A Dataset for Semantic Segmentation in the Presence of Unknowns

Zakaria Laskar[*,a], Tomáš Vojíř[*,a], Matej Grcic[*,b], Iaroslav Melekhov[•,c], Shankar Gangisetty[e], Juho Kannala[c,d], Jiri Matas[a], Giorgos Tolias[a], and C.V. Jawahar[e]

[a]VRG, FEE, Czech Technical University in Prague, Czechia
[b]University of Zagreb Faculty of Electrical Engineering and Computing, Croatia
[c]Aalto University, Finland
[e]IIIT Hyderabad, India
[d]University of Oulu, Finland

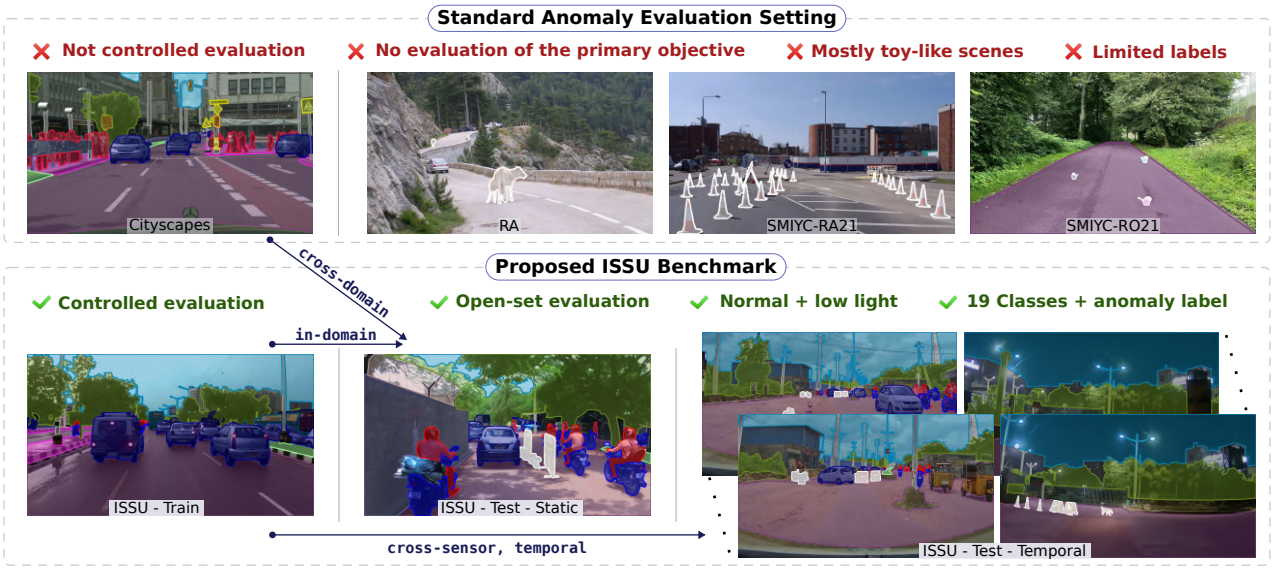

Figure 1. Standard benchmarks cannot separate the effects of domain shift, lighting conditions, and anomaly size during evaluation. The proposed dataset allows controlled evaluation of these effects and supports evaluation of both closed-set and anomaly segmentation.

Abstract

Before deployment in the real-world deep neural networks require thorough evaluation of how they handle both knowns, inputs represented in the training data, and unknowns (anomalies). This is especially important for scene understanding tasks with safety critical applications, such as in autonomous driving. Existing datasets allow evaluation of only knowns or unknowns - but not both, which is required to establish "in the wild" suitability of deep neural network models. To bridge this gap, we propose a novel anomaly segmentation dataset, ISSU, that features a diverse set of anomaly inputs from cluttered real-world environments. The dataset is twice larger than existing anomaly segmentation datasets, and provides a training, validation and test set for controlled in-domain evaluation. The test set consists of a static and temporal part, with the latter comprised of videos. The dataset provides annotations for both closed-set (knowns) and anomalies, enabling closed-set and open-set evaluation. The dataset covers diverse conditions, such as domain and cross-sensor shift, illumination variation and allows ablation of anomaly detection methods with respect to these variations. Evaluation results of current state-of-the-art methods confirm the need for improvements especially in domain-generalization, small and large object segmentation. The code and the dataset are available at https://github.com/vojirt/benchmark_issu.

1. Introduction

Many successful computer vision applications rely heavily or entirely on deep neural networks trained on extensive, fully or partially labeled training data [6, 9, 16, 32]. The

* Equal Contribution. The corresponding author *zakaria.nits@gmail.com*
• The work was done prior to joining Amazon

validation part of the training process provides performance estimates for situations well covered in the training data.

However, when exposed to data not well represented in training, predictions of a deep neural network model may become arbitrary. Therefore, a crucial capability for these models is the ability to detect such *unknown* or *anomalous inputs*[1]. There are multiple reasons why an input may be "unknown" – it might be a rare case, belonging to the long-tail missed for statistical reasons, a result of domain shifts such as introduction of a novel classes (*e.g.*, a segway on a highway) or a result of optical sensor defects (*e.g.*, a broken or dirty lens in a surveillance camera).

Deep neural networks, which lack the ability to recognize unknowns, assign to these anomalous inputs a label that corresponds to one of the known classes of the training set, potentially with high confidence [13]. This may result in suboptimal or even dangerous behavior in deployed systems. Thus, the importance of anomaly detection is critical in safety-sensitive applications, such as autonomous driving, where an undetected anomaly could lead to accidents.

