Segment Every Out-of-Distribution Object

Supplementary Material

6. Supplementary

We propose the first prompt-based OoD detection method. Our core idea has two main aspects: **1)** Generating prompts directed at OoD objects using information from the anomaly score map and **2)** employing a prompt-based segmentation model to provide accurate masks for OoD objects. In this phase, the segmentation model, refined through prompts, accurately identifies and segments out OoD objects, enhancing the overall detection accuracy and efficiency. Together, these novel steps demonstrate exceptional performance in the field of OoD detection, offering a new perspective for the identification of OoD objects. In this supplementary, we include more details on the following aspects:

- We present the implementation details of acquiring training data using the OE method in Section 6.1.
- We delineate the specifics of generating OoD object masks in Section 6.2.
- We provide a detailed description of the primary evaluation metrics used in our experiments, elucidating the significance of each metric and the performance of our S2M method across these metrics in Section 6.3.
- We detail the efficiency analysis, demonstrating the operational effectiveness of our approach in Section 6.4.
- We employ FastSAM in stead of the standard SAM in Section 6.5.
- We utilize an entropy-based anomaly score in Section 6.6.
- We present visualizations of some S2M results in Section 6.10.

6.1. Details of Outlier Exposure

During the preparing of training dataset, we use the OE [23, 32] strategy to generate the OoD training images. We use objects in COCO dataset as OoD objects, and use images in Cityscapes dataset as background. We exclude those objects from the COCO that are also included in the Cityscapes. The left column of Fig. 11 shows the generated training image. Then, we use RPL [32] to get the anomaly score on these training images. The anomaly scores of these training images are shown in the middle column. The original anomaly scores, which generally lie between -20 and 10, are not suitable for visualization. For visualization purposes, we have normalized these scores to a scale of 0 to 255 for each image. It should be noted that the training process uses the original anomaly scores, not the normalized ones. Right column show the training label of OoD objects. We generated the smallest bounding boxes based on the masks of the OoD objects, which serve as the training labels. During the training of the prompt generator, we utilize the anomaly scores as inputs and employ the generated boxes as prompts.

6.2. Details of Mask Generation

During the inference, we use the produced box prompts to generate masks of OoD objects. The prompt generator is designed to process anomaly scores as input, thereby generating box prompts that highlight OoD objects. In addition, it concurrently produces confidence scores associated with these prompts. To enable a direct comparison between our S2M method and current mainstream approaches using the same metrics, the corresponding confidence scores of these prompts are assigned to the pixels in the generated masks for the OoD objects. For areas with multiple overlapping masks, the pixel values are assigned based on the lowest confidence score among the box prompts that produced these overlapping masks. We employ this strategy with the intention of lowering the false positive rate. Ultimately, the output of our S2M methods is a map where pixel values ranging from 0 to 1. In this map, a pixel value of 0 indicates ID areas, while any other values correspond to OoD regions.

6.3. Evaluation Metrics

During the experimental process, we employed three evaluation metrics. The first metric, IoU, is used to assess the accuracy of OoD object detection at a specific threshold. However, since IoU does not reflect the robustness of different methods to threshold selection, we introduce the second metric AuIoU. AuIoU provides a comprehensive measure of the model's accuracy in detecting OoD masks across various threshold levels, reflecting the ease of selecting the most suitable threshold. A higher AuIoU score indicates greater ease in selecting the optimal threshold. The third metric, mean F1 score, which takes into account both precision and recall, thus providing a more holistic assessment of the prediction results. Across all the three metrics, the proposed S2M outperforms the state-of-the-art OoD detection methods with a large margin.

IoU is a widely used evaluation metric in semantic segmentation. It is employed to assess the accuracy of the model in detecting OoD objects in comparison with the given labels. In this study, we ensure that for all methods which produce anomaly scores, the reported IoU represents the best IoU achieved by the optimal threshold on the specific dataset. For the proposed S2M, the reported IoU is calculated without the need for a threshold. During the computation pro-

cess, we utilized all produced box prompts, obtaining the IoU by taking the intersection of the masks generated from these prompts.

The average IoU of our S2M method is at least 7.52% higher than the other methods listed in Table 1. This demonstrates that our method not only outperforms mainstream methods but also achieves superior performance without the necessity of a threshold. This result can be visualized in the Fig. 10. S2M achieves the highest IoU on the SMIYC validation dataset without a threshold. This indicates that our S2M method is more suitable for real-world application scenarios.

