Maskomaly:
Zero-Shot Mask Anomaly Segmentation

Jan Ackermann[1]
ackermannj@ethz.ch

Christos Sakaridis[2]
csakarid@vision.ee.ethz.ch

Fisher Yu[1,2]
fisheryu@ethz.ch

[1] Department of Computer Science
ETH Zurich,
Switzerland

[2] Computer Vision Lab
ETH Zurich,
Switzerland

Abstract

We present a simple and practical framework for anomaly segmentation called Maskomaly. It builds upon mask-based standard semantic segmentation networks by adding a simple inference-time post-processing step which leverages the raw mask outputs of such networks. Maskomaly does not require additional training and only adds a small computational overhead to inference. Most importantly, it does not require anomalous data at training. We show top results for our method on SMIYC, RoadAnomaly, and StreetHazards. On the most central benchmark, SMIYC[1], Maskomaly outperforms all directly comparable approaches. Further, we introduce a novel metric that benefits the development of robust anomaly segmentation methods and demonstrate its informativeness on RoadAnomaly.

1 Introduction

Anomaly detection is the task of identifying whether one data point belongs to a set of inlier classes that have been seen during the training. Recently, Fan *et al*. [13] have demonstrated how difficult anomaly detection is from a theoretical viewpoint. Nevertheless, it is an essential component of many real-world systems operating in safety-critical settings, such as autonomous cars. In order to achieve fully automated driving, autonomous cars need to understand when an anomaly is present and where the anomaly is located in the scene. The latter task is significantly more challenging, as a model needs not only to output one single anomaly score per scene but a dense map of pixel-level scores. With more diverse datasets [4, 7, 30, 36] and more complex training paradigms [1, 8, 35, 36], current approaches have already achieved high scores on relevant benchmarks. However, the current state of the art has not yet reached a level of accuracy that would allow deployment in real-world settings. Given these facts, one might expect that an even more complex training pipeline is needed to achieve results of practical utility.

Nonetheless, we propose a very simple yet effective zero-shot method which requires no training on anomalous data. Our method, Maskomaly, builds upon mask-based semantic segmentation networks, such as Mask2Former [12], by post-processing their raw mask predictions to compute a dense anomaly heatmap, only adding a small computational overhead at inference. To the best of our knowledge, we are the first, along with the concurrent work in [25, 40], to explore the utility of mask-based semantic segmentation networks for anomaly segmentation. Our key insight is that mask-based networks trained for standard semantic segmentation already learn to assign certain

[1]Public Benchmark at https://segmentmeifyoucan.com/leaderboard

Figure 1: Maskomaly results on the SMIYC [7] test set. Segmentations are shown in red and obtained by thresholding at 0.9.

masks to anomalies. Even though such masks are discarded by default when generating semantic predictions, we show that they can be leveraged for inference on images potentially containing anomalies to achieve state-of-the-art results in anomaly segmentation. Although Maskomaly is an intuitive extension of Mask2Former, we show that computing the anomaly heatmap from the mask outputs of the latter is nontrivial and justify our proposed components through proper ablations. On multiple anomaly segmentation benchmarks, Maskomaly beats all state-of-the-art methods that do not train with auxiliary data and most methods that perform such training. We evidence the generality of our key insight and method across different backbones. Finally, we present a new metric that promotes robust anomaly prediction methods tailored for real-world settings.

2   Related Work

Anomaly Detection: Identifying samples that deviate from a known probability distribution is an old problem [26], generally referred to as anomaly detection. Early works identify the uncertainty in the model and utilize ensembling [32], input perturbations [35], and max-softmax probability [28]. More encouraging results are achieved by methods that estimate the likelihood with a generative model [39, 42, 46] or that train a model discriminatively with negative data [2, 23, 29, 33, 47]. This is also referred to as training with auxiliary data.