Autonomous driving is a very complex task, with one of its core elements being the perception of the environment surrounding the vehicle, often referred to as scene understanding. Scene understanding is typically defined as a closed-set semantic segmentation task, where each pixel in an image is assigned to one of K known classes. Progress in this area has been greatly advanced by large semantic segmentation datasets [6, 20, 26, 30] along with the development of powerful deep learning models [4, 5] specifically designed for semantic segmentation. However, these datasets overlook the anomaly detection problem. Neither anomalous data nor evaluation protocols are provided with their test sets, limiting the ability to evaluate segmentation models in real-world settings where unknowns may occur.

To address these limitations, specialized datasets focusing on anomaly detection in driving scenarios have been developed, including LostAndFound [22], Fishyscapes [1], RoadAnomaly [17], and SMIYC [3]. However, these datasets typically use a binary evaluation approach ("known" *vs*. "anomaly"), where all pixels belonging to closed-set classes are assigned a single "known" label, while unknowns are labeled as "anomaly". This approach diverges from open set $K+1$ evaluation ("closed-set" *vs*. "anomaly"), which is essential for real-world applications. Moreover, these datasets lack in-domain training data and are often collected under controlled conditions - usually in clear daylight and in simplified environments such as empty roads or parking areas. This setup leads to limited scene diversity and the absence of clutter from other traffic actors.

In this paper, we introduce ISSU, a fully annotated semantic segmentation dataset that provides *both* closed-set and anomaly labels for images in the test set. This allows for the joint evaluation of closed-set and open-set semantic segmentation, with the anomaly label forming an additional class. Our dataset comprises real-world images collected from roads in India, which, due to its unstructured traffic conditions, present a wide range of anomalies in diverse sizes, shapes, lighting conditions, and complex backgrounds cluttered with on-road traffic agents. ISSU consists of three parts: ISSU-Train, ISSU-Test-Static and ISSU-Test-Temporal. ISSU-Train includes training and validation sets, while ISSU-Test-Static forms the test split for controlled in-domain evaluation (*i.e.*, with train and test data from the same distribution). ISSU-Test-Temporal contains temporal test data in the form of short video clips collected using a different sensor setup than ISSU-Train. Semantic annotations with anomaly labels, along with the specific design of the dataset, allow controlled evaluations that isolate the effects of various nuisance factors, such as different anomaly sizes, lighting conditions, and camera sensors. Beyond standard evaluations, our dataset enables cross-domain evaluations, facilitating an analysis of how models trained on datasets from structured environments, such as Cityscapes [6], generalize to unstructured settings, and vice versa.

Dataset-Year	Size (annotated)	Weather/Env. Cond.	Location	Clutter
AppoloScape'18 [15]	145k	Diverse	China	High
Mapillary Vistas'17 [20]	25k	Diverse	World	Diverse
BDD100K'20 [30]	10k	Diverse	US	Low
IDD'19 [26]	10k	Good	India	High
Cityscapes'16 [6]	5k	Good	Europe	Low
WildDash 2'22 [31]	4.3k	Diverse	World	Diverse
ACDC'21 [24]	4k	Adverse	Europe	Low
ISSU-Train'24	3.4k	Diverse	India	High

Table 1. **Comparison of existing datasets for semantic segmentation** for driving scenarios.

The contributions are as follows.

1. We introduce the first real-world segmentation dataset with both closed-set and anomaly labels with defined static and temporal test splits.
2. We present a comprehensive evaluation of the best performing state-of-the-art anomaly segmentation models, based on the standard and most commonly used SMIYC benchmark leaderboard[2], covering approaches from pixel-based to mask-based methods.
3. We provide in-depth analysis of how in-domain, cross-domain, cross-sensor, lighting variations, and anomaly size affect performance. Our results indicate that current methods struggle under these challenging conditions, highlighting the need for further research.
4. The proposed ISSU-Test-Temporal, which consists of short video clips, opens up new directions for future research - particularly in test-time adaptation of anomaly segmentation models in real-world.

[1] We use these term interchangeably.

[2] https://segmentmeifyoucan.com/leaderboard

Dataset-Year	Domain	Size	Anom. Size	Modality	%Anom. Pixels	%Non-Anom. Pixels	Classes	Weather/Env. Cond.	Clutter	oIoU
Street-hazards'22 [14]	Synthetic	1500	Diverse	Static	1.00	98.90	13	Day	Low	✓
Fishyscapes-static'21 [1]	Hybrid	1000	Diverse	Static	2.10	85.80	2	Diverse	Low	✗
LostAndFound'16 [22]	Real	1000	Small	Static	0.12	39.10	2	Day	None	✗
RoadAnomaly'19 [17]	Real	60	Diverse	Static	9.85	33.16	2	Day	High	✗
Fishyscapes-LaF'21 [1]	Real	275	Small	Static	0.23	81.13	2	Day	None	✗
SOS'22 [18]	Real	1129	Diverse	Temporal	0.21	23.30	2	Day	None	✗
WOS'22 [18]	Real	938	Diverse	Temporal	0.88	41.80	2	Day	None	✗
SMIYC-RoadAnomaly'21 [3]	Real	100	Diverse	Static	13.80	82.20	2	Day	Low	✗
SMIYC-RoadObstacle'21 [3]	Real	327	Small	Temporal†	0.12	39.10	2	Diverse	None	✗
ISSU-Test-Static'24	Real	980	Diverse	Static	2.18	89.60	20	Diverse	High	✓
ISSU-Test-Temporal'24	Real	1140	Diverse	Temporal	1.20	85.60	20	Diverse	High	✓

Table 2. **Comparison of existing datasets for anomaly detection** in driving scenarios. Datasets are compared in terms of dataset properties (*Domain, Size, Modality*, number of *Classes*), anomaly statistics (*Anom*aly size, *%Anom*aly and *Non-Anom*aly pixels), diversity of conditions (*Weather/Environment*, *Clutter*) and support for open-set evaluation (*oIoU*). † indicates low frame-per-second in sequences. The *void* class is not considered in the class count reported in the table.

