A Dataset for Semantic Segmentation in the Presence of Unknowns
Supplementary Material

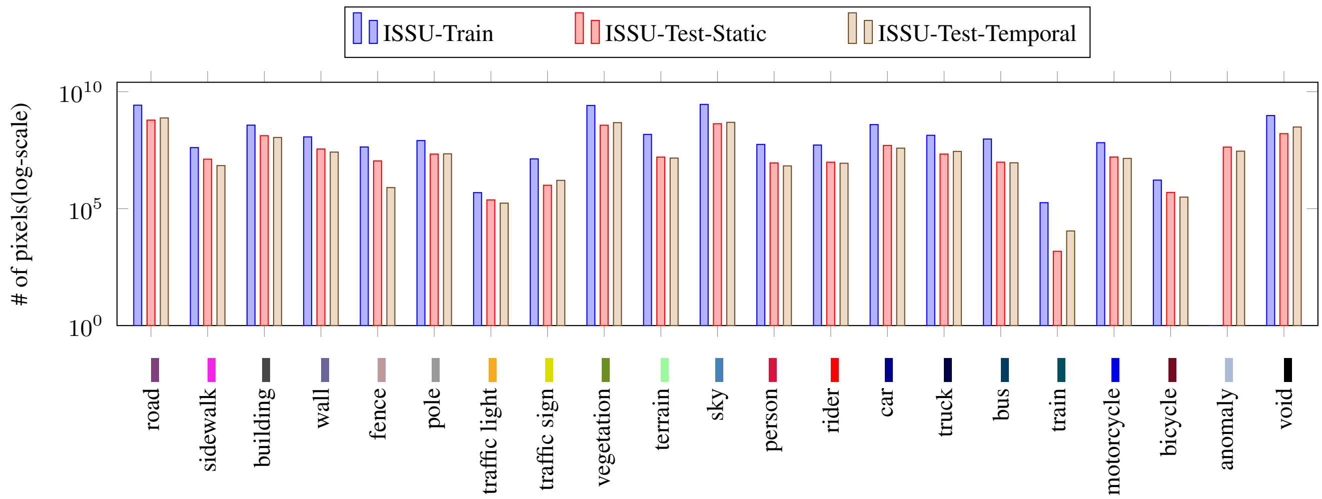


Figure 6. **Dataset statistics**. The number of annotated pixels per class and their associated class labels for each part (ISSU-Train, ISSU-Test-Static, and ISSU-Test-Temporal) of the proposed dataset.

In the supplementary, we provide additional results in Sec. 7, implementation details of benchmarked methods in Sec. 8, dataset composition process in Sec. 9 and comparison with existing anomaly segmentation datasets in Sec. 10.

7. Additional Results

In this section, we provide extended results and additional analysis. Section 7.1 shows the train and test statistics, Section 7.2 presents the results of the cross-domain evaluation, while Secs. 7.3 and 11 provides qualitative examples of undetected anomalies which are the primary contributors to the high FPR metric. Furthermore, Secs. 7.4 and 7.5 include additional ablation studies and detailed analyses.

7.1. Statistics

The number of pixels (log-scale) per class in ISSU-Train, ISSU-Test-Static, ISSU-Test-Temporal is shown in Fig. 6. As can be seen, the distribution of pixel counts per class is similar between train and test splits. Additionally, Tab. 6 provides statistics on the number of normal and adverse images across different ISSU splits.

7.2. Cross-domain Results

Complete results for the cross-domain evaluation, *i.e.*, training on CityScapes and evaluating on the proposed ISSU, are provided in Tab. 7 and Tab. 8 for the road anomaly and road obstacle evaluation protocols, respectively.

The effects of cross-domain evaluation are less pronounced for the road obstacle evaluation, *i.e.*, where only the road region and anomalies are considered due to the high visual similarity of road regions across domains. In this setting, pixel-level methods demonstrate better robustness.

Dataset	Day	Lowlight
ISSU-Train	2690	746
ISSU-Test-Static	848	132
ISSU-Test-Temporal	868	270

Table 6. **Statistics of Day and Lowlight across the train and test splits of the proposed dataset**.