Anomaly Segmentation   aims to predict a dense map of outlier probabilities for a given image. Works such as [4, 41] provide methods that average over multiple predictions. Another line of work estimates the uncertainty by computing the dissimilarity between the input and a resynthesized image [14, 36, 44]. The synthesized output is conditioned on the semantic segmentation of the input. A recent work [1] trains a network to predict errors of a frozen semantic segmentation network. To stimulate failure, they attack the network with FGSM [20]. GMMSeg [34] uses GMMs to capture class-conditional densities and train the dense representation in a discriminative manner. Methods that train on auxiliary data include Pebal [43], which jointly optimizes their novel pixel-wise anomaly abstention learning and energy-based models, DenseHybrid [24], which utilizes discriminative and generative modeling simultaneously, and Max. Entropy [8], which learns to predict high entropy at anomalous regions and uses a meta-classifier to reduce the number of false positives.

Mask-Based Approaches: The outputs of mask-based segmentation networks have been leveraged in [27], which uses the uncertainty in the instance predictions to detect bounding boxes of anomalies. [6] predicts whether an anomaly is present based on the count of instance proposals. Unlike these methods, our method predicts a dense anomaly map for an image, assigning a continuous score to each pixel. Moreover, [19] segments anomalies by rejecting regions with known objects. Distinctly, we do not train one binary classifier per semantic class but use one semantic segmentation model. Further, rejecting known regions is only one part of our method, as we also utilize "anomalous" masks as positive predictions. A concurrent work [25] considers Mask2Former for anomaly
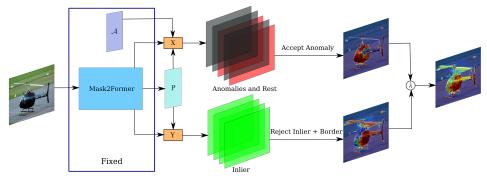
Figure 2: Maskomaly computes the anomaly segmentation by combining two separate predictions. Based on the outputs of class distribution P, module Y chooses which masks belong to inliers and constructs a prediction by assigning low values within inlier masks or across their shared borders. X constructs a second prediction by assigning high probability to pixels included in masks which were found to predict anomalies on a validation set. This set of masks is denoted by $\mathcal{A}$. The final prediction is obtained by interpolating the two initial ones with parameter λ.

segmentation. They contribute a score combined with a standard detector [30]. Furthermore, another work [40] contributes a new metric for training to detect outliers. By contrast, our work contributes a complete post-processing method, does not use auxiliary data, and views masks as cohesive units.

3 Method

Given a set of C semantic inlier classes, our goal is to find an algorithm that receives an RGB image of dimension $H \times W \times 3$ and outputs a dense anomaly segmentation $o = [0,1]^{H \times W}$. It should assign high values to anomalous regions and low values to inliers. Anomalies are objects that do not fit into any of the C classes. For our method, we utilize intermediary outputs of any mask-based segmentation networks. In particular, we work with a set of soft-masks $M = \{m_1,...,m_N\}$ and a set of probability vectors $P = \{p_1,...,p_N\}$, where N is the number of masks and $C+1$ is the label of the void class. A pixel $[i,j]$ belongs to a mask m_k if and only if $k = \text{argmax}_{1 \le n \le N} m_n[i,j]$. Moreover, a mask m_k belongs to an inlier class if and only if $\text{argmax}_{1 \le l \le C+1} p_k[l] \ne C+1$. In the following sections, we use Mask2Former [12] with a Swin-L backbone and trained on Cityscapes [13] because it was the best-performing mask-based segmentation network at the time of writing. However, our method is not restricted to a specific segmentation model.