2. Related Work

Semantic segmentation driving datasets aggregate images from the driver's front view and label them into the 19 most relevant classes to driving tasks (such as road, curb, pedestrian, *etc*.), as originally proposed in [6]. Some datasets include additional class labels tailored to the specific locations where the data was collected, such as a "tricycle" class in the China region [15]. However, all of them adhere to the basic 19 classes for compatibility reasons. More recent datasets focus on increasing task difficulty by capturing scenes on a larger scale [15, 20], incorporating unstructured traffic environments [26], or including adverse driving conditions [7, 24, 30, 31].

Despite these advancements, all datasets ignore pixels outside of the predefined training classes, and their evaluation protocols assess only closed-set performance, *i.e*., performance on the classes specified during training. This lack of annotation for unknown objects in test sets and the closed-set evaluation methodology limit their ability to validate models in realistic scenarios involving unknown objects. In contrast, ISSU allows the evaluation of semantic perception models in the presence of unknowns by providing labels that include an unknown class, thus supporting an open-set evaluation. The statistics of the commonly used driving semantic segmentation datasets are shown in Tab. 1.

Anomalies in road-driving scenes. Limited evaluation of standard semantic segmentation road-driving datasets gave rise to specialized datasets that benchmark the detection of unknowns as a standalone task [1, 3]. The Fishyscapes [1] benchmark evaluates obstacle detection in a subset of the LostAndFound [22] dataset and a subset of Cityscapes *val* injected with synthetic anomalies. The SMIYC [3] benchmark is fully based on real-world images and validates the detection of anomalies on drivable surfaces as well as on the whole images. Several other standalone test datasets were proposed in conjunction with novel methods, such as RoadAnomaly [17] which was later merged into the SMIYC benchmark or synthetic Street-hazards [14] which is not widely used due to large domain shift between not photorealistic synthetic and real-world images. Most recently, WOS and SOS [18] datasets were introduced. These datasets include video sequences but only focus on the evaluation on drivable regions of interest. Unlike all these datasets, ISSU contains labels with 19 known classes and an unknown class which enables evaluation of anomaly detection performance in various regions of interest (such as the whole image, drivable surface only or anything in-between), and joint evaluation of the performance in open-set setting ($K+1$ class evaluation).

Acquiring detailed annotations (*e.g*., 19 known classes and an anomaly class) requires extensive manual effort. Thus, several datasets [2, 14] attempt to simplify the labeling efforts by simulating real-world traffic in synthetic environments. However, the quality of synthetic images diverges from the real-world data, leading to domain shifts that complicate the evaluation. Existing road driving anomaly datasets are summarized in Tab. 2.

3. The Proposed ISSU Dataset

Unstructured driving environments, such as those seen on Indian roads, are challenging for the task of semantic segmentation. The density of on-road and near-road traffic agents such as cars, pedestrians, road-side shops create a cluttered environment as shown in Fig. 2.

3.1. Dataset Composition

We compose our dataset using images collected on Indian roads [21, 25, 26] with new and detailed annotations of known class and anomaly labels (*cf*. Sec. 3.3). The dataset consists of three parts, a training set (ISSU-Train), and two test sets (ISSU-Test-Static and ISSU-Test-Temporal).

Figure 2. **Examples of anomalies (shown in white) in the ISSU dataset.** The anomalous examples are ordered from small (left) to very large (right). Top: examples of anomalies of different size and shape at approximately the same distance from the ego-vehicle in ISSU-Test-Static. Bottom: temporal view of an anomaly observed at different time-steps in ISSU-Test-Temporal.

Training set (ISSU-Train). It consists of images collected from different parts of Indian cities [25, 26]. The training set images contain *only* objects from the known classes.

Static test set (ISSU-Test-Static). It consists of images collected in the same way as the training set, but the test set images contains both known and anomalous objects.

Temporal test set (ISSU-Test-Temporal). It consist of short video clips that are also collected on Indian roads [21]. Images in each clip contains both known and anomalous objects. This set is collected using a consumer grade dashcam [21] which are ubiquitous but may produce lower image quality, *e.g.*, due to firmware issues[3]. On the other hand, training sets ISSU-Train and CityScapes consist of images captured using higher quality cameras.

Challenging examples. We focus on studying the affect of challenging viewing conditions, such as extreme lighting conditions and weather variations. To achieve this, all three aforementioned dataset parts include several images and video clips collected in lowlight or in rainy conditions. Detailed examples are shown in the supplementary. This subset has many challenges, such as light burst from oncoming cars and low visibility. The images collected in rainy conditions contain rain droplets and wiper movements, making the segmentation task even more challenging.

3.2. Training and Evaluation Setups

In-domain Static Evaluation. Training on ISSU-Train and testing on ISSU-Test-Static account for an *in-domain Static evaluation setup*.

Cross-domain Static Evaluation. Training models on the CityScapes dataset, and testing on ISSU-Test-Static forms a *cross-domain Static evaluation setup*. The comparison with the in-domain setup allows us to evaluate the impact of such a domain shift in anomaly segmentation performance.

Cross-sensor Temporal Evaluation. Testing on ISSU-Test-Temporal allows *cross-sensor temporal evaluation* of methods trained on ISSU-Train or CityScapes due to the quality discrepancy between such sensors. Modeling domain shifts from such image corruptions isw an active field of research [12]. We argue that it is necessary to benchmark anomaly segmentation methods in such real-world settings. This setup can be evaluated in an in-domain or cross-domain fashion, *i.e*. training models on the CityScapes / ISSU-Train datasets, and testing on ISSU-Test-Temporal forms a *cross / in-domain Temporal evaluation setup*.