AuIoU. Area under IoU curve (AuIoU) is calculated by the area under the IoU curve with different thresholds. Here we define th as the threshold. TP_{th}, FP_{th}, FN_{th} represent the pixel numbers of True Positives, False Positives, and False Negatives when the threshold is th. True Positives (TP) are pixels correctly identified as OoD, False Positives (FP) are in-distribution pixels incorrectly identified as OoD, and False Negatives (FN) are OoD pixels that are not identified as such. With the above definitions, AuIoU can be computed as,

$$AuIoU = \frac{1}{n} \sum_{th=th_0}^{th_n} \left(\frac{TP_{th}}{TP_{th} + FP_{th} + FN_{th}} \right) \tag{9}$$

where n is the total number of steps and th_0 is the smallest threshold and th_n is the largest threshold. In our experiments, we fixed the value of n at 100, set th_0 to 0, and incrementally increased it to $th_n = 0.99$ with a step size of 0.01. A straightforward interpretation of AuIoU is the area under the IoU curve as depicted in Fig. 10. A higher AuIoU signifies that the model achieves better overall results across various thresholds, indicating that it is easier to obtain an appropriate threshold for the model. This is important in real-world application scenarios where determining the optimal threshold is inherently challenging.

The average AuIoU of our S2M method, as shown in Table. 1, is 41.37% higher than that of RPL. This suggests that RPL is sensitive to threshold selection. This perspective is also intuitively substantiated by observing the IoU curves in Fig. 10. The IoU curve for RPL shows that only a limited range of thresholds result in an IoU above 50%, suggesting that RPL has a narrow range of thresholds where it can achieve optimal performance. This finding highlights the challenges RPL faces in determining an appropriate threshold for optimal performance, a significant limitation in practical applications where flexibility and adaptability in threshold settings are crucial. In contrast, our S2M method demonstrates superior performance in the accurate detection of OoD objects, working effectively without the need for threshold selection, in contrast to the limitations faced by RPL.

Mean F1. The mean F1 score is calculated as the average of F1 scores obtained at various threshold levels. It is the harmonic mean of precision and recall, used to measure the accuracy and completeness of a model's predictions for the positive class. $Precision_{th}$ represent the precision when threshold is th. With the above definitions, mean F1 can be computed as,

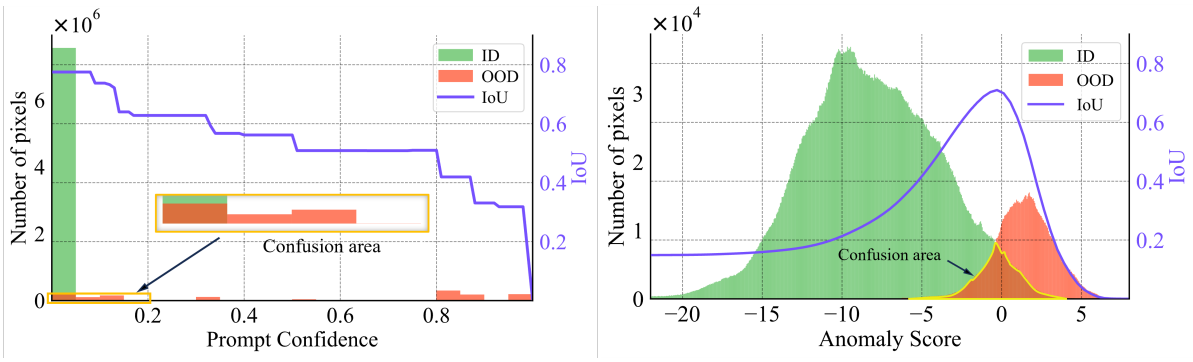
$$\text{mean F1} = \frac{1}{n} \sum_{th=th_0}^{th_n} \left(2 \times \frac{Precision_{th} \times Recall_{th}}{Precision_{th} + Recall_{th}} \right) \tag{10}$$

This metric is especially valuable in scenarios where an optimal threshold has not been pre-established. A high F1 score indicates that the model achieves a favorable balance between precision and recall, suggesting it is proficient in correctly classifying positive cases while minimizing the number of false positives and false negatives. This implies the model's effectiveness in handling cases where both the accuracy of the positive predictions and the completeness of capturing all positive instances are critically important.