7.3. Qualitative Results

To analyze the high FPR metric, particularly for Mask2Former-based methods, we conducted a visual analysis of the results from the RbA method (representative of Mask2Former-based approaches) in Fig. 7. By setting the anomaly score threshold such that the TPR metric reaches 95%, we observe several examples of fully (or partially) undetected anomalous instances. This behavior leads to a high FPR at this operating point, as the method includes many known-class pixels to correctly classify the "hard" anomalous cases.

We considered cross-sensor (in-domain Temporal) and cross-domain setups, and qualitatively compared two methods: PixOOD and RbA (✓) in Fig. 8. The results are shown for very large and small anomalies with TP, FN and FP

	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 ↑
pixel-level	JSR-Net†	✗	3.60	55.71	5.06	45.57	8.07	36.64	2.21	69.51	5.10	19.70	3.03	15.02
	DaCUP†	✗	5.16	50.69	16.35	46.35	8.81	35.45	2.61	66.03	13.88	22.63	3.87	16.61
	PixOOD	✗	11.44	73.73	33.19	56.30	20.36	52.84	4.81	80.74	25.53	48.67	14.72	46.99
mask-level	RbA	✗	43.31	97.30	70.47	57.17	4.12	55.24	15.66	98.46	46.21	41.33	1.15	40.56
	EAM	✗	51.49	96.32	68.83	65.58	4.82	61.98	30.28	96.12	53.51	56.04	2.86	51.87
	Pebal	✗	38.80	96.62	71.09	57.17	5.29	55.51	14.79	96.84	46.86	41.33	3.07	40.69
	RbA	✓	56.39	80.75	78.98	57.50	11.88	55.12	24.64	91.56	54.40	43.72	3.18	41.97
	EAM	✓	54.54	95.40	71.74	66.80	7.94	63.44	35.57	96.42	61.97	57.33	2.72	53.16
	Pebal	✓	48.32	64.88	79.66	57.50	34.20	55.36	16.11	79.54	55.01	43.72	8.31	42.07
	UNO	✓	55.54	92.96	79.15	68.11	12.03	65.58	37.24	92.37	70.35	57.24	6.56	54.63
	M2A	✓	37.48	79.82	69.00	50.59	26.40	48.23	10.66	91.92	33.16	33.99	16.76	33.33

Table 7. **Cross-domain evaluation of road anomaly, closed-set and open-set**.

	Method	OOD Data	Static		Temporal	
			AP ↑	FPR_T ↓	AP ↑	FPR_T ↓
pixel-level	JSR-Net†	✗	80.70	11.91	25.45	41.63
	DaCUP†	✗	85.95	9.23	69.52	20.42
	PixOOD	✗	92.30	5.10	84.34	10.84
mask-level	RbA	✗	62.40	99.11	32.48	99.28
	EAM	✗	57.96	93.83	37.15	95.44
	Pebal	✗	62.85	98.08	34.21	97.97
	RbA	✓	76.14	68.89	37.86	87.93
	EAM	✓	61.35	93.44	43.03	98.26
	Pebal	✓	73.58	40.79	29.36	67.09
	UNO	✓	66.25	90.81	49.10	90.50
	M2A	✓	63.29	45.84	30.74	81.35

Table 8. **Cross-domain evaluation of road obstacle**

pixels colored accordingly. These qualitative visualizations support the findings in Fig. 9 – pixel-level PixOOD struggles in detecting very large anomalies while being better than mask-level RbA in detecting small anomaly objects. The cross-sensor and cross-domain shift is challenging for both methods as shown by the known classes misclassified as anomalies (FP pixels).

7.4. Ablation: Anomaly Size

Figure 9 presents an ablation study of the performance of all methods with respect to different anomaly sizes. The findings, consistent across all methods, align with the results presented in the main paper (*cf*. Fig. 5).

