best match the query. For this, we utilize CLIP [48] to align image and text features into a joint embedding space where their similarity can be quantified. Using this approach, natural language supervision guides the model to understand that latent representations of semantically similar contents of images should be close in the embedding space.

We retrieve video sequences of OoD objects similar to a textual query as follows: In our database, each element comprises consecutive image crops of OoD road obstacles, which we identify and associate during the OoD detection and tracking steps. We then compare the embedding space

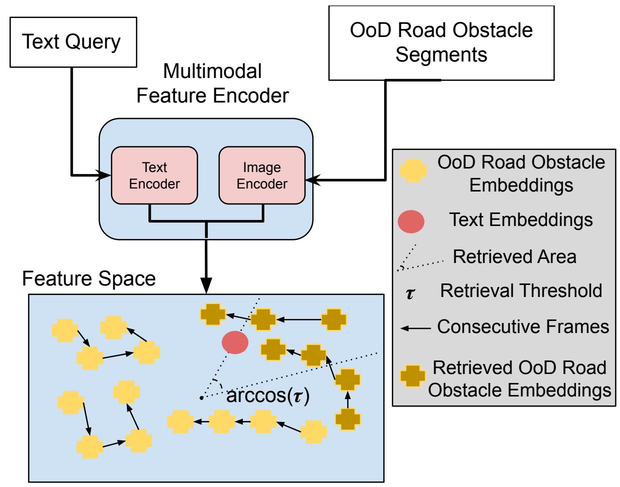

Figure 4. **OoD Road Obstacle Retrieval Overview:** Given detections of OoD road obstacles and a text query, both text and images are embedded in a single multi-model embedding space. Using this embedding space, all OoD road obstacles within a given threshold τ from the text query embedding are retrieved.

representation of the given text query to the by-frame OoD road obstacle crops. To determine the similarity between a video sequence and the text query, we aggregate the similarities of the sequence's individual OoD crops. We retrieve and present the most similar sequences to the user. Specifically, we use cosine similarity to measure the similarity between the embedding representation of the query text and the individually cropped OoD road obstacle detections. For each uniquely detected object in a given sequence, we measure the frame-to-text similarity. The highest similarity score among all the crops in a sequence determines the overall similarity score for a sequence of crops of an OoD road obstacle.

In more detail, for every cropped OoD road obstacle detection $\mathbf{x}_j$ in a detected sequence of crops $\mathbf{S}_k = \{\mathbf{x}_j\}_{j=1}^{n_k}$, the image embeddings g_j are computed as

$$\mathbf{g}_j = \mathcal{E}_{\text{image}}(\mathbf{x}_j) \in \mathbb{R}^d \tag{3}$$

where $\mathcal{E}_{\text{image}}$ is a Vision Transformer ViT-B/32 [16] image encoder. Then given a text query t, a text embedding $\mathbf{f}$ is computed as:

$$\mathbf{f} = \mathcal{E}_{\text{text}}(t) \in \mathbb{R}^d \tag{4}$$

where $\mathcal{E}_{\text{text}}$ is a Transformer text-encoder [59] with modifications described in [49]. To quantify the *semantic similarity* between an image-text pair, we measure the pairwise cosine similarities between their embeddings. Cosine similarity quantifies the angle between the representation vectors and is a typical similarity measure for text embedding space; it is calculated as follows:

$$s(g_j, f) = \frac{{\mathbf{g}_j}^\top \mathbf{f}}{\|\mathbf{g}_j\|_2 \, \|\mathbf{f}\|_2} \quad \in [-1, 1] \tag{5}$$

A sequence $\mathbf{S}_k$ is considered a *positive* match to a text query $\mathbf{f}$ if, for any of the frames in the sequence, the similarity score of its embeddings exceeds a chosen similarity threshold $\tau \in [-1, 1]$, *i.e.* if

$$\exists\, \mathbf{g}_j \in \mathbf{S}_k : s(\mathbf{g}_j, \mathbf{f}) \geq \tau \, . \tag{6}$$

Note that retrieving the image with the highest similarity to the text query is sufficient to retrieve the entire corresponding OoD road obstacle sequence as the remaining images of the sequences are associated by tracking information, *cf*. Figure 4

