Training-Free OOD Object Detection Leveraging Pre-trained Segmentation Model Competency

Kimin Yun[‡,*]
Visual Intelligence Lab.
ETRI
Daejeon, South Korea
kimin.yun@etri.re.kr

Jeonghoon Song[‡,†]
Department of AI and Big Data
Soonchunhyang University
Asan, South Korea
sjhon121215@sch.ac.kr

Yuseok Bae
Visual Intelligence Lab.
ETRI
Daejeon, South Korea
baeys@etri.re.kr

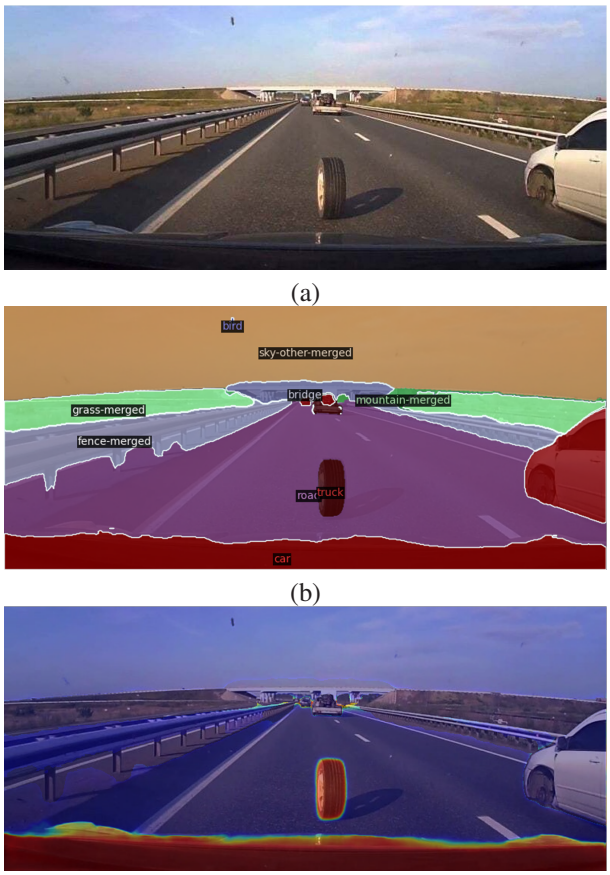
***Abstract*—This paper addresses a method for detecting Out-of-Distribution (OOD) objects by leveraging pre-trained segmentation models. The research on OOD detection focuses on the ability to accurately identify and classify untrained classes as "unknown." Previous methods rely on strong augmentation to detect OOD objects. However, these augmentation-based methods assume specific characteristics of OOD objects and can suffer from overfitting due to the limited datasets. In this study, we propose a training-free OOD object detection approach that infers OOD objects from a pre-trained segmentation model without additional training. Specifically, our method combines two modules: rejection of regions confidently recognized by the model (inliers) and selection of masks that capture the characteristics of OOD objects (outliers). Furthermore, using a class-agnostic segmentation model, the probability of OOD objects is refined at the object level, enhancing performance. Our method shows competitive performance on datasets where OOD objects are encountered in autonomous driving contexts. Additionally, we show the utility of measuring the model's competency to recognize what it knows and doesn't know from the perspective of the pre-trained model.**

***Index Terms*—Deep learning, object segmentation, autonomous driving, out-of-distribution**

I. Introduction

There has been growing interest in the capabilities of artificial intelligence models. Typically, these models are trained on large-scale datasets through supervised learning, achieving high performance on well-defined target applications. While this approach has driven the success of machine learning models, recent research has increasingly focused on understanding the extent of knowledge within these trained models and assessing their competencies in various scenarios [1], [2]. In this context, it is crucial to determine whether an input is Out-of-Distribution (OOD) from the perspective of the model's competency.

In the application of autonomous driving, detecting OOD objects—those that the model has not encountered during training—is particularly important, as it plays a critical role in tasks such as collision avoidance and planning. However, due to the limited data available for OOD objects, accurately modeling their shapes can be challenging. Earlier approaches attempted to detect OOD objects by modeling the background using video information [3]–[5], but more recent methods focus on leveraging the internal competency of pre-trained segmentation models to infer OOD objects [6]–[8].