7.5. Ablation: Effect of Anomaly Sizes to Metrics

The component-level F1 metric was introduced by Chan *et al*. [3] to account for small-sized anomalies. Correlation plot in Fig. 10 between pixel-level metric, AP and component-level F1, shows that both these metrics are highly correlated. We hypothesize this is due to the diversity of anomaly size in our dataset. Detailed component-level metrics - F1, sIoU and PPV are provided in Tab. 9 for completeness following common practice [3].

In order to show the correlation between the F1 and AP metrics in the proposed dataset, we fit a regression line that minimizes the total squared difference (SSR) between the observed data points (x_i, y_i) and the predicted values $y_{\text{pred},i}$, *i.e*., $y_{\text{pred}} = mx + c$, where m is the slop, and c is the intercept. The correlation coefficient R^2 measures how well the regression line explains the variability of the data. The R^2 value is defined as:

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} \tag{1}$$

where SST is a total sum of squares that measures the variability in the data relative to the mean, *i.e*., $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$; the residual sum of squares, SSR is a measure of the discrepancy between the actual data points and the values predicted by a regression model. It quantifies the amount of variation in the dependent variable y that the model does not explain, *i.e*., $\text{SSR} = \sum_{i=1}^{n}(y_i - y_{\text{pred},i})^2$

8. Implementation Details

Pixel-level baselines. We implement JSR-Net[4] and DaCUP[5] baselines by extending the publicly available code releases. Both baselines extend the DeepLabV3 segmentation model with specialized plug-in modules for anomaly detection. Thus, we follow the optimization procedure and hyperparameters reported in the original papers [27] and [28]. Similarly, we extend the publicly available code of the PixOOD[6] baseline. This baseline relies on a generic feature extractor, so we use ViT-L trained with DINOv2 as suggested in [29]. Other hyperparameters follow the reported values as well.

[4] https://github.com/vojirt/JSRNet
[5] https://github.com/vojirt/DaCUP
[6] https://github.com/vojirt/PixOOD

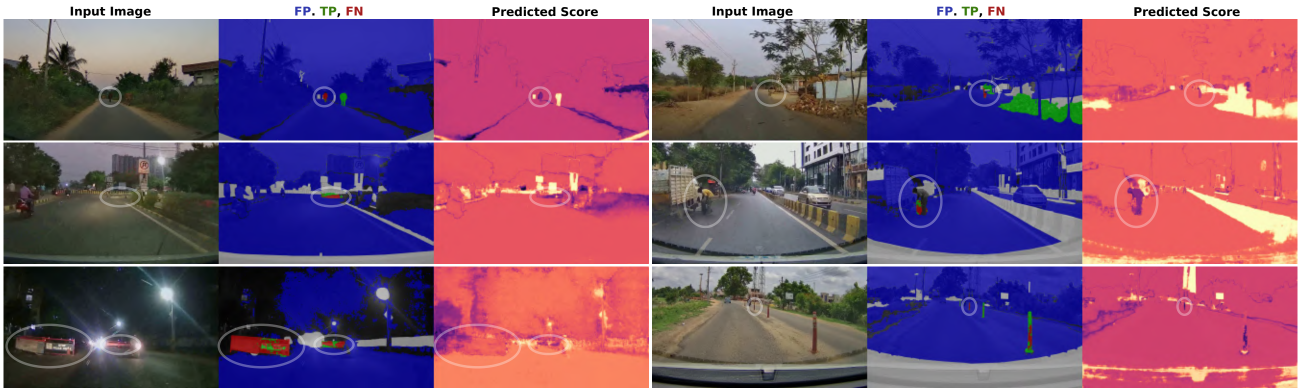
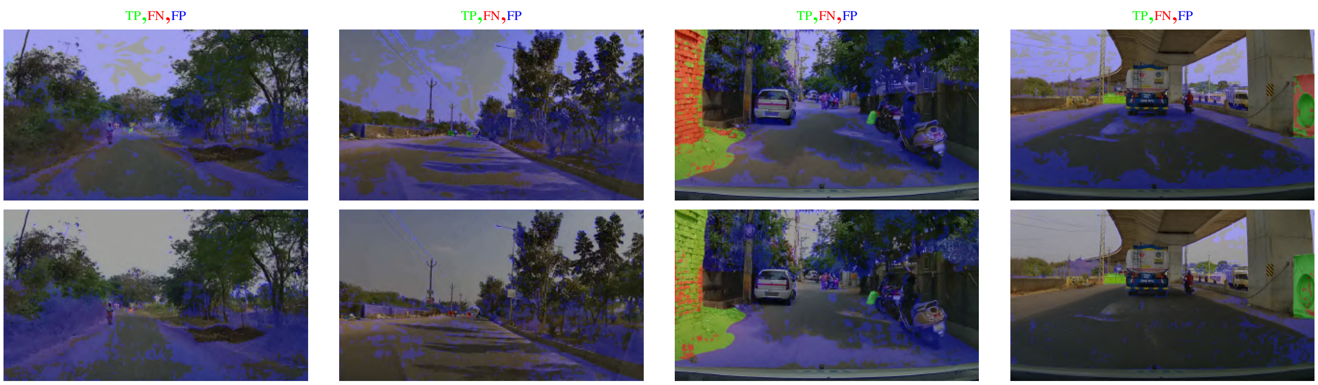