4. Experiments

This section presents our experimental findings and setup. Since this specific task has not been addressed in previous literature, there are no standard baselines available to compare against. Therefore, we present two main experiments: (1) an investigation into the importance of object-level processing instead of direct image-level processing for retrieval and (2) an ablation study of the individual components of our proposed method. The investigation into object-level processing compares the approach of segmenting, tracking, and retrieving based on cut-outs of OoD road obstacles against direct retrieval on entire images. Additionally, the effects of tracking are evaluated. The ablation study consists of three experiments. In the first experiment, we evaluate the efficacy of our proposed method for the task of OoD retrieval. We report the results of our proposed method for segmentation, tracking, and retrieving OoD road obstacles using two different OoD segmentation networks. Additionally, we compare the retrieval performance against the same approach but using perfect detections. The second and third experiments evaluate the effects of the RoI segmentation and meta-classification on the detection, tracking, and retrieval performance, respectively.

Datasets: We perform experiments on the publicly available Street Obstacle Sequences (SOS), Carla-WildLife (CWL), and Wuppertal Obstacle Sequences (WOS) [41]. The SOS dataset contains 20 real-world video sequences with 13 different OoD objects. The CWL dataset contains 26 synthetic video sequences with 18 different OoD objects. WOS contains 44 real-world video sequences with seven different OoD objects. In all the above, we consider OoD objects as objects not included in Cityscapes labels. We target retrieving all occurrences of the different OoD objects from the three datasets.

OoD Segmentation Evaluation: We follow the standard evaluation protocol for the pixel-level performance measures adopted from [7, 47]. Namely, these are the Area Under Precision-Recall Curve (AUPRC) and the False Positive Rate at 95% of True Positive Rate (FPR_{95}). From a practitioner's perspective, it is often sufficient only to recognize a fraction of the pixels of an OoD object to detect and localize them. For evaluating the component-level performance of the OoD segmentation model, the averaged component-wise score $\overline{F}_1$ [10] serves as our main evaluation metric. The threshold for considering true positive and false positive is set to be the threshold that optimizes the pixel-wise F_1 score. We note that the standard evaluation protocol for OoD segmentation only evaluates predictions that fall into a labeled region of interest [10, 47]. Since expensive ground truth segmentation labels cannot be assumed to be available for large-scale OoD analysis, this assumption must be relaxed for applications that utilize the OoD predictions for downstream tasks. We report the $\overline{F}_1$ score on the predicted road regions instead of ground truth regions of interest.

Tracking Evaluation: We evaluate the object tracker performance using the common multiple object tracking (MOT) metrics [4]. These metrics quantify the algorithm's ability to accurately detect the number of objects present and determine the position of each object. The Multiple Object Tracking Accuracy (MOTA) is a metric that evaluates the tracking algorithm's performance in detecting objects and maintaining their trajectories, regardless of the precision with which the object positions are estimated. On the other hand, the Multiple Object Tracking Precision (MOTP) assesses the tracker's ability to accurately estimate the positions of objects, irrespective of its detection capabilities.

Retrieval Evaluation: To evaluate our retrieval performance, we provide a textual query, in our case, the name of the ground truth classes from the OoD datasets, and we evaluate how well our method succeeds in retrieving the matching OoD road obstacle. We use instance-based precision and recall as metrics. A retrieved instance (such as an image crop supposed to contain an OoD road obstacle object) is called true positive if the majority of the pixels within the corresponding image bounding box semantically belong to the query. Consequently, precision is the fraction of retrieved instances that match the query, and recall is the fraction of all instances in the dataset according to the query, which are correctly retrieved. As the retrieval performance depends on a similarity threshold, we report the Precision-Recall Curve for all queries in each dataset.