3.1 Masks as Binary Classifiers

Our first idea is conceptually simple: The segmentation network outputs class probabilities and mask scores for each query. For the C known classes in an image, we expect it to output at least one mask. We confirmed this intuition by evaluating the average IoU of all semantic classes on the test set of Cityscapes [13]. Due to the high IoU scores, we interpret each mask as a binary classifier. For a mask m_i, this means that there is a set $s_i^{T_{\text{mask}}}$ of 2D coordinates that belong to it with probability higher than threshold T_{mask} and its complement $\overline{s}_i^{T_{\text{mask}}}$. Under this assumption, a pixel belongs to an anomaly if and only if it is inside $O = \bigcap_{i \in \mathcal{I}} \overline{s}_i^{T_{\text{mask}}}$. This is the intersection of all complements of inlier masks. Here, $\mathcal{I}$ denotes the set of indices of all masks which belong to an inlier class according to P. A straightforward way to predict anomalies would be to assign $o[i,j] = \mathbb{1}\{[i,j] \in O\}$. To avoid overconfident predictions, we scale the output by the uncertainty of the dominant mask and receive a softer assignment:

$$o_{\text{reject}}[i,j] = \min_{n \in \mathcal{I}}(1 - m_n[i,j] \cdot \max_{1 \le l \le C} p_n[l]). \tag{1}$$

We observe that other methods assign high outlier scores at the boundaries of two inlier classes. Working with masks, we find these areas by computing the pairwise intersection of inlier class masks after thresholding them at T_b and assign these areas an outlier score ε_b.

3.2 Anomalous Masks

Despite the simplicity and theoretical effectiveness of the above rejection-based approach, we must account for possible failure cases, such as large domain shifts. Similar to [25], we observed that on FishyScapes LaF [4], over 50% of the outlier pixels belong to an inlier mask. Although the analysis also shows that inlier pixels belong to inlier masks with over 90%, these findings suggest the sub-optimal performance of o_{reject}.

Algorithm 1 Maskomaly, Hyperparameters $= \{T_{mask}, T_b, \varepsilon_b, \lambda\}$

1: **function** MASKOMALY($M, P, \mathcal{A}, \mathcal{I}$)
2:    **for** $1 \leq n \leq N$ **do**   ▷ Add inlier masks to $\mathcal{I}$
3:        **if** $(\text{argmax}_{1 \leq l \leq C+1} p_n[l]) \neq C+1 \wedge (\text{max}_{1 \leq l \leq C+1} p_n[l]) \geq T_{mask}$ **then**
4:            $\mathcal{I} \leftarrow \mathcal{I} \cup \{n\}$
5:    **for** $i, j \in H \times W$ **do**   ▷ Reject inlier pixels
6:        $o_{reject}[i,j] \leftarrow \min_{n \in \mathcal{I}}(1 - m_n[i,j] \cdot \max_{1 \leq l \leq C} p_n[l])$
7:    **for** $\{k,n\} \subseteq \mathcal{I}$ **do**   ▷ Reject borders
8:        $b \leftarrow (m_k > T_b) \odot (m_n > T_b)$
9:        $b \leftarrow \min(1 - b + \varepsilon_b, 1)$
10:       $o_{reject} = \min(o_{reject}, b)$
11:   **for** $i, j \in H \times W$ **do**   ▷ Accept anomalous predictions
12:       $o_{accept}[i,j] \leftarrow \max_{n \in \mathcal{A}}(m_n[i,j] \cdot p_n[C+1])$
13:   **return** $\lambda \cdot o_{reject} + (1 - \lambda) \cdot o_{accept}$ ▷ Interpolate reject and accept scores

---

We go one step further and analyze the behavior and impact of each mask individually. As mask-based segmentation frameworks include mask-loss [11, 12], it is natural to expect that some queries specialize in predicting a single semantic class. We call a query specialized for class c if it predicts a single class c with probability $1 - \varepsilon$ in over T_{query} of the cases. We verified this expectation by evaluating the class probabilities on a test set of Cityscapes [13] with $T_{query} = 0.9$ and $\varepsilon = 0.1$ and found queries that predict classes. This gives rise to whether there are also queries that specialize in detecting void-class objects or anomalies.