Figure 7. Qualitative results of the RbA(✗) at 95% TPR threshold. The figure shows examples of anomalies that are not detected (fully or partially) at this threshold where most of the image pixels are falsely labeled as anomalies, resulting in very high FPR at 95% TPR metric. The pixel classifications at the 95% TPR threshold are coded by color overlay in the middle images – false positive (blue), true positive (green), false negative (red), void (white) and true negative (without overlay).



(a) in-domain Temporal, small anomaly

(b) in-domain Temporal, very large anomaly

(a) cross-domain, small anomaly

(b) cross-domain, very large anomaly

Figure 8. **Qualitative Results** shown for PixOOD (1^{st} and 3^{rd} row) and RbA (✓) (2^{nd} and 4^{th} row) across in-domain Temporal (cross-sensor) and cross-domain setups for small and very large anomalies. Anomaly detection threshold is set based on operation point 95%TPR

Mask-level baselines. All mask-level baselines extend the Mask2Former architecture with anomaly detection capabilities. In the case of EAM and UNO[7] we use the default Mask2Former upsampling and SWIN-L backbone pretrained on ImageNet-22k, as suggested in the corresponding manuscripts [8, 11]. In the case of the RbA[8] baseline, we use SWIN-B and a single transformer decoder layer. This architecture was validated as optimal for RbA [19]. We use the same architecture when adapting the pixel-level baseline PEBAL to mask-level predictions. Finally, we use a