The average mean F1 of our S2M method on five datasets in Table 1 is 35.64% higher than Synboost, which shows the best performance in mean F1 among mainstream methods. This indicates that our S2M method excels in balancing precision and recall, particularly in terms of accurately and comprehensively predicting positive classes. Specifically, the higher mean F1 score suggests that the S2M method is more effective in reducing both false positives (incorrectly marking negative instances as positive) and false negatives (missing true positive instances), thereby surpassing other mainstream methods in overall performance. This advantage is crucial as it demonstrates the reliability and accuracy of the S2M method across various application scenarios.

6.4. Details of Efficiency Analysis

When comparing our S2M-B, used in our main experiments, with the RPL method, we observe that the total running time for S2M-B is only 0.059s longer than RPL, a modest increase considering its additional capabilities. The efficiency of S2M-B can be attributed to its dual-component structure which has shown in Fig. 5. Firstly, it includes a mainstream OoD detector that generates an anomaly mask. Secondly, it features the SAM, which utilizes the original image and a box prompt to create precise OoD masks. A significant advantage of this setup is the efficiency in processing time. The operation of SAM on the image can be overlapped with the running time of the RPL, as these two processes can be executed in parallel. Once the box prompts are generated, they can be directly inputted into the decoder, together with the processed original image, to produce the



(a) S2M (b) RPL

Figure 10. **Anomaly score distribution and IoU curve.** We magnify the confusion area as depicted in Figure (a), to provide a clearer and more detailed view.

Method	SMIYC anomaly			SMIYC obstacle		
	AuIoU	IoU	mean F1	AuIoU	IoU	mean F1
RPL+CoroCL	22.26	68.80	31.55	3.85	**28.66**	5.72
S2M (FastSAM-s)	38.55	63.37	41.41	**21.66**	27.47	**25.00**
S2M (FastSAM-x)	**60.58**	**81.09**	**64.18**	20.53	24.08	23.27

Table 5. This table presents the accuracy measurements of S2M utilizes FastSAM on SMIYC. It highlights the S2M method with FastSAM, which utilizes anomaly scores from RPL+CoroCL as inputs. The parameters of FastSAM-x is 68M and FastSAM-s is 11M.

final outcomes. Therefore, our method introduces minimal latency overhead compared to the baseline RPL.

6.5. S2M with FastSAM

As a faster version of SAM that performs comparably, FastSAM [51] can also be used as a promptable segmentation model in our S2M. FastSAM offers two unique model sizes: the compact and swifter FastSAM-s, based on YOLOv8s with an 11M model size, and the more extensive FastSAM-x, based on YOLOv8x with a 68M model size. We leverage FastSAM as the segmentation model and conduct experiments across all datasets using models trained with 2% noise. From Table. 5, we can find that the FastSAM also show an acceptable result on various metrics. S2M with FastSAM-x performs better on SMIYC anomaly validation dataset than RPL, with AuIoU 38.32% higher, IoU 12.29% higher and mean F1 32.63% higher than RPL method. And the S2M with FastSAM-s performs better on SMIYC obstacle validation dataset, with AuIou 17.81% higher and mean F1 19.28% higher than RPL, but IoU 1.19% lower than RPL. Here we use the IoU of RPL with the best performance on the validation dataset. The running time of S2M (FastSAM-s) and S2M (FastSAM-x) is shown in Table. 6. Due to the fast encoder speed and parallel way of segmentation model, the running time of S2M (FastSAM-s) and S2M (FastSAM-x) mainly influenced by the prompting process. However, the performance of FastSAM is lower than SAM with the same input. After visualization we found that SAM shows a stronger robustness to noisy box prompts than FastSAM. That is the reason that S2M with SAM performs better than FastSAM.

Methods	RPL	S2M (FastSAM-s)	S2M (FastSAM-x)
running time (s)	0.2166	0.2415	0.2336

Table 6. In these time measurements, we have excluded the dataloader aspect of the RPL model from our analysis, while including the set image process of the FastSAM model for consideration which run in a parallel way.

6.6. S2M With Entropy Based Anomaly Score

The anomaly scores in our methods, derived from RPL, have been computed using an energy-based approach. To demonstrate the generalization capability of our method, we have also conducted experiments using anomaly scores calculated via an entropy-based method [6]. As previously mentioned, we employ RPL [32] to generate anomaly scores for training images using an entropy-based method, while maintaining all other settings unchanged. Given that the range of entropy-based anomaly scores approximately lies between 0 and 1, we amplify the anomaly score of each pixel by a factor of 20 during training and inference to facilitate the model's ability to distinguish between in-distribution and out-of-distribution pixels. The results of entropy-based anomaly score of RPL and S2M based entropy anomaly score are shown in Table 7. The table demonstrates that S2M, when utilizing anomaly scores calculated via the entropy-based method, also exhibits improved performance compared to using the original anomaly scores.