The main challenge of finding void class queries is differentiating between queries that predict coherent masks of void class objects and queries that predict noise. We found that we cannot distinguish them reliably based on their class probabilities. Instead, we compute the average IoU of each mask with the anomalies on a validation set and extract all queries that generate masks with an average IoU larger than a threshold T_{IoU}. In Section 5.4, we show the quantitative results of this method. With this set of masks $\mathcal{A}$, we compute a different prediction that accepts predictions from masks that are associated with anomalies:

$$o_{accept}[i,j] = \max_{n \in \mathcal{A}} (m_n[i,j] \cdot p_n[C+1]). \tag{2}$$

Finally, we present Maskomaly in Alg. 1, combining the modules from Sec. 3.1 and 3.2.

4 Maximal Detection Margin

Chan *et al.* [7] argue that for practical reasons, we need to consider more metrics than the standard Average Precision, FPR95, and AuROC. Fig. 3 illustrates the issue with Average Precision. Despite Max. Entropy achieving a high AP of 95.5%, it assigns high scores to large parts of the image

Figure 3: Comparison of Max. Entropy [8] (top) with Maskomaly (bottom) on an image from SMIYC [7]. The first column shows dense predictions; the others show predictions thresholded at 0.3, 0.5, and 0.8. ME achieves 95.5% AP and Maskomaly 98.9%.

that are not anomalous. Although it is helpful that a metric is independent of a threshold, it is not applicable in practice. In practice, we need to fix a threshold, and with high probability, it should separate the data well. Our last contribution is a new metric for anomaly detection that encourages more robust methods.

We present Maximal Detection Margin (MDM) that requires a threshold T_{margin} and a metric d which measures the difference between a binary ground-truth map B and a soft prediction A thresholded at multiple levels. MDM is computed as the length of the largest interval in which all thresholds achieve at least T_{margin} regarding metric d. Further, it is normalized to $[0,1]$ and computed by:

$$\text{MDM}_d^{T_{margin}}(A,B) = \max_{0 \leq x < y \leq 1} (y-x) \cdot \mathbb{1}\{\forall z \in [x,y] : d(A > z, B) > T_{margin}\}. \tag{3}$$

5 Experiments

5.1 Experimental Setup

Datasets: **SegmentMeIfYouCan** (SMIYC) [7] is a road-anomaly dataset consisting of 100 test and 10 validation images of anomalies in street scenes with pixel-level annotations. The images show diverse perspectives and anomaly types and have been sourced from the internet. The authors withhold the test set; the scores are only accessible by submitting the method to the official benchmark. **FishyScapes** [4] Static is a dataset based on inserting objects from Pascal VOC [16] into Cityscapes [13] images. Its test set includes 1000 images, and its validation set includes 30. We only evaluated our model on the validation set. **RoadAnomaly** [36] consists of 60 images, of which 30 are also part of SMIYC's test set. It shows many scenes with anomalies on streets sourced from the internet. **StreetHazards** [30] is a synthetic dataset recorded inside the Carla Simulator [15]. It includes 1500 images and is incredibly challenging due to the high diversity of anomaly types, sizes, and locations.

Metrics: We evaluate other approaches and ours on the standard metrics for anomaly segmentation, Average Precision (AP), False Positive Rate at True Positive Rate of 95% (FPR95), and Area under Receiver Operator Characteristic (AuROC). On SMIYC, we include their three proposed metrics sIoU gt, PPV, and mean F_1 [7]. On RoadAnomaly, we also evaluate with our novel MDM metric.

Hyperparameters: For our experiments, we use the semantic segmentation Mask2Former model [12] with a Swin-L [38] backbone which was trained on CityScapes [13]. Cheng *et al.* describe their complete training setup in their work [12]. For all experiments we use the same SMIYC [7] validation set.

Our method is not sensitive to the specific values of its hyperparameters. This is mainly because Mask2Former makes robust predictions. Experiments show that on RoadAnomaly [36],

Table 1: Benchmark results on SMIYC [7]. The best method is marked in bold, and the second best is underlined. We separate methods that use auxiliary data during training.