4.1. Object-level vs Image-level Processing

In the first experiment, we present a comprehensive evaluation of our proposed method for object retrieval. We compare our approach of segmenting, tracking, and embedding cut-outs of OoD road obstacles against the conventional approach of retrieving directly on embeddings of the full im-

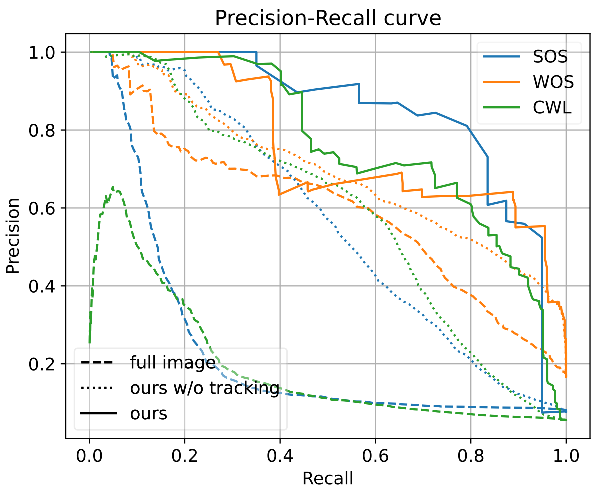

Figure 5. **Precision-Recall curve for each dataset**: Each curve illustrates the trade-off between precision and recall for varying thresholds. Dashed curves represent the baseline approach of retrieving based on full-scale driving scenes. Solid curves represent our method of retrieving based only on a cut-out of the OoD road obstacle with tracking information, and dotted curves represent our method but without tracking information.

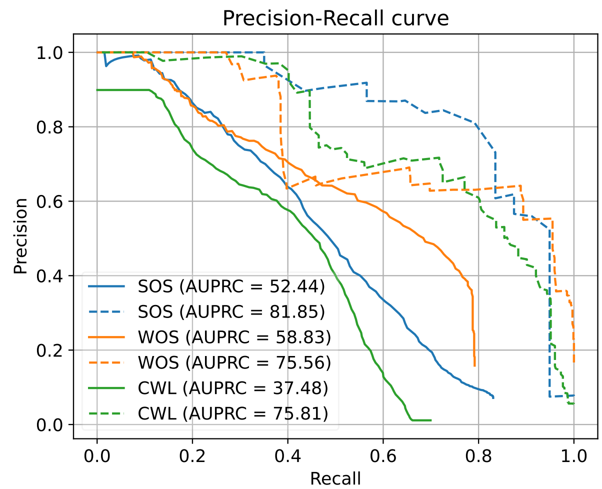

Figure 6. **Average precision-recall curve for each dataset**: The area under the curve (AUPRC) provides a comprehensive measure of the retrieval performance for each dataset. Solid curves represent our method, where we segment and track the video streams, and dashed curves represent the retrieval performance on ground truth detections and tracking.

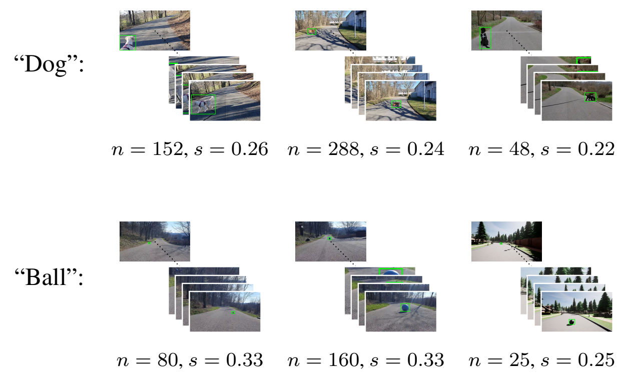

Figure 7. Examples of retrieved video sequences with the corresponding query, sequence length (n), and similarity score (s). From left to right, the images correspond to the first and last frame of the sequence.

age. Our approach is rooted in our observation that object-level information is necessary for retrieving OoD road obstacles in complex driving scenes where OoD road obstacles make up the minority of the full driving scene. Furthermore, we examine the impact of tracking on retrieval performance. For this experiment, we assume optimal conditions where all OoD road obstacles were detected and tracked correctly.