[7] https://github.com/matejgrcic/Open-set-M2F

[8] https://github.com/NazirNayal8/RbA

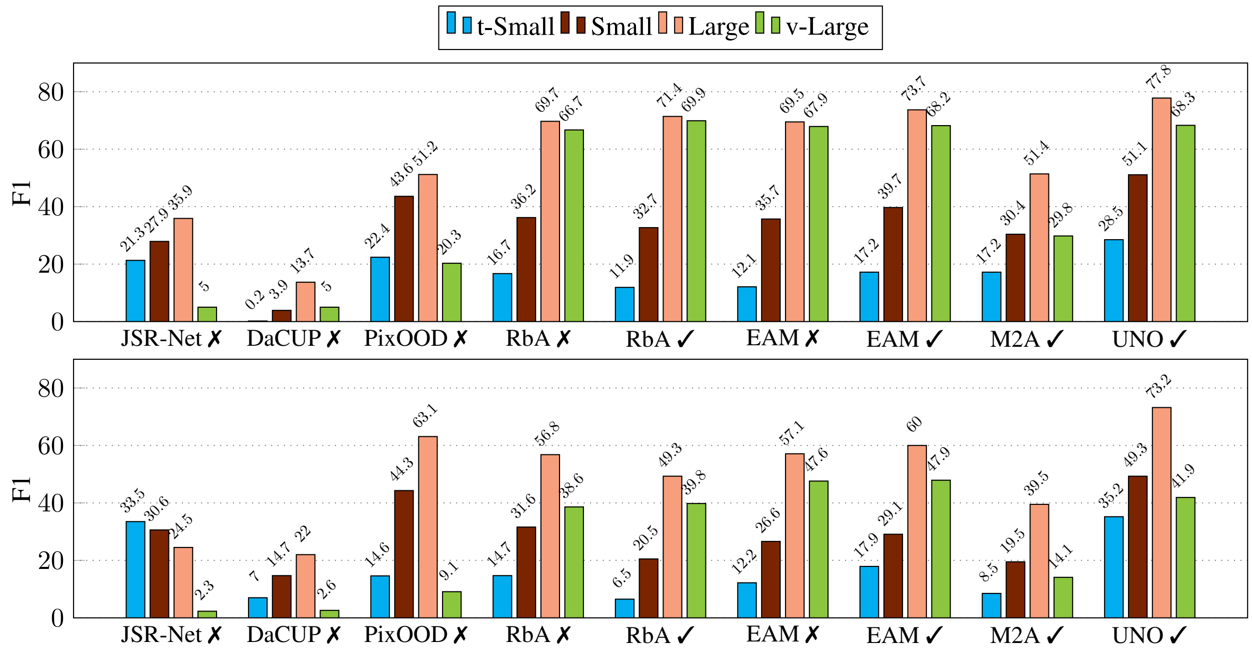


Figure 9. **Ablation for different anomaly sizes**. Top (bottom) plot shows results for ISSU-Test-Static (ISSU-Test-Temporal), respectively. The different anomaly sizes are defined in Fig. 3. The corresponding tick (✓/ ✗) defines trained with / without OOD data.

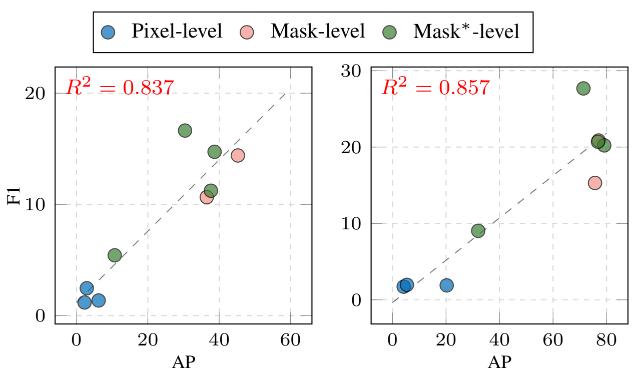


Figure 10. **Correlation of AP-F1**. We fit a regression line and report the correlation coefficient R^2 between the F1 and AP metrics. The correlation coefficient is defined as $R^2 = 1 - \text{SSR}/\text{SST}$ (*cf*. Sec. 7.5) showing how well the regression line explains the variability of the data. The reported values (0.837 and 0.857) indicate a strong correlation for both datasets.

frozen ResNet-50 feature extractor pretrained on ImageNet for the Mask2Anomaly baseline. Again, this backbone was validated as optimal for Mask2Anomaly [23]. We use the default hyperparameter values reported in the corresponding manuscripts for all baselines.

9. Dataset Composition

ISSU-Train and ISSU-Test-Static are composed from the train and validation sets of IDD [26] which already has semantic segmentation annotations as per level-4 IDD label hierarchy that consists of 30 classes. We mapped these classes to CityScapes (C), anomaly (A) and void (V) classes as shown in Tab. 10. Certain IDD classes are mapped to multiple classes, however, the mapping is such that an input pixel can only map to one of the 3 classes (C / A / V) making the assignment unique.