Method	Segmentation Framework	Aux. data	Anomaly Track				
			AP↑	FPR95↓	sIoU gt↑	PPV↑	mean F$_1$↑
PEBAL [43]	DLV3+ [10]	✓	49.1	40.8	38.8	27.2	14.5
SynBoost [14]	VPLR [48]	✓	56.4	61.9	34.67	17.8	10.0
DenseHybrid [24]	DLV3+ [10]	✓	78.0	9.8	54.2	24.1	31.1
Max. Entropy [8]	DLV3+ [10]	✓	85.5	15.0	49.2	39.5	28.7
JSRNet [44]	DLV3 [9]	✗	33.6	43.9	20.2	29.3	13.7
DenseHybrid [21]	DLV3+ [10]	✗	51.5	33.2	-	-	-
Image Resyn. [36]	PSP [47]	✗	52.3	25.9	39.7	11.0	12.5
ObsNet [1]	DLV3+ [10]	✗	75.4	26.7	44.2	52.6	45.1
EAM [25]	M2F [12]	✗	76.3	93.9	-	-	-
Maskomaly (ours)	M2F [12]	✗	93.4	6.9	55.4	51.5	49.9

over 85% of the masks are assigned to a class with a confidence of ≥ 0.9. Thus, $T_{mask} \leq 0.9$ is a good choice. 93% of the pixels in the entire set of soft mask membership maps for the aforementioned masks have a score of ≤ 0.1, and from the remaining pixels, 80% have a score of ≥ 0.9. Thus, the "border" pixels can be separated well from the "core" mask pixels based on their mask membership scores, and any $T_b \geq 0.1$ is a good choice for doing so. W.r.t. λ, if we assume "accept" and "reject" are two per-pixel anomaly classifiers that correctly classify pixels with $p > 0.5$, then λ should be chosen based on the most likely prior (non-anomalies) due to symmetry. Hence, *any $\lambda > 0.5$ is a good choice to optimize for AP.*

In our experiments we set $\lambda = 0.6$, $T_{IoU} = 0.25$, $T_{mask} = 0.3$, $T_b = 0.1$, $\varepsilon_b = 0.001$, and $T_{query} = 0.9$. We initialize $\mathcal{I}$ to include queries influenced by ambiguous labeling of ground [13]. We chose the indices based on IoU with street predictions on the SMIYC [7] validation set.

5.2 Comparison with the State of the Art

We include baselines and state-of-the-art methods with segmentation modules and backbones of similar predictive power in our evaluation. Most approaches, including DenseHybrid [24], ObsNet [1], and Max. Entropy use DeepLabV3+ [10] with a WideResNet38 [48] backbone. EAM [25] uses Mask2Former [12]. All of these frameworks achieve an IoU of around 83% on Cityscapes [10, 12, 48].

Results on SMIYC are shown in Table 1. Our method beats all directly comparable approaches in AP, FPR95, sIoU, and mean F$_1$. It is especially noteworthy that our method does not use auxiliary data and improves substantially upon previous such methods as EAM [25] and RbA [40] which were both designed around Mask2Former [12] as well. Our performance is even comparable to theirs when trained with anomalous data. Overall, our method ranks third in AP on the entire public SMIYC benchmark. Maskomaly's results on the validation set are visualized and compared to other state-of-the-art methods in Fig. 4. Our method can deal with a challenging variety of sizes, shapes, and counts of anomalies. ObsNet [1] fails to detect multiple anomalous regions, such as the complete giraffe, the phone booth, and 4 of the birds. Max. Entropy [8] predicts anomalous regions outside the true anomalous objects, such as the people in the tank's background and the street next to the phone booth.