3.3. Annotation

The annotation process is performed to assign three types of labels: i) semantic class labels representing the known set of classes in the training set, ii) anomaly label denoting unknowns, and iii) void label for pixels that should not be taken into account during evaluation.

Labels of known classes. We follow the 19 CityScapes labels to define our known classes, but make adjustments for the context of Indian roads. As Indian roads often have blurry road boundaries, we assign both the road and the nearby drivable region as road class label. Unlike semantic segmentation datasets [26], we exclude traffic cones and short on-road traffic-poles from the known "traffic-sign" class following standard anomaly segmentation datasets [3].

Anomaly label. We conducted a rigorous multi-step annotation process (details are provided supplementary Sec. 9), identifying anomalies as objects outside the known 19

[3]https://dashcamtalk.com/forum/threads/ddpai-mini3-video-quality.37223/

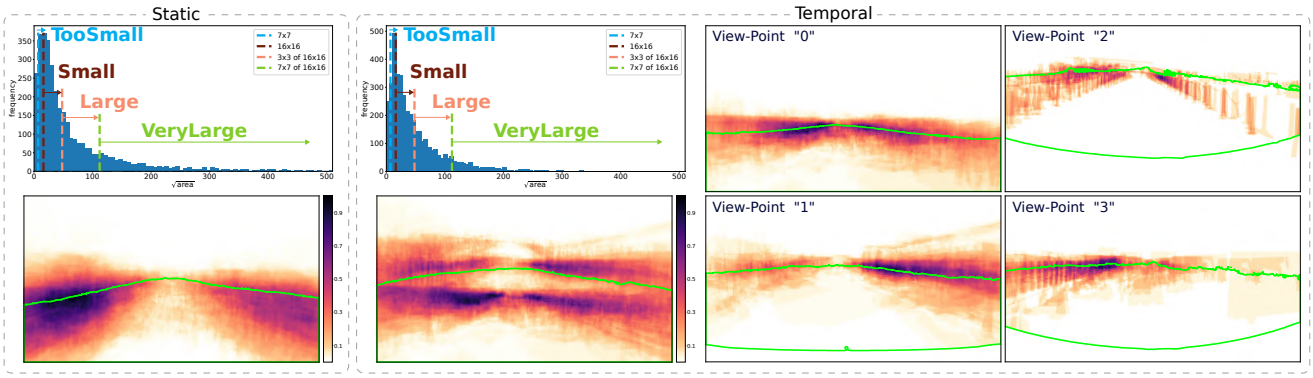

Figure 3. **Distributions of anomalies with respect to their size and spatial location within images.** The anomalies are quantized to four different size intervals that are used in the ablation. Anomalies less than 7×7 (black dashed line) are ignored during all evaluations. The spatial distributions are visualized as a probability heatmap for each image location. Green line outlines road pixels that appeared in more than 50% of dataset images. For temporal dataset the spatial distribution is also visualized for different view-points.

CityScapes classes and within a pre-defined region of interest (ROI) - on or within 2 meters of the road (based on visual inspection). Although this process may not cover all unknowns, it was a deliberate design choice to avoid unknowns outside the region of interest as they are less likely to affect the ego-vehicle. In addition, unknowns within the ROI, but frequently observed (*e.g.* auto-rickshaws, banner) were also not included as anomalies. All such unknowns, but non-anomaly objects, were labeled as void in both train and test sets. Examples of anomaly objects in our datasets include tires, bins, water-tanks, construction material, road barricades, road-maintenance dugouts, animals, road-side vendor items such as fruits *etc*., traffic cones and traffic-poles, pile of stones, mud and sand, tubs, rope, deep potholes. The process involved 7 annotators over a span of 2 months. To ensure that the team of annotators is familiar with the task, they received appropriate training until they achieved 95% accuracy with respect to the CityScapes labels. Examples of anomalies are shown in Fig. 2

Void label. Pixels that are not assigned to the labels of known classes or the anomaly label, according to the aforementioned guidelines, are assigned to the void label.

3.4. Evaluation protocols

We establish four distinct evaluation protocols, each focusing on different aspects of the problem. This is enabled by the proposed ISSU dataset, because it includes images annotated with K known classes along with an anomaly class.

Road obstacles evaluation protocol considers the driving surfaces to be the area of interest for evaluation. Therefore, pixels that are not annotated as *road* or as anomalies are assigned to the void label during this evaluation setup.

Road anomaly evaluation protocol benchmarks performance across all non-void pixels, *i.e*., pixels labeled as any of $K + 1$ classes. In contrast to the previous protocol, this one also accounts for errors occurring outside the driving regions, which increases the task difficulty.

Open-set evaluation protocol validates the recognition of known classes in the presence of anomalies. This protocol penalizes *both* misclassifications among known classes and incorrect detection of anomalies.

Closed-set evaluation protocol assesses classification performance solely on the K known classes and maps the anomaly label to the void label during this evaluation protocol. This standard evaluation protocol estimates the capabilities of trained models in an ideal setting and can serve as an upper bound for open-set evaluation performance.

3.5. Metrics

Average precision (AP) quantifies anomaly detection performance by measuring the area under the precision-recall curve. This threshold-free metric is used to evaluate performance in road obstacle and road anomaly protocols.

FPR$_T$ measures the false positive rate for the threshold that yields the true positive rate of 95%. This is particularly important in safety-critical applications that demand high sensitivity and recall of all anomalous objects.

TPR$_F$ measures the true positive rate for the threshold that yields the false positive rate of 5%. This metric forms a complementary operation point to the previous one and targets applications that require high precision, *i.e*. low false positive detection.