(a)

(b)

(c)

Fig. 1. Comparison of semantic segmentation and anomaly detection results: (a) Input image, (b) Semantic segmentation, (c) OOD object detection.

For example, as shown in Figure 1, even in unusual scenarios, such as a car wheel rolling on the road, a semantic segmentation model may classify the regions into specific classes. In such cases, the model might mistakenly label the OOD object, like the wheel, as a known class, such as 'Truck'

[*] corresponding author.
[‡] These authors contributed equally to this work.
[†] This work was done during his internship at ETRI.

(Figure 1(b)). This misclassification occurs because traditional models often rely on detection confidence, leading to unseen objects being classified as 'background' and thus ignored. This presents a significant risk, especially in safety-critical applications like autonomous driving. Therefore, the goal of this paper is to leverage the model's inherent knowledge to detect and identify regions it considers anomalous or 'unknown,' as demonstrated in Figure 1(c). Through this map, the model can provide crucial information about objects it has not encountered before, thereby enhancing safety and reliability.

Conventional methods to OOD detection have involved additional training with synthetic data to differentiate the characteristics of inliers and outliers, leading to the retraining of networks. Energy-based models, which have shown good performance, operate by analyzing the distribution of class probabilities, assigning lower energy to well-learned inliers and higher energy to outliers. However, these methods require additional training and often produce a high number of false positives at the pixel level. Recently, approaches leveraging transformer-based networks to detect OOD masks at the mask level have shown promising results. These methods exploit the capabilities of networks trained on vast datasets to detect OOD objects without the need for further training, making effective use of the model's inherent knowledge.

In this paper, we extend the mask-level transformer-based methods using two key strategies. First, we identify candidate regions for out-of-distribution detection by combining two types of masks: a rejection-based mask that excludes regions where the model is confident, and an unknown mask that captures areas with low probabilities of belonging to any known class while also being confused with all pre-trained classes. Furthermore, similar to human perception, where an object can be recognized as a distinct entity without knowing its exact nature, we enhance the detection of unknown object regions by integrating with the Segment Anything model, which excels at identifying object boundaries in a class-agnostic manner. Experimental results on the SegmentMeIfYouCan (SMIYC) [9] and RoadAnomaly [10] datasets, which simulate abnormal object detection in autonomous driving scenarios, demonstrate that our proposed method effectively detects OOD objects by leveraging the model's inherent knowledge without additional training. Moreover, our method is adaptable to any pre-trained model, making it applicable not only for detecting unknown objects on the road but also as a tool for estimating areas or objects that the detector is uncertain about, thus serving as a valuable resource for further training or model improvement.

II. Related Works

Out-of-Distribution (OOD) detection evaluates a model's ability to recognize and classify data outside its training distribution as "unknown." Recent advances have emphasized energy-based models, which assess the distribution of class probabilities rather than relying solely on confidence scores. In these models, inliers are assigned low energy and exhibit a peaked distribution, while outliers receive high energy and tend toward a uniform distribution.

Traditional methods have enhanced OOD detection by incorporating synthetic training data and refining anomaly detection at the pixel level. For instance, PEBAL [11] effectively differentiates inliers from outliers by adjusting energy levels with soft margin parameters and introducing smoothness and sparsity regularization for outlier regions.

Balanced Energy [12] extends PEBAL by incorporating prior knowledge from the dataset to assign different weights to each class during loss computation. This approach balances energy distribution across classes during OOD detection. However, it requires the estimation of prior probability from sampled data during training.

Residual Pattern Learning (RPL) [13] integrates residual and contrastive learning techniques to address the degradation of inlier class performance when retraining with synthetic data. RPL preserves inlier weights and introduces residual learning for OOD data, while contrastive learning aligns synthesized images with their originals in feature space, separating unrelated backgrounds.

Rejected by All (RbA) [6] uses Mask2Former [14], a state-of-the-art segmentation algorithm. In this approach, transformer decoder queries represent class probabilities, and RbA interprets regions where all queries are negative as unknown objects. Despite limited data, the model is fine-tuned to classify outlier pixels as belonging to unknown classes.

Similarly, Maskomaly [7] employs a rejection-based method to exclude regions where the model lacks confidence, assuming that masks from certain queries represent unknown objects without additional training. Maskomaly selects the top four queries from the SegmentMeIfYouCan (SMIYC) [9] validation dataset that maximize performance, but this approach can lead to overfitting to dataset-specific characteristics.