Methods	FS Static			FS Lost&Found			SMIYC-Anomaly			SMIYC-Obstacle			RoadAnomaly		
	AuIoU	IoU	mean F1	AuIoU	IoU	mean F1	AuIoU	IoU	mean F1	AuIoU	IoU	mean F1	AuIoU	IoU	mean F1
RPL (Entropy)	8.81	14.65	14.61	2.25	3.63	3.83	30.03	47.02	40.96	0.72	1.45	1.35	16.66	26.23	25.10
S2M (Entropy)	**67.48**	**72.18**	**73.81**	**28.19**	**33.17**	**34.86**	**40.11**	**55.67**	**50.80**	**6.08**	**42.17**	**8.25**	**25.14**	**30.87**	**31.41**

Table 7. RPL (Entropy) represent RPL+CoroCL methods with entropy-based anomaly score. S2M (Entropy) represent S2M with anomaly score from RPL (Entropy). The S2M method demonstrates strong **generalization capability**, effectively detecting OoD objects even when processing anomaly scores calculated using the entropy-based method.

Methods	FS Static			FS Lost&Found			SMIYC-Anomaly			SMIYC-Obstacle			RoadAnomaly		
	AuIoU	IoU	mean F1	AuIoU	IoU	mean F1	AuIoU	IoU	mean F1	AuIoU	IoU	mean F1	AuIoU	IoU	mean F1
Mask2Anomaly	6.00	11.60	10.57	0.68	1.57	1.29	44.08	81.31	53.37	8.07	42.41	11.90	24.31	**56.74**	32.80
Mask2Anomaly*	**57.77**	**62.34**	**62.90**	**25.74**	**27.94**	**30.94**	**65.61**	**83.90**	**72.69**	**54.69**	**60.58**	**61.46**	**47.18**	53.39	**52.56**
RbA	3.28	8.07	5.99	0.45	1.52	0.86	28.03	70.28	39.07	2.49	23.18	4.22	16.43	54.99	24.32
RbA*	**50.41**	**57.16**	**56.66**	**24.46**	**27.27**	**29.12**	**33.33**	**76.23**	**41.48**	**41.99**	**52.09**	**48.40**	**44.13**	**55.36**	**51.30**

Table 8. Mask2Anomaly* and RbA* denote the application of our S2M methodology utilizing the anomaly scores from Mask2Anomaly and RbA, respectively. Mask2Anomaly and RbA experiments is conducted with the models provided by authors. The results indicate that our method significantly improves the performance of stronger results.

	SMIYC Anomaly			SMIYC Obstacle		
Method	AuIoU	IoU	meanF1	AuIoU	IoU	meanF1
vanilla RPL	22.26	68.80	31.55	3.85	28.66	5.72
S2M (RPL w. PEBAL's generator)	**59.78**	**73.85**	**68.41**	**13.96**	**29.41**	**20.24**

Table 9. Compare the results of RPL and SAM with prompt generator trained on anomaly score from Pebal.

6.7. S2M with Advanced SOTA Methods

To demonstrate the general improvement capability of our method, we conducted experiments using the anomaly scores from Mask2Anomaly [40] and RbA [37]. These experiments were carried out with the models provided by the authors, applying our S2M based on their anomaly scores. The results, shown in Table 8, indicate that our method consistently enhances the performance.

6.8. Reuse Prompt Generator without Training

Results presented in Table 9 demonstrate that a prompt generator, when trained on PEBAL's [46] anomaly scores and evaluated on RPL' [32], still achieves superior performance compared to RPL. The initial training of PEBAL's generator utilized anomaly scores from PEBAL, which has different domain of anomaly score from RPL. Applying PEBAL's generator directly on RPL's anomaly scores, without any modification, typically yields suboptimal results. In our experiment, we scale up the anomaly score of RPL by a factor of 20. This adjustment contributes to better performance. The result suggests that our prompt generator can be effectively used without the need for retraining.

6.9. Input Contains No OoD Objects

We test our S2M (RPL) method on Cityscapes validation datasets, which comprises 600 images without OoD objects and used as ID dataset during training. The result shows that our prompt generator did not detect any box prompt in all 600 images, indicating that S2M can effectively discern images without OoD objects.