F1 score [3] combines the anomaly detection metrics (recall and precision) and extends them from the pixel level to the component level. By grouping connected neighboring anomalous pixels into cohesive components, this approach provides instance-level performance estimates.

Intersection-over-Union (IoU) quantifies recognition performance by measuring the overlap between predicted and ground-truth segments. Since traffic scene segmentation is a multiclass classification task, we report the macro-averaged IoU over the classes of interest. We use IoU to assess the

	Method	OOD Data	Static						Temporal					
			Road Anomaly			Closed & Open-set			Road Anomaly			Closed & Open-set		
			AP ↑	FPR_T ↓	TPR_F ↑	IoU ↑	$oIoU_T$ ↑	$oIoU_F$ ↑	AP ↑	FPR_T ↓	TPR_F ↑	IoU ↑	$oIoU_T$ ↑	$oIoU_F$ ↑
In-domain														
pixel-level	JSR-Net†	✗	4.2	56.1	3.8	56.8	9.8	44.1	2.3	58.7	2.5	37.3	6.3	29.2
	DaCUP†	✗	5.4	100	20.4	57.0	9.0	47.0	2.9	100	11.1	37.3	6.9	30.4
	PixOOD	✗	20.3	39.4	50.9	65.8	47.8	60.9	6.2	56.5	26.2	55.1	32.9	52.2
mask-level	RbA	✗	75.7	73.4	93.6	73.1	36.4	66.4	36.5	94.9	75.2	57.8	5.5	54.3
	EAM	✗	77.1	5.9	94.4	73.4	66.5	67.4	45.2	92.9	81.8	59.1	6.1	55.5
	Pebal	✗	69.9	9.2	93.3	73.1	64.2	67.2	32.4	92.6	74.9	57.8	7.8	55.4
	RbA	✓	79.1	3.9	95.9	72.9	67.8	66.5	37.7	29.4	76.6	57.8	41.7	55.6
	EAM	✓	76.8	4.2	96.1	73.8	68.4	67.6	38.7	91.4	84.4	59.5	6.7	55.7
	Pebal	✓	64.5	4.4	95.7	72.9	67.8	67.1	23.6	24.7	77.1	57.8	46.2	55.7
	UNO	✓	71.4	3.0	96.9	73.7	68.4	65.9	30.4	89.7	84.8	59.0	9.8	55.2
	M2A	✓	32.0	66.9	71.1	53.9	31.5	49.5	10.7	78.6	47.5	40.3	16.9	34.1
Cross-domain														
	PixOOD	✗	11.4	73.7	33.2	56.3	20.4	52.8	4.8	80.7	25.5	48.7	14.7	46.9
	RbA	✗	43.3	97.3	70.5	57.2	4.1	55.2	15.7	98.5	46.2	41.3	1.1	40.6
	RbA	✓	56.4	80.7	78.9	57.5	11.9	55.1	24.6	91.6	54.4	43.7	3.2	41.9
	UNO	✓	55.5	92.9	79.1	68.1	12.0	65.6	37.2	92.4	70.3	57.4	6.6	54.6

Table 3. Results for road anomaly, closed-set and open-set evaluation protocols under in-domain and cross-domain evaluation setups. The T (F) subscript for oIoU metric refers to operating point (anomaly score threshold) for which the methods achieves 95% TPR (5% FPR).

performance in closed-set evaluation.

Open-Intersection-over-Union (oIoU) [10] evaluates recognition performance of known classes in the presence of anomalous instances. Unlike the standard IoU metric, oIoU incorporates false positives and false negatives committed by the anomaly detector. *The difference between IoU and oIoU highlights the performance gap between closed-set and open-set deployments.*

3.6. Statistics

The train set of ISSU-Train comprises 3436 images, while the validation set comprises 762 images. The test set ISSU-Test-Static contains 980 annotated images. ISSU-Test-Temporal, which consists of video clips, includes a total of 21118 images, of which 1140 are annotated. The unannotated images are released to facilitate future research in using temporal images for online test-time adaptation of anomaly segmentation models to mitigate the challenges of domain shifts. The number of pixels (log-scale) per class in ISSU-Train, ISSU-Test-Static, ISSU-Test-Temporal is shown in Fig. 6 of the supplementary. As can be seen, the distribution of pixel counts per class is similar between train and test splits. Additionally, Tab. 6 in supplementary provides statistics on the number of images captured under normal daylight and adverse lowlight conditions across different ISSU splits. The frequency histogram of anomaly sizes is shown in Fig. 3, illustrating significant variations in anomaly sizes. Examples of images showing these variations are presented in Fig. 2. The spatial distribution of the anomalies, along with the approximate road regions, is shown in the bottom left of Fig. 3 for both ISSU-Test-Static and ISSU-Test-Temporal. The plot shows that the anomalies are distributed across various regions of the road.

	Method	OOD Data	Static		Temporal	
			AP ↑	FPR_T ↓	AP ↑	FPR_T ↓
In-domain						
pixel-level	JSR-Net†	✗	85.7	8.4	52.1	26.5
	DaCUP†	✗	85.5	100	56.8	100
	PixOOD	✗	93.1	4.3	83.1	10.1
mask-level	RbA	✗	92.7	77.5	53.4	98.9
	EAM	✗	94.5	2.2	70.0	98.1
	Pebal	✗	92.3	3.4	54.2	95.6
	RbA	✓	95.8	1.7	57.2	33.7
	EAM	✓	95.6	1.6	62.1	96.2
	Pebal	✓	92.5	1.9	48.9	23.8
	UNO	✓	94.0	1.2	56.1	92.3
	M2A	✓	48.9	78.5	30.0	79.5
Cross-domain						
	PixOOD	✗	92.3	5.1	84.3	10.8
	RbA	✗	62.4	99.1	32.5	99.3
	RbA	✓	76.1	68.9	37.9	87.9
	UNO	✓	66.3	90.8	49.1	90.5

Table 4. Results for road obstacle evaluation protocols under in-domain and cross-domain setups.