The results for the precision-recall curve for each of the methods and datasets are shown in Figure 5. The results demonstrate our method's significant advantage in performance over the baseline approach. This is primarily attributed to the fact that OoD road obstacles typically only occupy a minor portion of the overall driving scene. Therefore, relying solely on full-frame retrieval leads to inferior performance. The results also show that tracking plays a role in improving the retrieval results. This is because far-away detections are more challenging to retrieve than closer ones. Therefore, creating a link between detections closer to the camera and far away detections via tracking improves the retrieval performance.

4.2. Ablation Study

We conduct an ablation study to understand the contribution and significance of the individual components of our method. In the first experiment, we evaluate the efficacy of our proposed method for the task of OoD road obstacle retrieval. Figure 6 shows our retrieval performance measured by the area under the precision-recall curve for the different datasets compared to the setting with perfect OoD detections. We note that the tracking performance of the proposed lightweight algorithm is almost perfect when evaluated on ground truth detection. Therefore, to achieve better tracking performance, we require either a more robust tracking method that can compensate for the errors in detection or enhance the segmentation model to reduce false positives. We found that, although the performance is reasonable when we assume that all OoD road obstacles are detected, there is still room for improvement. Regarding our OoD segmentation method, despite utilizing a state-of-the-art network, it falls short of capturing all instances of

		Segmentation			Tracking		Retrieval
Dataset	Method	AUPRC ↑	FPR_{95} ↓	$\overline{F}_1$ ↑	MOTA ↑	MOTP ↓	AUPRC ↑
SOS	Entropy max	85.20	1.30	50.40	0.32	12.45	37.53
	RbA	**89.47**	**0.33**	**53.58**	**0.36**	**5.93**	**52.44**
WOS	Entropy max	**94.92**	**0.59**	30.13	0.13	51.17	26.03
	RbA	93.76	0.81	**48.52**	**0.23**	**16.88**	**58.83**
CWL	Entropy max	79.54	1.38	47.64	0.48	18.91	26.33
	RbA	**86.93**	**0.59**	**60.17**	**0.52**	**7.01**	**37.48**

Table 1. OoD object segmentation, tracking, and retrieval results.

	Segmentation	Tracking		Retrieval
Dataset	$\overline{F}_1$ ↑	MOTA ↑	MOTP ↓	AUPRC ↑
SOS	68.94 (+15.36)	0.68 (+0.32)	3.17 (-2.76)	65.01 (+12.57)
WOS	73.85 (+25.33)	0.46 (+0.23)	7.40 (-9.48)	64.59 (+5.76)
CWL	63.89 (+3.72)	0.58 (+0.06)	6.62 (-0.39)	40.87 (+3.39)

Table 2. OoD object segmentation, tracking, and retrieval results under perfect region of interest, with comparative performance gains in comparison to RbA in Table 1.

OoD objects (as indicated by dashed curves not reaching a recall value of one). Table 1 summarizes our evaluation results for OoD segmentation, tracking, and retrieval across three video datasets using different segmentation networks. The table shows a strong correlation between the object-level OoD segmentation network performance ($\overline{F}_1$ score) and the tracking and retrieval performance. This signifies the importance of OoD segmentation for retrieval. Figure 7 shows qualitative examples of retrieved video sequences.

Perfect Regions of Interest: The segmentation model predicts the road as well as OoD obstacles. All OoD obstacles in the non-drivable area are excluded from the final predictions using the predicted road area as the RoI. However, after analyzing our method, we identified a pattern of multiple false positives occurring on the sidewalk. This observation can be attributed to the fact that parts of the sidewalk are often incorrectly predicted as road, resulting in an inaccurate RoI. Therefore, we evaluate our proposed method under perfect RoI masks obtained from the ground truth road and OoD road obstacle segmentation masks. We evaluate the performance gain in segmentation, tracking, and retrieval. Table 2 shows the results of this experiment and highlights the potential additive performance gains.

We note that the improvement is due to the decrease in the number of false positive predictions, which leads to better tracking and, therefore, better retrieval scores. Since the tasks of tracking and retrieval depend on the threshold-selected segmentation of OoD road obstacles, we exclude the threshold-independent pixel-wise OoD scores evaluation metrics (AUPRC and FPR_{95}) from the remainder of the evaluations.