The proposed method builds on RbA's rejection-based strategy and Maskomaly's training-free approach but avoids overfitting by selecting different queries for each image. Additionally, it improves the estimation of unknown object boundaries by incorporating a class-agnostic segmentation model, enabling more accurate object-level decisions.

III. Method

A. Rejection by Inliers

Our proposed method is fundamentally based on a transformer segmentation model. As illustrated in Figure 2, we employ the Mask2Former model [14], which takes an image as input and outputs both a membership map and class probabilities for each query. For instance, if the model has N total queries and is trained on K classes, the class probabilities have dimensions of $N \times K$, and the membership map has dimensions of $K \times H \times W$.

Previous method [6] has analyzed how object queries in mask classification behave like one-vs-all classifiers, operating almost independently when segmenting different masks. Based on this observation, the rejection-based method initially defines the entire image region as an anomalous and then

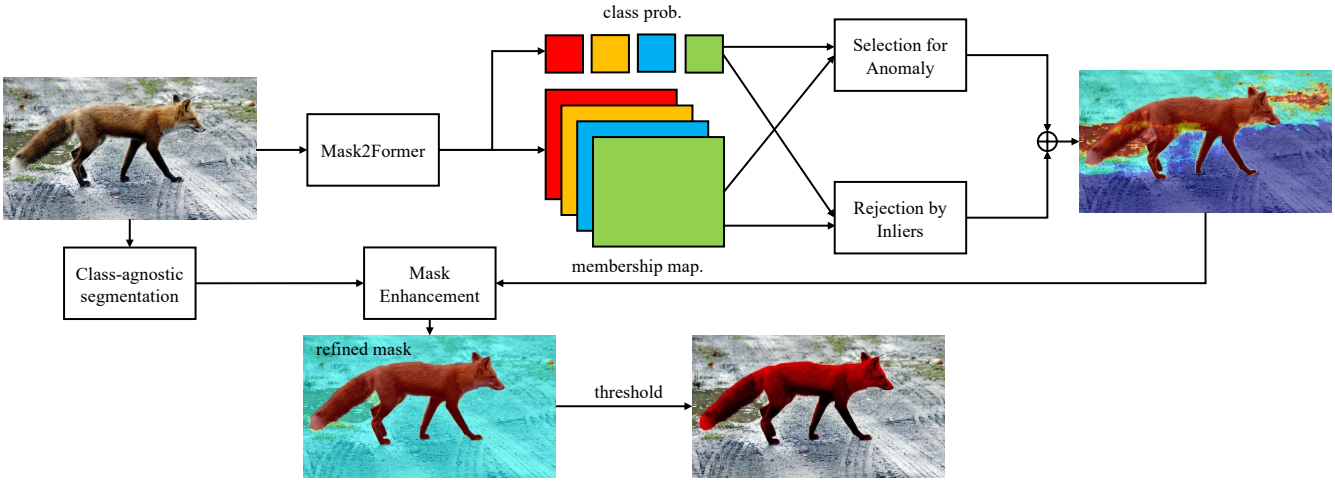


Fig. 2. Overall framework of the proposed method.

gradually removes the mask for regions confidently predicted by each query. This approach can be mathematically expressed as follows:

$$m_{\text{reject}}[x, y] = \min_{q \in Q} \left(1 - m_q[x, y] \cdot \max_{1 \le k \le K} p_q[k]\right), \quad (1)$$

where Q denotes the set of query indices, and x and y represent coordinates in the image. This method is conceptually similar to those employed in Maskomaly and RbA.

B. Selection for Anomalous Region

In the original Maskomaly approach, clues for identifying anomalous regions were derived by referencing the ground truth of the validation dataset. This allowed the model to infer queries that represent anomalous areas. However, while the ground truth can be useful for evaluating the model's performance in specific scenarios, using it to select specific queries introduces issues. These issues include dataset overfitting and unrealistic settings for practical applications.

To address this, we propose a method for dynamically selecting queries likely to represent anomalous regions based on the input image. According to the analysis in the RbA paper, outliers are often identified by queries with very low confidence and are rarely voted on by object queries. Additionally, it is a common assumption in many OOD detection studies that OOD objects should exhibit a uniform probability distribution across classes.