4. Baselines

The baselines were selected as the top performing methods on the standard and the most frequently used SMIYC [3] benchmark. We broadly categorized them into two groups

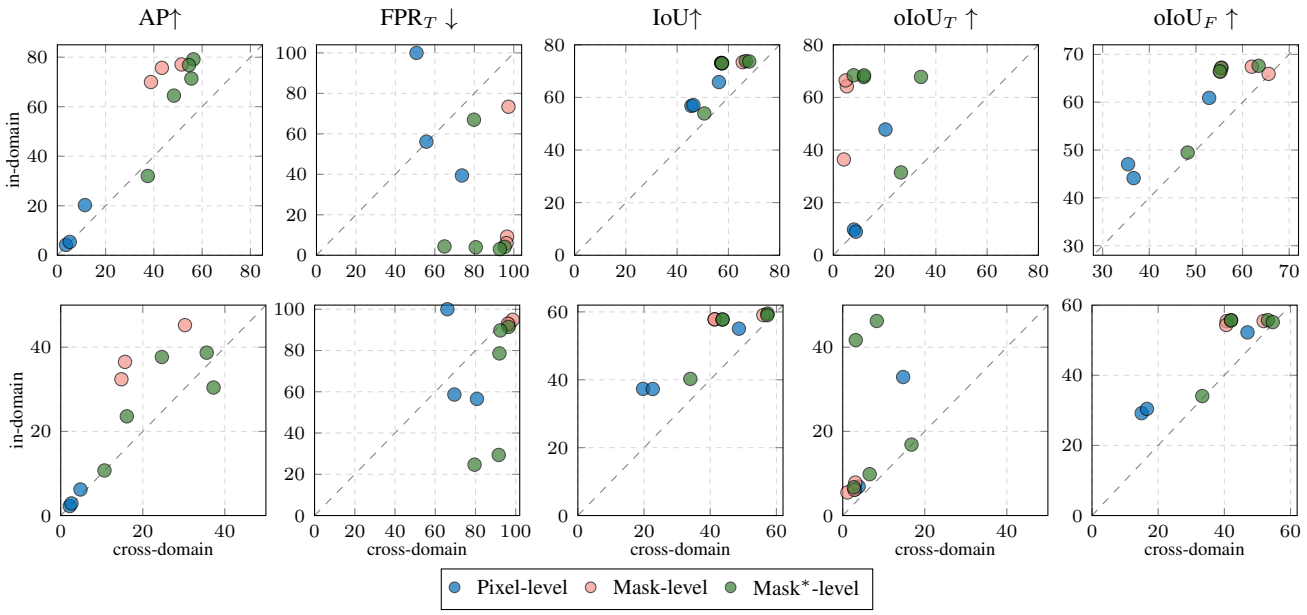

Figure 4. **Cross-domain *vs*. In-domain performance in road anomaly evaluation protocol**. Top row – Static, bottom row – Temporal. Mask* are mask-based methods trained with OOD data. The $y = x$ reference line shows relative gain or drop. The T (F) subscript for oIoU metric refers to operating point (anomaly score threshold) for which the methods achieves 95% TPR (5% FPR).

based on the granularity of regions for which an anomaly score is predicted, *i.e.*, pixel-level and mask-level. The methods are briefly described in the following paragraphs.

Pixel-level baselines. We consider two reconstruction-based methods, JSR-Net [27] and DaCUP [28] that localizes anomalies as poorly reconstructed pixels. We also include recent PixOOD [29] that uses a statistical decision strategy in pre-trained representation to detect anomalies.

Mask-level baselines. We consider baselines that extend the mask-level classifier [5]. A seminal mask-level approach EAM [11] assigns anomaly scores to masks instead of pixels and aggregates decisions to recover dense predictions. RbA [19] considers regions rejected by all masks as anomalous, while Mask2Anomaly (M2A) [23] adapts the model architecture to enhance anomaly detection. Finally, UNO [8] revisits the $K+1$ classifier built on top of the mask classifier and combines negative class recognition with prediction uncertainty to improve anomaly detection.

5. Experimental results

Road Anomaly Evaluation results are presented in Tab. 3 and Tab. 7 in supplementary. For the in-domain Static evaluation setup, most Mask2Former (mask-level) methods trained with auxiliary out-of-domain (OOD) data achieve good performance across three anomaly detection metrics: FPR$_T$ ($< 5\%$), TPR$_F$ ($> 90\%$) and AP ($> 70\%$). Due to the lack of any "objectness" priors, the pixel-level methods classify many random pixels as anomalous with high confidence, resulting in a poor anomaly detection metrics.

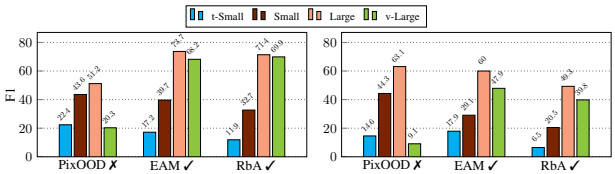

Figure 5. **Ablation of different anomaly sizes.** The plot shows results in ISSU-Test-Static (left) and ISSU-Test-Temporal (right) for road anomaly evaluation protocol under in-domain setup.

In contrast, results on the challenging in-domain Temporal setup show that both pixel-level and mask-level methods have high FPR$_T$ and low AP. This indicates that domain shifts due to differences in sensor quality adversely affect anomaly detection performance.