Table 2 shows our performance compared to state-of-the-art methods on RoadAnomaly [36] and the validation set of FishyScapes [4] Static. We observe that on RoadAnomaly, our method outperforms all previous ones that have not trained with auxiliary data by a margin of 4.2% on

Table 2: Results on RoadAnomaly [36] and FishyScapes [4] Static validation set. The best method is marked in bold, and the second best is underlined. We separate methods that use auxiliary data during training.

| Method | Segmentation Framework | Aux. data | RoadAnomaly | | | | FishyScapes Static | |
			AP↑	FPR95↓	$MDM_{F_1}^{60}$↑	$MDM_{F_1}^{70}$↑	AP↑	FPR95↓
SynBoost [14]	VPLR [48]	✓	38.2	64.8	0.0	0.0	66.4	25.6
PEBAL [43]	DLV3+ [10]	✓	45.1	44.6	-	-	**92.1**	**1.5**
DenseHybrid [24]	DLV3+ [10]	✓	63.9	43.2	-	-	60.0	4.9
Max. Entropy [8]	DLV3+ [10]	✓	**79.7**	19.3	**25.2**	**9.2**	76.3	7.1
DenseHybrid [24]	DLV3+ [10]	✗	35.1	43.2	-	-	54.7	15.5
ObsNet [1]	DLV3+ [10]	✗	54.7	60.0	5.1	0.0	9.4	47.7
GMMSeg [34]	SF [45]	✗	57.7	44.3	-	-	<u>82.6</u>	-
EAM [25]	M2F [12]	✗	66.7	13.4	-	-	**87.3**	**2.1**
Maskomaly (ours)	M2F [12]	✗	<u>70.9</u>	**11.9**	**67.1**	**35.9**	69.5	<u>14.4</u>
Maskomaly [opt]	M2F [12]	✗	**80.8**	<u>12.0</u>	<u>62.6</u>	<u>20.8</u>	68.8	15.0

Table 3: Results on StreetHazards [30] test set. The best method is marked in bold, and the second best is underlined. We separate methods that use auxiliary data during training.

| Method | Segmentation Framework | Backbone | Aux. data | SH trained | StreetHazards (SH) | | |
					AP↑	FPR95↓	AUC↑
OH*MSP [8]	LDN [51]	DenseNet-169	✓	✓	18.8	30.9	89.7
Max. Entropy [8]	DeepLabV3+ [10]	WideResNet-38	✓	✗	19.3	22.9	91.6
DenseHybrid [24]	LDN [51]	DenseNet-121	✓	✓	**30.2**	**13.0**	**95.6**
SO+H [22]	LDN [51]	DenseNet-121	✗	✓	12.7	25.2	91.7
ObsNet [1]	DeepLabV3+ [10]	WideResNet-38	✗	✗	12.8	25.3	90.5
DML [6]	PSP-Net [47]	ResNet-50	✗	✓	14.7	**17.3**	<u>93.7</u>
DenseHybrid [24]	LDN [51]	DenseNet-121	✗	✓	<u>19.7</u>	<u>17.4</u>	**93.9**
Maskomaly (ours)	Mask2Former [12]	Swin-L	✗	✗	**23.5**	19.6	91.7

AP and 1.5% FPR95. When considering our method's optimal configuration [opt] as described in Section 5.4, we even beat Max. Entropy by 1.1%. On FishyScapes Static, our method ranks second in AP and FPR95 from the methods that do not train with auxiliary data. Table 2 also includes the results of four methods on our new metric. Here, we chose F_1 because it balances between recall and precision. As it is nontrivial to define a reasonable threshold for F_1, we took the scores from Pascal VOC [17] and Cityscapes [13]. We can observe the informativeness of our new metric. Although Max. Entropy outperforms our method in terms of AP by 8.8%, the unoptimized version of our methods achieves a 41.9% and 26.7% higher score for $MDM_{F_1}^{60}$ and $MDM_{F_1}^{70}$ respectively.

Table 3 shows the quantitative results of our method compared to state-of-the-art methods on StreetHazards [30]. Our method beats all methods on AP but the version of DenseHybrid which is trained on StreetHazards and with auxiliary data. We achieve 4.2% higher AP than Max. Entropy which was also not trained on the StreetHazards training set. Our method also achieves 3.8% better AP than any method that does not train with auxiliary data even though these methods do not suffer from distribution shift.