Meta Classification [11, 52] poses an additional but negligible computational overhead to the OoD prediction pipeline that significantly reduces false positives. We evaluate the impact of omitting meta-classification from our pipeline. Table 3 presents our findings on the impact of

	Segmentation	Tracking		Retrieval
Dataset	$\overline{F}_1$ ↑	MOTA ↑	MOTP ↓	AUPRC ↑
SOS	19.68 (-33.90)	-1.57 (-1.93)	30.41 (+24.48)	24.29 (-28.15)
WOS	28.31 (-20.21)	-0.99 (-1.22)	16.66 (+0.22)	32.55 (-26.28)
CWL	22.98 (-37.19)	-0.69 (-1.21)	14.21 (+7.20)	35.95 (-1.53)

Table 3. OoD object segmentation, tracking, and retrieval results without meta classification, with comparative performance loss compared to the RbA method in Table 1.

meta-classification on segmentation, tracking, and retrieval performance. As expected, removing meta-classification reduces the ($\overline{F}_1$ score) due to an increased number of false positives, resulting in a considerable drop in tracking and retrieval performance.

5. Conclusion

This work presents a first approach for the novel task of text-to-video OoD road obstacle retrieval. Our primary aim is to address the question of "*Have we ever encountered this before?*", a critical question arising during the life cycle of AI components in real-world automated driving scenarios. Addressing this question helps advance the development of automated driving systems by enabling them to adapt their driving policies in constantly changing environments. By leveraging single-frame OoD segmentation, object tracking, and multi-modal embedding of OoD road obstacles, our method provides an effective and efficient solution to retrieve relevant video data in response to practical deployment issues. The empirical results showcase the clear advantages of our object-level processing approach over the baseline that relies solely on complete image information. By exploring the retrieval task's dependence on segmentation and tracking, we uncover valuable insights into enhancing performance. Specifically, we note the need for better post-segmentation methods to eliminate false-positive predictions, *i.e.* the prediction of the drivable area as a region of interest for OoD road obstacles and meta-classification for automated segment-wise false-positive removal. We believe this work lays the groundwork for further research into the issue of OoD road obstacle retrieval for a fast response to AI-related safety concerns during deployment. In doing so, our contribution targets resolving real-world challenges arising in the life cycle of data-driven AI components in automated driving perception systems.

Acknowledgment. R. Chan acknowledges financial support by the Ministry of Culture and Science of the German state of North Rhine-Westphalia within the KI-Starter (grant no. 005-2204-0023) research funding program. F. Güney acknowledges financial support from the Royal Society Newton Fund Advanced Fellowship (NAF\R\2202237).

References

[1] Relja Arandjelovic and Andrew Zisserman. Multiple queries for large scale specific object retrieval. In *BMVC*, volume 2, page 6, 2012. 3

[2] Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2574–2583, 2017. 2

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 3

[4] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 6

[5] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9739–9748, 2020. 3

[6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 3

[7] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 6

[8] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6247–6257, 2020. 3

[9] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv preprint arXiv:1905.00737*, 2019. 3

[10] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. 3, 6

[11] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 5128–5137, 2021. 3, 4, 8

[12] Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[13] Wei Chen, Yu Liu, Weiping Wang, Erwin M. Bakker, Theodoros Georgiou, Paul W. Fieguth, Li Liu, and Michael S. Lew. Deep image retrieval: A survey. *ArXiv*, abs/2101.11282, 2021. 3

[14] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 4

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 4

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[17] Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2687–2704, may 2022. 3

[18] Tim Fingscheidt, Hanno Gottschalk, and Sebastian Houben. *Deep neural networks and data for automated driving: Robustness, uncertainty quantification, and insights towards safety*. Springer Nature, 2022. 2

[19] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 3

[20] 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. 3

[21] Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun. 3d traffic scene understanding from movable platforms. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):1012–1025, 2013. 3

[22] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[23] Matej Grcić, Josip Šarić, and Siniša Šegvić. On advantages of mask-level recognition for outlier-aware segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2936–2946, June 2023. 3

[24] Sergio Guadarrama, Erik Rodner, Kate Saenko, Ning Zhang, Ryan Farrell, Jeff Donahue, and Trevor Darrell. Open-vocabulary object retrieval. In *Robotics: science and systems*, volume 2, page 6, 2014. 3