Results in cross-domain Static and cross-domain Temporal setups shows a significant performance drop for all methods compared to the in-domain setup (*cf*. Fig. 4) resulting in low AP and high FPR$_T$. The complementary TPR$_F$ metric shows that some mask-based methods such as UNO could detect 79% and 70% of anomalies for cross-domain Static and Temporal setups respectively. Successfully detecting the remaining anomalies, to achieve 95% TPR, results in high FPR$_T$ ($> 90\%$). This is because the methods include many known-class pixels as true-positives to correctly classify the hard anomalous cases. Qualitative examples of hard anomalies are presented in Sec. 7.3 (Figs. 7 and 8) and Sec. 11 (Figs. 13 to 15) in supplementary.

Closed and Open-set Evaluation (Tab. 3). The IoU metric

	Method	OOD Data	Day						Lowlight					
			ISSU-Test-Static			ISSU-Test-Temporal			ISSU-Test-Static			ISSU-Test-Temporal		
			AP ↑	FPR_T ↓	$\overline{F1}$ ↑	AP ↑	FPR_T ↓	$\overline{F1}$ ↑	AP ↑	FPR_T ↓	$\overline{F1}$ ↑	AP ↑	FPR_T ↓	$\overline{F1}$ ↑
pixel-level	JSR-Net	✗	4.26	57.45	1.84	2.28	75.69	0.83	4.20	39.94	0.61	2.28	75.69	0.83
	DaCUP	✗	5.09	100.00	1.67	2.90	100.00	2.25	7.80	38.80	3.13	2.80	100.00	2.78
	PixOOD	✗	34.24	32.46	2.28	16.07	53.57	1.50	5.63	60.59	0.65	2.16	65.76	0.56
mask-level	RbA	✗	77.06	5.59	16.34	39.82	95.77	11.38	67.99	97.44	9.28	24.16	96.20	6.81
	EAM	✗	77.72	5.09	21.34	47.48	95.57	15.06	73.92	95.58	14.61	37.68	89.63	11.50
	Pebal	✗	71.27	6.37	21.60	34.95	97.45	11.84	61.61	89.31	9.78	21.70	87.91	7.18
	RbA	✓	79.64	3.44	22.14	40.75	23.03	12.00	75.40	74.38	12.30	28.56	74.54	8.64
	EAM	✓	77.26	3.63	22.21	39.62	94.27	15.28	74.47	14.65	14.26	37.81	84.35	12.06
	Pebal	✓	65.39	3.71	0.00	28.55	22.19	0.00	58.92	16.92	0.00	12.20	28.80	0.00
	UNO	✓	71.71	2.82	29.19	31.78	41.77	18.60	74.47	14.65	14.26	30.10	88.17	14.58
	Mask2Anomaly	✓	35.13	61.02	9.45	11.71	77.05	5.61	19.97	91.35	6.17	7.78	86.29	4.32

Table 5. **Ablation of different lighting conditions.** Results for in-domain road anomaly evaluation protocol under day and light-adverse conditions (night, rain, fog, dawn).

for ISSU is for all methods about 10% lower than respective IoU achieved in CityScapes considering in-domain Static setup (75.88% vs. 65.83% for PixOOD and $83.5 - 83.7\%$ vs. $\sim 73\%$ for most Mask2former methods). This difference is significantly higher (20-30%) for in-domain Temporal and cross-domain setups. This highlights the difficulty of the cluttered traffic environment in India and challenging domain shifts. Open-set IoU (oIoU) follows similar conclusions as road anomaly evaluation protocols. Results on in-domain Temporal and cross-domain setups show significant difference between closed set IoU (IoU) and open-set IoU at 95% TPR ($oIoU_T$). This is attributed to misdetection of anomalies as known classes and known classes as anomalies. The drop is less significant between IoU and $oIoU_F$ but results in significantly lower TPR_F. It is to be noted, zero drop (IoU = oIoU) can be achieved by not detecting any anomalies (TPR/FPR = 0%). Thus it is important to analyze both TPR_F (FPR_T) and $oIoU_F$ ($oIoU_T$). The open-set results signify the importance of evaluating semantic segmentation in real-world setting by jointly evaluating closed-set segmentation in the presence of anomalous object.

Road Obstacle Evaluation (Tab. 4). When the evaluation is limited to road regions, the pixel-level methods generally are much better at generalizing from CityScapes to the ISSU, resulting in a much lower FPR metric for both static and temporal datasets. However, for in-domain Static evaluation setup the mask-level methods are able to outperform other method, mainly due to strong priors baked in object-wise mask predictions that seems to be more robust to detecting entire anomaly instances. In the temporal part where the different sensors act as a form of a domain-shift the Mask2former based methods struggle to localize all anomalies resulting in high FPR. The effects of domain shift are less pronounced in this setup due to the uniformity of roads, as shown in Tab. 8.

Ablations: anomaly sizes. The effect of anomaly sizes is shown in the Fig. 5, where specific anomaly size ranges (defined in Fig. 3) are considered. The size intervals were motivated by the dataset statistics and spatial resolution of the most commonly used backbone architectures. We use the F1 metric, which is designed to measure instance-level performance. The metric is generally improved with larger anomaly sizes, except for the largest anomalies, where the methods struggle to accurately and fully segment the very large instances. This is again more apparent for pixel-level methods. Consistently with the evaluation limited to road region, the temporal dataset with the additional challenges of different sensors negatively effects Mask2former based methods significantly more than the pixel-level methods. The results for all methods are in the supplementary Fig. 9.

Ablations: lighting variations. This ablation compares the performance of the methods under lighting variations - day (clear weather and good lighting conditions) and lowlight (*e.g.* fog, rain, dawn). The results are presented in Tab. 5 which shows that both the pixel-based and mask-based methods struggle under lowlight conditions.