Based on these observations, we select queries likely to represent anomalous masks according to two criteria. First, to ensure the query has a low probability of belonging to any inlier class, we use the following criterion:

$$S = \{q \in Q \mid (1 - p_q[K+1]) < T_{\text{void}}\}, \quad (2)$$

where $K+1$ indicates the void (background) class.

Second, we select queries that not only have a low probability of belonging to any inlier class but also exhibit a distribution as close to uniform as possible. This is done using the following criterion:

$$S_{\text{selected}} = \arg \min_{q \in S, |S_{\text{selected}}| = N} \text{Var}\left(p_q[0], p_q[1], \ldots, p_q[K]\right), \quad (3)$$

Using these selected queries, we identify complementary regions for the anomalous area in a manner opposite to the rejection-based method. Similar to the approach used in Maskomaly, we combine the m_{reject} and m_{accept} masks linearly to generate a heatmap-like region that highlights the OOD objects.

$$m_{\text{accept}}[i, j] = \max_{s \in \mathcal{S}_{\text{selected}}} \left(m_s[i, j] \cdot p_s[K+1]\right). \quad (4)$$

$$m_{\text{init}} = \lambda \cdot m_{\text{reject}} + (1 - \lambda) \cdot m_{\text{accept}} \quad (5)$$

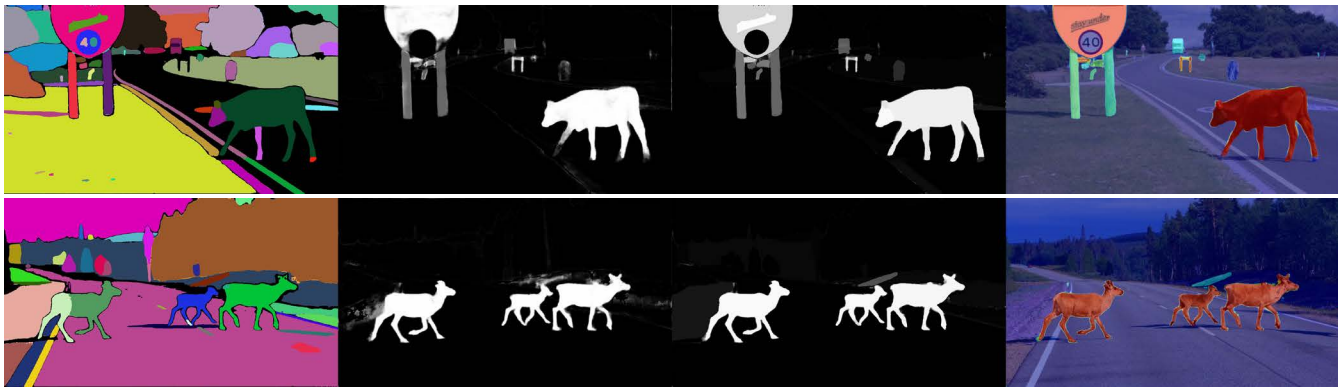
C. Combining with Segment Anything

The final mask is obtained by binarizing the heatmap in Eq. 5 through thresholding. However, this approach may result in a significant number of false positives. To improve object-level decision-making, we recognize that while the exact shape of an OOD object is unknown, the inherent characteristic that objects typically have closed boundaries remains consistent.

To leverage this property, we utilize the Segment Anything Model (SAM) [15], which performs class-agnostic segmentation. The SAM model generates masks for any object in the image based on user input prompts, regardless of class. When these prompts are applied uniformly across the entire image, SAM provides candidate masks for all objects present in the scene. This allows us to group anomalous values into distinct object-level segments, reducing false positives and leading to more accurate object-level inferences.