6.10. Visualizations of Segmentation Result

We visualize the OoD mask generated by our S2M methods on Road Anomaly, Fishyscapes and SMIYC in Fig. 12, Fig. 13 and Fig. 14.

Validation on Road Anomaly demonstrates the precision of S2M. Our method accurately detects OoD objects while ensuring that ID objects are not mistakenly identified as OoD, as shown in the first row of Fig. 12. S2M gives a precise mask of horse and excludes the people nearby. S2M is also capable of generating precise masks for multiple OoD objects, as demonstrated in the second and fourth rows.

Validation on the Fishyscapes dataset highlights the precision of S2M in detecting small anomalies. Our method excels in accurately identifying small OoD objects when the anomaly scores are optimal, as illustrated in the first row of Fig. 13. This capability is crucial for scenarios involving diminutive and subtle anomalies. Furthermore, S2M efficiently detects semi-transparent, synthetically created OoD objects, showcasing its robustness and precision in complex scenarios. This is effectively demonstrated in the fourth and fifth rows, where S2M successfully delineates these challenging objects without compromising accuracy.

The SMIYC dataset exemplifies the efficacy of our approach in addressing the diverse and dynamic nature of road obstacles. The comprehensive environment of SMIYC allows for the evaluation of our method's ability to detect a wide range of OoD objects on roadways, from tiny to larger, more conspicuous obstacles.