6. Conclusions

We presented a new dataset and a benchmark for anomaly segmentation in a real-world setting. The results show that cross-domain generalization remains a challenge for current state-of-the-art anomaly segmentation methods. When trained on in-domain data, the performance of these models improves by a significant margin. This forms a strong baseline for future work in cross-domain generalization and adaptation of anomaly segmentation models. The results also showed that existing methods struggle in the presence of lower sensor quality, lower visibility, and small anomaly size. The diverse conditions provided by our benchmark offer a timely test-bed for anomaly segmentation research.

Acknowledgments

The authors acknowledge support from respective sources as follows. Zakaria Laskar: Programme Johannes Amos Comenius (no. CZ.02.01.01/00/22 010/0003405). Giorgos Tolias: Junior Star GACR (no. GM 21-28830M). Tomáš Vojíř and Jiri Matas: Toyota Motor Europe and by the Czech Science Foundation grant 25-15993S. Shankar Gangisetty and C.V. Jawahar: iHub-Data and Mobility at IIIT Hyderabad. Matej Grcic: Croatian Recovery and Resilience Fund - NextGenerationEU (grant C1.4 R5-I2.01.0001). Iaroslav Melekhov and Juho Kannala: Research Council of Finland (grants 352788, 353138, 362407), Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Allice Wallenberg Foundation.

We thank Ram Sharma from CVIT, IIIT Hyderabad, Mahender Reddy and the annotation team in Annotations and Data Capture-Mobility, IIIT Hyderabad for their work in the annotation process.

References

[1] Hermann Blum, Paul-Edouard Sarlin, Juan I. Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *IJCV*, 2021. 2, 3, 6, 10

[2] Daniel Bogdoll, Iramm Hamdard, Lukas Namgyu Rößler, Felix Geisler, Muhammed Bayram, Felix Wang, Jan Imhof, Miguel de Campos, Anushervon Tabarov, Yitian Yang, Hanno Gottschalk, and J. Marius Zöllner. Anovox: A benchmark for multimodal anomaly detection in autonomous driving. *CoRR*, abs/2405.07865, 2024. 3

[3] 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 *NeurIPS Dataset and Benchmarks*, 2021. 2, 3, 4, 5, 6, 10

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2

[5] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 7

[6] 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 *CVPR*, 2016. 1, 2, 3

[7] Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *IJCV*, 2020. 3

[8] Anja Delić, Matej Grcic, and Siniša Šegvić. Outlier detection by ensembling uncertainty with negative objectness. In *BMVC*, 2024. 7, 3, 6

[9] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2010. 1

[10] Matej Grcic and Sinisa Segvic. Hybrid open-set segmentation with synthetic negative data. *IEEE TPAMI*, 2024. 6

[11] Matej Grcic, Josip Saric, and Sinisa Segvic. On advantages of mask-level recognition for outlier-aware segmentation. In *CVPR*, 2023. 7, 3

[12] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 4

[13] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 2

[14] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *Int. Conf. on Mach. Learn.* PMLR, 2022. 3, 10

[15] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The ApolloScape Open Dataset for Autonomous Driving and Its Application. In *IEEE TPAMI*, pages 2702–2719, 2020. 2, 3

[16] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset V4. *IJCV*, 2020. 1

[17] Krzysztof Lis, Krishna Kanth Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *ICCV*, 2019. 2, 3

[18] 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 *ACCV*, 2022. 3, 10

[19] Nazir Nayal, Misra Yavuz, João F. Henriques, and Fatma Güney. Rba: Segmenting unknown regions rejected by all. In *ICCV*, 2023. 7, 3

[20] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 2, 3

[21] Chirag Parikh, Rohit Saluja, C. V. Jawahar, and Ravi Kiran Sarvadevabhatla. IDD-X: A multi-view dataset for ego-relative important object localization and explanation in dense and unstructured traffic. In *ICRA*, pages 14815–14821. IEEE, 2024. 3, 4, 5

[22] 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 *Int. Conf. on Intelligent Robots and Systems*, 2016. 2, 3, 10

[23] Shyam Nandan Rai, Fabio Cermelli, Dario Fontanel, Carlo Masone, and Barbara Caputo. Unmasking anomalies in road-scene segmentation. In *ICCV*, 2023. 7, 4

[24] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: the adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021. 2, 3

[25] Furqan Ahmed Shaik, Abhishek Reddy Malreddy, Nikhil Reddy Billa, Kunal Chaudhary, Sunny Manchanda, and Girish Varma. IDD-AW: A benchmark for safe and robust segmentation of drive scenes in unstructured traffic and adverse weather. In *WACV*, 2024. 3, 4, 5

[26] Girish Varma, Anbumani Subramanian, Anoop M. Namboodiri, Manmohan Chandraker, and C. V. Jawahar. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *Winter Conf. Appl. of Comput. Vis.*, 2019. 2, 3, 4

[27] Tomas Vojir, Tomáš Šipka, Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino, and Jiri Matas. Road Anomaly Detection by Partial Image Reconstruction With Segmentation Coupling. In *ICCV*, pages 15651–15660, 2021. 7, 2

[28] Tomáš Vojíř and Jiří Matas. Image-Consistent Detection of Road Anomalies As Unpredictable Patches. In *WACV*, pages 5491–5500, 2023. 7, 2

[29] Tomáš Vojíř, Jan Šochman, and Jiří Matas. PixOOD: Pixel-Level Out-of-Distribution Detection. In *ECCV*, 2024. 7, 2, 6

[30] 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 *CVPR*, 2020. 2, 3

[31] Oliver Zendel, Matthias Schörghuber, Bernhard Rainer, Markus Murschitz, and Csaba Beleznai. Unifying panoptic segmentation for autonomous driving. In *CVPR*, pages 21351–21360, 2022. 2, 3

[32] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *IJCV*, 2019. 1