The main requirement for the annotation is to ensure only the test set ISSU-Test-Static contains anomalies. This is done by first identifying the anomaly objects in IDD and creating two subsets: one that does not include any of the listed anomaly objects forming ISSU-Train and the remaining subset forms ISSU-Test-Static. To identify anomaly objects, we asked the annotators to find images with objects in IDD classes that are mapped to A, lies within 2 meters of the road and likely to cause damage or alter the trajectory of a vehicle. A list of such objects are mentioned in Sec. 3.3 and shown in Fig. 11. The shortlisted images with anomaly objects are used to form ISSU-Test-Static and the remaining subset constitutes ISSU-Train. Objects in A that are outside 2 meters of the road, or unlikely to adversely affect a vehicle, are annotated as void. Similarly, objects in ”traffic-sign” IDD class are mapped to both C and A. Objects that are mapped to A consist of traffic cones and traffic poles that are considered anomalies in existing anomaly

Road Anomaly								
	Method	OOD Data	Static			Temporal		
			F1 ↑	sIoU ↑	PPV ↑	F1 ↑	sIoU ↑	PPV ↑
pixel-level	JSRNet†	✗	3.2 / 1.7	13.8 / 14.6	8.2 / 3.2	1.2 / 1.2	12.0 / 13.5	4.4 / 2.6
	DaCUP†	✗	1.2 / 2.0	8.7 / 7.0	7.5 / 6.6	0.9 / 2.5	4.3 / 8.3	5.4 / 6.1
	PixOOD	✗	1.8 / 1.9	15.6 / 27.5	13.4 / 7.6	1.4 / 1.4	14.1 / 24.7	7.8 / 3.7
mask-level	RbA	✗	11.2 / 15.3	28.5 / 36.7	19.9 / 18.2	5.7 / 10.7	17.5 / 25.6	12.4 / 17.8
	EAM	✗	20.2 / 20.9	29.3 / 35.8	23.2 / 23.2	11.7 / 14.4	19.1 / 25.4	18.0 / 20.3
	Pebal	✗	11.8 / 17.6	27.2 / 34.2	21.7 / 23.0	6.3 / 11.3	14.1 / 23.2	17.2 / 20.5
	RbA	✓	9.6 / 20.2	33.2 / 36.9	15.5 / 25.7	5.3 / 11.2	20.0 / 21.6	11.9 / 23.9
	EAM	✓	21.5 / 20.7	30.4 / 39.1	25.2 / 23.0	10.6 / 14.7	26.0 / 27.8	13.5 / 20.2
	Pebal	✓	13.0 / 0.0	29.4 / 0.0	25.2 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
	UNO	✓	27.8 / 27.7	27.8 / 44.3	43.3 / 29.1	18.6 / 16.6	22.8 / 37.9	28.4 / 17.8
	M2A	✓	10.8 / 9.0	27.3 / 25.5	18.3 / 17.5	4.4 / 5.4	8.8 / 16.1	15.3 / 15.4
Road Obstacle								
	Method	OOD Data	Static			Temporal		
			F1 ↑	sIoU ↑	PPV ↑	mF1 ↑	sIoU ↑	PPV ↑
pixel-level	JSRNet†	✗	31.2 / 24.3	55.4 / 62.4	33.2 / 23.8	11.5 / 18.6	25.4 / 45.1	28.1 / 26.6
	DaCUP†	✗	28.1 / 28.2	62.8 / 53.0	22.3 / 24.5	31.0 / 24.1	47.7 / 41.6	32.7 / 25.5
	PixOOD	✗	28.0 / 27.9	58.9 / 64.6	27.3 / 22.9	33.6 / 28.2	50.5 / 53.7	35.8 / 27.6
mask-level	RbA	✗	25.6 / 29.5	37.7 / 51.7	38.6 / 30.1	13.7 / 22.5	26.0 / 41.8	25.1 / 27.4
	EAM	✗	36.4 / 32.7	31.7 / 55.5	51.8 / 28.4	19.3 / 27.5	28.8 / 43.3	27.7 / 28.2
	Pebal	✗	25.6 / 30.8	37.7 / 51.1	38.7 / 32.4	13.7 / 22.5	26.6 / 42.5	25.0 / 27.2
	RbA	✓	17.5 / 37.0	40.9 / 50.9	24.6 / 40.1	9.3 / 23.5	26.6 / 39.1	18.0 / 31.9
	EAM	✓	36.3 / 41.4	35.5 / 56.7	47.9 / 38.7	23.6 / 28.0	30.8 / 48.4	32.4 / 26.8
	Pebal	✓	19.0 / 37.4	40.3 / 50.5	26.8 / 41.1	10.8 / 25.5	23.1 / 35.3	22.8 / 38.6
	UNO	✓	38.5 / 41.5	29.0 / 62.1	67.9 / 35.0	26.2 / 31.1	31.9 / 52.2	34.6 / 28.7
	M2A	✓	24.5 / 22.4	36.9 / 39.3	35.0 / 31.0	12.4 / 15.4	21.9 / 25.6	25.3 / 31.2

Table 9. Component-level metrics for road anomaly (top) and obstacle (bottom) tracks in the form **cross-domain/in-domain**.

segmentation datasets [3].