When the input image I is processed through the Segment Anything model (SAM), it produces the following segment regions:

$$SAM(I) = r_1, r_2, \ldots, r_N$$



(a) (b) (c) (d)

Fig. 3. Integration with the Segment Anything Model: (a) SAM result, (b) Initial mask, (c) Refined mask, and (d) Anomalous region overlaid on the input image.

where $r_i \in \{0,1\}^{H\times W}$ represents the i-th segment region (binary mask). Next, the values of the initial mask m_{init} within each segment are averaged and assigned to the refined mask m_{refined} as shown in the equation below:

$$m_{\text{refined}}[x,y] = \sum_{i=1}^{N} r_i(x,y) \cdot \frac{\sum_{x',y'} m_{\text{init}}[x',y'] \cdot r_i(x',y')}{\sum_{x',y'} r_i(x',y')} \tag{6}$$

Figure 3 shows the application of the Segment Anything model. As shown in Figure 3(a), the SAM model extracts class-agnostic masks from the input image. These masks are then used to refine the initial mask obtained through our proposed method, as shown in Figure 3(b). Specifically, for each segment identified by SAM, the values from (b) are averaged within that segment and assigned to the corresponding mask. This process results in a heatmap that is better aligned with object contours, as shown in Figures 3(c) and (d).

This approach not only reduces false positives but also improves the model's performance by making object-level decisions, even when only part of an object is identified as unknown. By combining the initial mask with the object boundaries identified by SAM, we enhance the overall accuracy of OOD detection.

IV. Experiments

A. Experimental Setup

We use the SegmentMeIfYouCan (SMIYC) [9] and RoadAnomaly [10] datasets, both of which contain road anomalies in street scenes with pixel-level annotations. We evaluate the performance of various approaches, including ours, using the Average Precision (AP) metric for anomaly segmentation. For our experiments, we employed the Mask2Former model [14] with a Swin-L [16] backbone, which was trained on the CityScapes dataset for semantic segmentation to ensure a fair comparison with existing methods.

For our experiments, we set the hyperparameters as follows: $T_{void} = 0.99$ in Eq. (2), $\lambda = 0.6$ in Eq. (5), and $N = 2$ in Eq. (3). All other key parameters and techniques, including the method for handling ambiguous labels in the ground truth, were applied in the same manner as in the Maskomaly [7].

B. Quantitative Results

We compare our method against baselines and state-of-the-art methods, using semantic segmentation models like DeepLabV3 [18] and Mask2Former [14]. The segmentation frameworks employed by each method are shown in Table I.

Table I shows the quantitative results of our approach on the SMIYC [9] and RoadAnomaly [10] datasets. Without relying on additional data and training, our method achieved the second-highest performance on SMIYC, just behind Maskomaly, and the highest performance on RoadAnomaly. As noted earlier, Maskomaly's approach directly selects queries based on the validation dataset to identify anomalous mask regions, which can result in overfitting and create an unfair comparison. On the RoadAnomaly dataset, our method outperforms the baseline methods (RbA and Maskomaly). Also, it demonstrates performance nearly on par with RbA, even with its additional training.

C. Qualitative Results

Figures 4 and 5 show the qualitative results on the SMIYC and RoadAnomaly datasets, respectively. As shown in Figure 4, Maskomaly performs relatively well on the SMIYC dataset, but our proposed method, with the help of the Segment Anything model, further reduces false positives.

In the results on the RoadAnomaly dataset, illustrated in Figure 5, Maskomaly encounters issues where areas like building walls and signposts are incorrectly detected as anomalous regions. This is because Maskomaly relies heavily on the SMIYC validation set to infer the mask regions of OOD objects. In contrast, our method addresses this issue by modeling anomalous regions through query selection tailored to the characteristics of each image.

Figure 6 shows the results from Mask2Former, highlighting both the semantic segmentation and the anomalous regions detected by our method with a model trained on the COCO

TABLE I
BENCHMARK RESULTS ON SMIYC [9] AND ROADANOAMLY DATASET [10]. WE SEPARATE METHODS THAT REQUIRES ADDITIONAL TRAINING WITH AUXILIARY DATA.

Method	Segmentation Framework	Aux. Data	AP for SMIYC	AP for RoadAnomaly
PEBAL [17]	DLV3+ [18]	✓	49.1	45.1
SynBoost [19]	VPLR [20]	✓	56.4	38.2
DenseHybrid [21]	DLV3+ [18]	✓	78.0	63.9
Max. Entropy [22]	DLV3+ [18]	✓	85.5	79.7
EAM [8]	M2F [14]	✓	**93.8**	66.7
RbA [23]	M2F [14]	✓	90.9	**85.4**
DenseHybrid [24]	DLV3+ [18]	✗	51.5	35.1
ObsNet [25]	DLV3+ [18]	✗	75.4	54.7
EAM [8]	M2F [14]	✗	76.3	66.7
RbA [23]	M2F [14]	✗	86.1	78.5
Maskomaly [7]	M2F [14]	✗	**93.4**	70.9
Proposed method	M2F [14]	✗	91.7	**81.2**