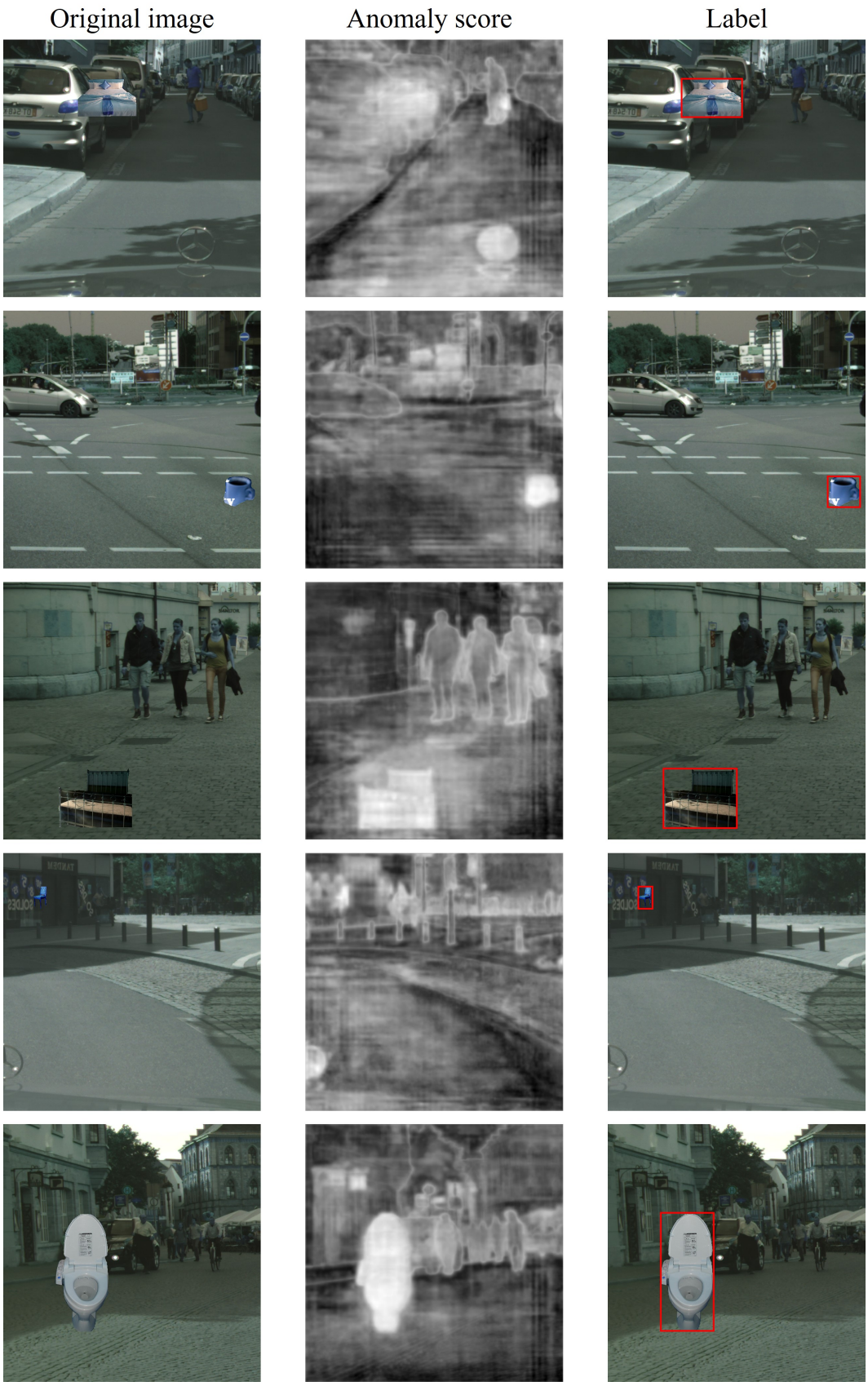


Figure 11. Visualization of training data.

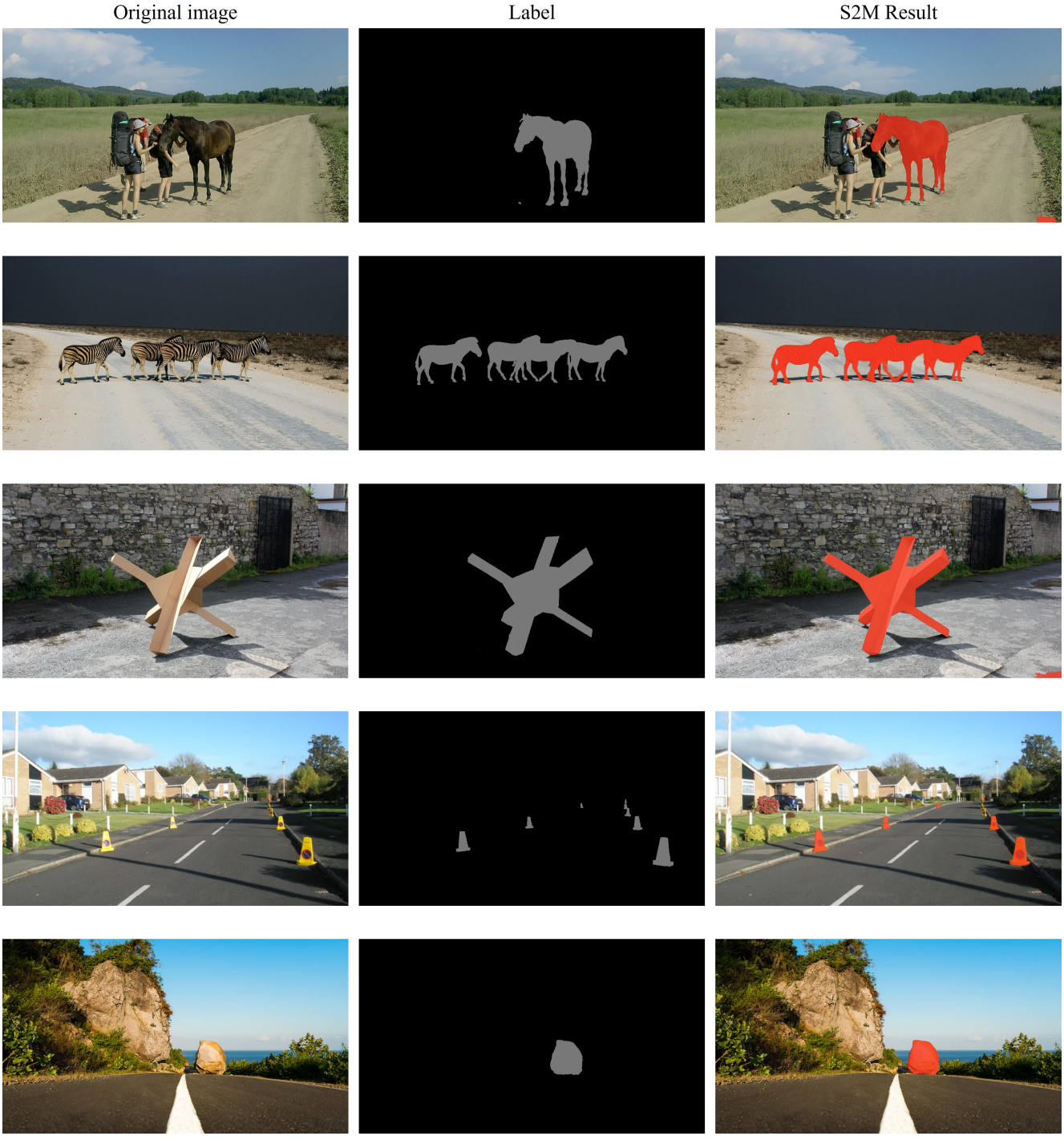


Figure 12. Visualization of S2M on Road Anomaly validation set. In the annotated images, pixels colored gray represent OoD objects, black pixels denote ID objects.