ISSU-Test-Temporal is composed using videos from IDD-X [21]. From the original 1140 videos, we selected a subset of 103 videos that depicted the anomaly objects present in ISSU-Test-Static. The particular clip showing the anomaly object is cropped and will be released as part of ISSU-Test-Temporal to facilitate methods to utilize temporal information. The clip is chosen in a way such that first and last frame in the clip observes the relevant anomaly. The average clip length is 8.5 seconds at a frame rate of 25 FPS resulting in around 21K images. For each clip, we selected around 10 frames for anomaly and closed-set label annotation. The frame selection is done in a way to ensure the anomaly is approximately observed at uniform temporal and spatial resolutions with respect to the ego-vehicle. The selected subset of frames are annotated into one of the 3 classes (C / A / V).

To include images with challenging lighting conditions, we expanded ISSU-Train and ISSU-Test-Static with images from IDD-AW [25]. The images in IDD-AW are also annotated as per level-4 IDD label hierarchy and consists of images collected in adverse weather conditions such as fog, rain, lowlight, snow. We excluded images collected in snow conditions due to the absence of anomaly objects. Similarly, ISSU-Test-Temporal also consists of rain and lowlight images present in the original IDD-X dataset. The number of such images with challenging lighting variations is listed in Tab. 6 and example images shown in Fig. 12.

10. Dataset comparison

In Tab. 11, we compare ISSU-Test-Static and ISSU-Test-Temporal with existing datasets based on the best performance achieved by any method on the respective datasets. The results indicate that for both evaluation protocols - road obstacle (RO) and road anomaly (RA), ISSU-Test-Static is comparably challenging to existing datasets. However, ISSU-Test-Temporal proves to be significantly more

Figure 11. **Qualitative results**. Example images with anomaly objects from ISSU-Test-Static (first 4 rows) and ISSU-Test-Temporal (bottom 4 rows).

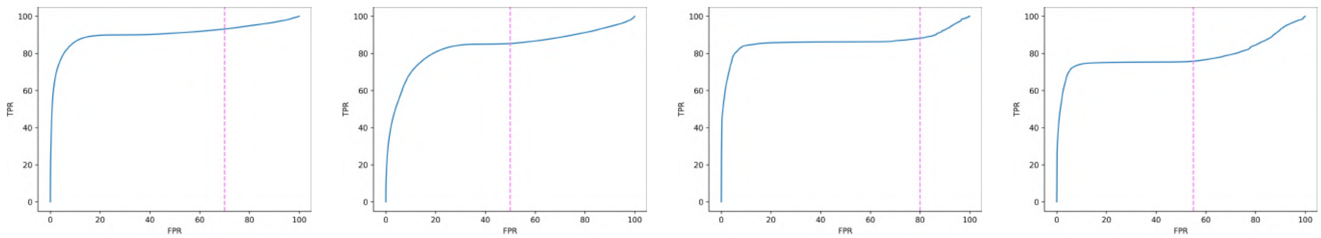
difficult, showing a notable gap in the best performance achieved.

The best values obtained for the metrics (F1 / AP / FPR) on the challenging SMIYC-RA'21 [3] , FSL&F [1] are (60.9 / 94.5 / 4.1), (- / 74.8 / 2.7) [8, 29] respectively. In comparison, the corresponding values on ISSU-Test-Static and ISSU-Test-Temporal are (27.7 / 79.2 / 3.0) and (18.5 / 45.2 / 24.7). Given in-domain training data, ISSU-Test-Static is as challenging as FSL&F while being significantly diverse (*cf*. Tab. 2). ISSU-Test-Temporal is much more challenging. A detailed comparison with other datasets is provided in Supplementary.