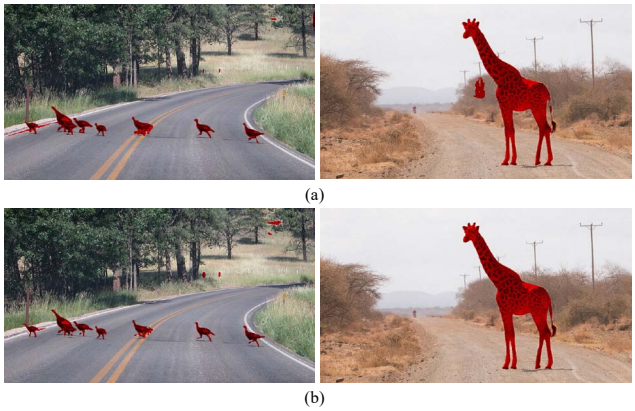


Fig. 4. Qualitative comparison of results on the SMIYC dataset [9]: (a) Maskomaly [7] and (b) Proposed method.

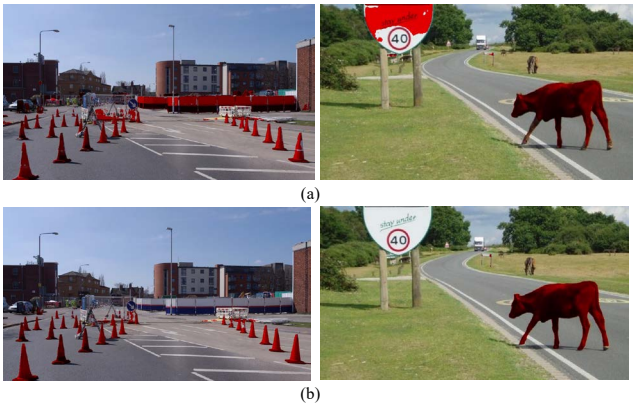


Fig. 5. Qualitative comparison of results on the RoadAnomaly dataset [10]: (a) Maskomaly [7] and (b) Proposed method.

dataset. As discussed in the Introduction, semantic segmentation (left column) segments the image based on the highest confidence among inlier classes, making it challenging to assess the model's true competency in recognizing certain regions. By applying our method (right column), OOD regions are detected based on the model's inherent competency, offering not only better OOD detection but also valuable insights

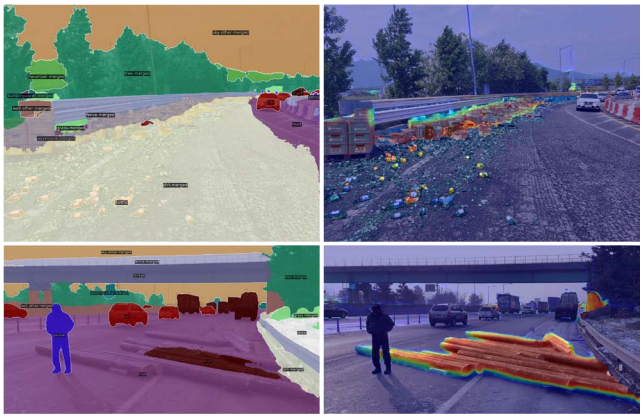


Fig. 6. Semantic segmentation results and anomaly detection for coco-trained model

for further model improvement.

V. LIMITATIONS AND FUTURE WORKS

This paper proposed a method for detecting OOD object regions by leveraging the model's inherent competency without additional training. However, the definition of OOD can be ambiguous and varies dependeing on the task. For example, a giraffe on a road might be considered OOD in conetext of autonomous driving but an inlier for a model trained on COCO, which could cause a misalignment with the intended task. Future research will focus on enhancing model competency, particularly in scenarios where some pre-trained models possess knowledge that others do not. Additionally, we plan to explore the use of multi-level features to improve the effectiveness of OOD detection.