[25] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[26] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4555–4564, 2016. 3

[27] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[28] 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, 2 edition, Dec. 2018. 3

[29] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15425–15434, October 2021. 2

[30] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. Object detection in videos with tubelet proposal networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 727–735, 2017. 3

[31] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2896–2907, 2017. 3

[32] 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. 3

[33] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 3

[34] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 33–40, 2016. 3

[35] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *Pattern Analysis and Machine Intelligence (PAMI)*, 2022. 2

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[37] Krzysztof Lis, Sina Honari, Pascal Fua, and Mathieu Salzmann. Detecting road obstacles by erasing them. *arXiv preprint arXiv:2012.13633*, 2020. 2

[38] Yang Liu, Idil Esen Zulfikar, Jonathon Luiten, Achal Dave, Deva Ramanan, Bastian Leibe, Aljoša Ošep, and Laura Leal-Taixé. Opening up open world tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19045–19055, 2022. 3

[39] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[40] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 1, 3

[41] Kira Maag, Robin Chan, Svenja Uhlemeyer, Kamil Kowol, and Hanno Gottschalk. Two video data sets for tracking and retrieval of out of distribution objects. In *Proceedings of the Asian Conference on Computer Vision*, pages 3776–3794, 2022. 2, 6

[42] Kira Maag, Matthias Rottmann, and Hanno Gottschalk. Time-dynamic estimates of the reliability of deep semantic segmentation networks. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 502–509. IEEE, 2020. 3

[43] Annika Mütze., Matthias Rottmann., and Hanno Gottschalk. Semi-supervised domain adaptation with cyclegan guided by downstream task awareness. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 80–90. INSTICC, SciTePress, 2023. 1

[44] Nazir Nayal, Mısra Yavuz, Joao F Henriques, and Fatma Güney. Rba: Segmenting unknown regions rejected by all. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 4

[45] Philipp Oberdiek, Matthias Rottmann, and Gernot A Fink. Detection and retrieval of out-of-distribution objects in semantic segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, pages 328–329, 2020. 2

[46] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18983–18992, 2023. 1

[47] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1099–1106. IEEE, 2016. 2, 6

[48] 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. 2, 3, 5

[49] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 5

[50] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[51] Matthias Rottmann, Pascal Colling, Thomas Paul Hack, Robin Chan, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020. 3, 4

[52] Matthias Rottmann and Marius Schubert. Uncertainty measures and prediction quality rating for the semantic segmentation of nested multi resolution street scene images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 4, 8

[53] Manuel Schwonberg, Joshua Niemeijer, Jan-Aike Termöhlen, Jörg P. schäfer, Nico M. Schmidt, Hanno Gottschalk, and Tim Fingscheidt. Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving. *IEEE Access*, 11:54296–54336, 2023. 1

[54] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20020–20029, 2022. 1

[55] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1

[56] Pierre Soille and Pierre Soille. Opening and closing. *Morphological image analysis: Principles and applications*, pages 105–137, 2004. 4

[57] Svenja Uhlemeyer, Julian Lienen, Eyke Hüllermeier, and Hanno Gottschalk. Detecting novelties with empty classes, 2023. 2

[58] Svenja Uhlemeyer, Matthias Rottmann, and Hanno Gottschalk. Towards unsupervised open world semantic segmentation. In *Uncertainty in Artificial Intelligence*, pages 1981–1991. PMLR, 2022. 2

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[60] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7942–7951, 2019. 3

[61] Tomáš Vojíř and Jiří Matas. Image-consistent detection of road anomalies as unpredictable patches. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5491–5500, January 2023. 3

[62] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*, 2022. 3

[63] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 3

[64] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pages 720–736. Springer, 2022. 1

[65] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 2

[66] Xiaoping Zhou, Xiangyu Han, Haoran Li, Jia Wang, and Xun Liang. Cross-domain image retrieval: methods and applications. *International Journal of Multimedia Information Retrieval*, 11(3):199–218, 2022. 3

[67] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pages 474–490. Springer, 2020. 3