Figure 12. **Qualitative results**. Example images with anomaly objects in challenging lighting conditions.

11. Additional Qualitative Results

We provide additional examples of failure cases for RbA (✓) and UNO (✓) in this section. First, we plot ROC curves of both methods in cross-domain Static and Temporal setups in Fig. 13. Across both setups, these methods attain a TPR of 80% at FPR $\leq$ 15%, beyond which the TPR deos not improve until a certain critical operating point is reached (indicated by vertical line in Fig. 13). Examples of anomalies detected beyond this critical operating point are presented in Fig. 14 and Fig. 15.



(a) RbA (✓) Static (b) RbA (✓) Temporal (c) UNO (✓) Static (d) UNO (✓) Temporal

Figure 13. **ROC curves** shown for RbA (✓) and UNO (✓) across **cross-domain** setup on Static and Temporal splits. X-axis: FPR_T, Y-axis: TPR_F. Anomalies not detected until the critical point indicated by vertical line are shown in 14 and 15.

(a) RbA (✓) Static

(a) RbA (✓) Temporal

Figure 14. **Cross-domain qualitative results** of RbA (✓) in (a) Static and (b) Temporal splits. Anomaly detection threshold is set based on Fig. 13 (a) and (b).

(a) UNO (✓) Static

(a) UNO (✓) Temporal

Figure 15. **Cross-domain qualitative results** of UNO (✓) in (a) Static and (b) Temporal splits. Anomaly detection threshold is set based on Fig. 13 (c) and (d).

Class	Mapping		
	CityScapes (C)	Anomaly (A)	Void (V)
road	✓		
parking	✓		
drivable fallback	✓		
sidewalk	✓		
non-drivable fallback			✓
person	✓		
animal		✓	
rider	✓		
motorcycle	✓		
bicycle	✓		
auto-rickshaw			✓
car	✓		
truck	✓		
bus	✓		
caravan			✓
vehicle-fallback		✓	✓
curb		✓	✓
wall	✓		
fence	✓		
guard rail		✓	✓
billboard			✓
traffic-sign	✓	✓	
traffic-light	✓		
pole	✓		
obs-str-bar-fallback		✓	✓
building	✓		
bridge			✓
vegetation	✓		
sky	✓		
fallback-background			✓

Table 10. **Dataset annotation protocol**. The mapping between the level-4 label hierarchy of IDD dataset and corresponding CityScapes (C), Anomaly (A), and Void (V) labels in our proposed datasets is indicated by the ✓ tick.

Datasets	Eval	F1↑	AP↑	FPR↓	oIoU$_T$ ↑
LostAndFound'16 [22]	RO	61.7	89.2	0.6	N/A
SOS'22 [18]	RO	53.6	89.5	0.3	N/A
WOS'22 [18]	RO	48.5	93.8	0.8	N/A
SMIYC-RoadObstacle'21 [3]	RO	75.0	95.1	0.1	N/A
Street-hazards'22 [14]	RA	N/A	58.1	13.0	59.8
Fishyscapes-static'21 [1]	RA	N/A	96.8	0.3	N/A
Fishyscapes-LaF'21 [1]	RA	N/A	74.8	1.3	N/A
SMIYC-RoadAnomaly'21 [3]	RA	60.9	94.5	4.1	N/A
ISSU-Test-Static'24	RO	41.5	95.8	1.2	N/A
ISSU-Test-Temporal'24	RO	31.1	83.1	10.1	N/A
ISSU-Test-Static'24	RA	27.7	79.2	3.0	68.4
ISSU-Test-Temporal'24	RA	18.5	45.2	24.7	46.2

Table 11. **The datasets performance comparison**. For different evaluation protocols - road obstacle (RO) and road anomaly (RA), best values obtained by any method across different metrics: F1, AP, FPR at 95%TPR (FPR), open-IoU at 95%TPR (oIoU$_T$) are presented.

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