VI. CONCLUSION

In this paper, we proposed a method for Out-of-Distribution (OOD) object detection that leveraged the inherent competency of transformer-based segmentation models. Our approach used a rejection-based strategy to eliminate regions confidently predicted by known classes, followed by dynamic query selection for identifying anomalous regions. This method addressed issues like overfitting by tailoring query selection to the input image. We also integrated our method with

the Segment Anything Model (SAM) to refine object-level OOD detection, reducing false positives and improving accuracy. Experiments on the SMIYC and RoadAnomaly datasets showed that our approach outperformed existing methods in anomaly segmentation, achieving higher Average Precision (AP) scores. Its flexibility allows adaptation to different pre-trained models without additional training, making it suitable for various applications, including autonomous driving. By focusing on the model's ability to recognize and handle unknown objects, our method contributes to enhancing the safety and reliability of AI systems deployed in real-world scenarios.

ACKNOWLEDGMENT

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00124 & RS-2022-II220124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities).

REFERENCES

[1] S. Pohland and C. Tomlin, "Understanding the dependence of perception model competency on regions in an image," in *World Conference on Explainable Artificial Intelligence*, pp. 130–154, Springer, 2024.

[2] H.-I. Kim, K. Yun, J.-S. Yun, and Y. Bae, "Customizing segmentation foundation model via prompt learning for instance segmentation," *arXiv preprint arXiv:2403.09199*, 2024.

[3] K. Yun, Y. Kwon, S. Oh, J. Moon, and J. Park, "Vision-based garbage dumping action detection for real-world surveillance platform," *ETRI Journal*, vol. 41, no. 4, pp. 494–505, 2019.

[4] K. Yun, H. Kim, K. Bae, and J. Park, "Unsupervised moving object detection through background models for ptz camera," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3201–3208, IEEE, 2021.

[5] K. Yun, H.-I. Kim, K. Bae, and J. Moon, "Background memory-assisted zero-shot video object segmentation for unmanned aerial and ground vehicles," *ETRI Journal*, vol. 45, no. 5, pp. 795–810, 2023.

[6] N. Nayal, M. Yavuz, J. F. Henriques, and F. Güney, "Rba: Segmenting unknown regions rejected by all," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[7] J. Ackermann, C. Sakaridis, and F. Yu, "Maskomaly: Zero-shot mask anomaly segmentation," in *The British Machine Vision Conference (BMVC)*, 2023.

[8] M. Grcic, J. Šaric, and S. Šegvic, "On advantages of mask-level recognition for open-set segmentation in the wild.," in *CVPR 2023 workshop on Visual Anomaly and Novelty Detection (VAND)*, 2023.

[9] R. Chan, K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, P. Fua, M. Salzmann, and M. Rottmann, "Segmentmeifyoucan: A benchmark for anomaly segmentation," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[10] K. Lis, K. Nakka, P. Fua, and M. Salzmann, "Detecting the unexpected via image resynthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2152–2161, 2019.

[11] Y. Tian, Y. Liu, G. Pang, F. Liu, Y. Chen, and G. Carneiro, "Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes," in *European Conference on Computer Vision*, pp. 246–263, Springer, 2022.

[12] H. Choi, H. Jeong, and J. Y. Choi, "Balanced energy regularization loss for out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15691–15700, 2023.

[13] Y. Liu, C. Ding, Y. Tian, G. Pang, V. Belagiannis, I. Reid, and G. Carneiro, "Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1151–1161, 2023.

[14] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[17] Y. Tian, Y. Liu, G. Pang, F. Liu, Y. Chen, and G. Carneiro, "Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes," in *European Conference on Computer Vision*, pp. 246–263, 2022.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[19] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-wise anomaly detection in complex driving scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16918–16927, 2021.

[20] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8856–8865, 2019.

[21] M. Grcić, P. Bevandić, and S. Šegvić, "Densehybrid: Hybrid anomaly detection for dense open-set recognition," in *European Conference on Computer Vision*, pp. 500–517, Springer, 2022.

[22] R. Chan, M. Rottmann, and H. 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*, pp. 5128–5137, 2021.

[23] N. Nayal, M. Yavuz, J. F. Henriques, and F. Güney, "Rba: Segmenting unknown regions rejected by all," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 711–722, 2023.

[24] M. Grcić and S. Šegvić, "Hybrid open-set segmentation with synthetic negative data," *IEEE transactions on pattern analysis and machine intelligence*, 2024.

[25] V. Besnier, A. Bursuc, D. Picard, and A. Briot, "Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15701–15710, 2021.