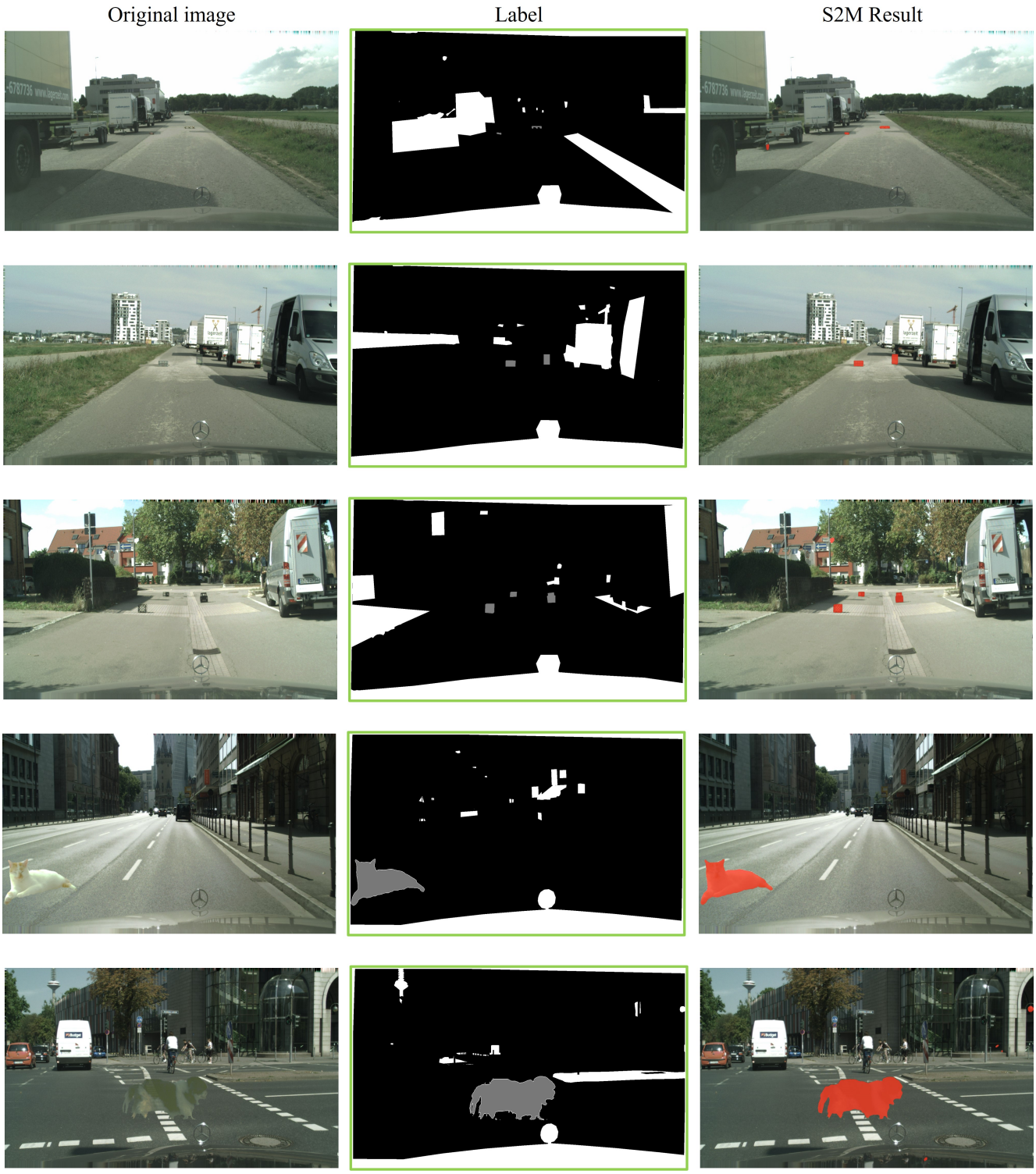


Figure 13. Visualization of S2M on Fishyscapes validation set. In the annotated images, pixels colored gray represent OoD objects, black pixels denote ID objects, and white pixels indicate regions to be ignored.

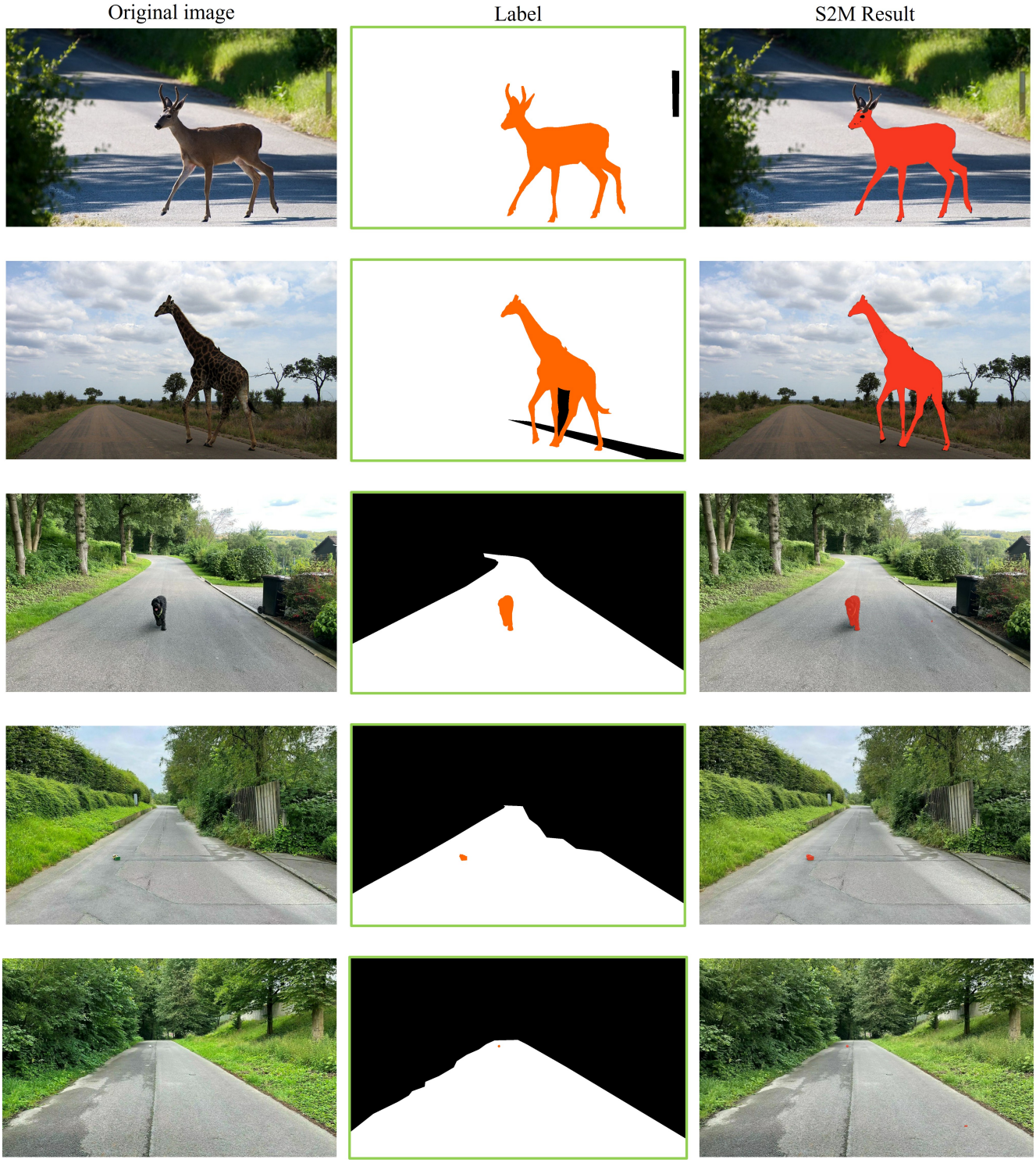


Figure 14. Visualization of S2M on SMIYC validation set. In the annotated images, pixels colored orange represent OoD objects, white pixels denote ID objects, and black pixels indicate regions to be ignored.