Inference Speed: We measure the computational overhead of our unoptimized implementation on an Intel I7-13700K. For this, we average over 1000 runs of our method for the images from the validation set of SMIYC [7]. Our method takes 182ms on average.
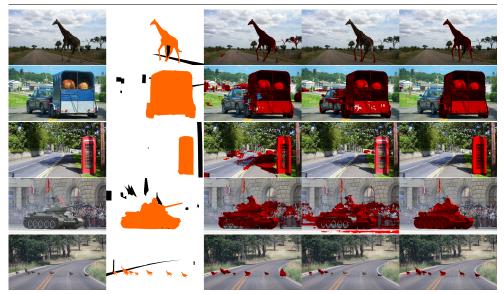
Figure 4: Qualitative results and comparison of our method on SMIYC [7] validation set. The first column shows the input. The second shows the ground truth, with the anomalies in orange and void in black. The last three columns show the predicted anomalies of Max. Entropy [7], ObsNet [1], and Maskomaly in red. We optimized the threshold to achieve the best F_1-score for each image individually and left out void zones.

5.3 Ablations

Individual Components: To understand the influence of the individual parts of the algorithm, we remove them one at a time and evaluate the performance of the ablated model. We also included a straightforward baseline that leverages Mask2Former outputs to emphasize more clearly the merit of Maskomaly. This baseline works by assigning

$$o_{\text{baseline}}[i,j] = \begin{cases} \max_{1 \le n \le N} m_n[i,j], & \text{if } \underset{1 \le l \le C+1}{\text{argmax}} \, p_{\text{argmax}_{1 \le n \le N} m_n[i,j]}[l] = C+1, \\ 1 - \max_{1 \le n \le N} m_n[i,j] & \text{otherwise.} \end{cases} \quad (4)$$

Intuitively speaking, the baseline checks for each pixel individually whether it belongs to an anomalous mask or not. After that, it assigns the probability of belonging to class $C+1$ or one minus the probability of belonging to its predicted inlier class.

Table 4 shows the performance of our ablated models and the baseline (Id 1) on the Road-Anomaly dataset and the test set of SMIYC. We can see that all ablated methods perform better than the baseline. Additionally, each individual component performs better than methods such as PEBAL [43] or DenseHybrid [24] without outlier supervision. Rejecting border regions reduces our FPR95 significantly. The largest step is achieved by combining the ideas of rejecting and accepting. This shows that the components complement each other well. Lastly, our mask-based approach allows us to act on masks as a coherent component. We gain significant AP and reduce the FPR95 greatly by initializing $\mathcal{I}$ with masks that we found to predict the ambiguous ground class [13], leading us to our complete Maskomaly method (Id 6). The impact of this initialization on RoadAnomaly [36] is especially high because all anomalies are located on streets or ground.

Fig. 5 shows a complete image of our performance. We can observe that the individual components have a lot less smooth curve than a method such as Max. Entropy. This is because we consider masks on a component level and possibly activate entire regions when increasing the threshold

Table 4: Results of our ablated models on RoadAnomaly [56] and SMIYC [7] test set. The individual components that are used in each version are ticked. The abbreviations we use are related to particular lines of code in Algorithm 1. Acc. refers to o_{accept} at lines 11–12, Rej. refers to o_{reject} at lines 5–6, Bord. refers to lines 7–10, and Init. $\mathcal{I}$ refers to lines 2–4. The best method is marked in bold.

					RoadAnomaly			SMIYC	
Method Id	Acc.	Rej.	Bord.	Init. $\mathcal{I}$	AP↑	FPR95↓	AUC↑	AP↑	FPR95↓
1	N/A	N/A	N/A	N/A	18.6	54.4	74.3	-	-
2	✓	✗	✗	✗	46.0	26.3	89.2	-	-
3	✗	✓	✗	✗	45.3	19.3	91.3	-	-
4	✗	✓	✓	✗	45.5	15.7	92.0	58.4	23.4
5	✓	✓	✓	✗	63.2	14.8	94.2	-	-
6	✓	✓	✓	✓	70.9	11.9	95.5	93.4	6.9

Table 5: Benchmark results on SMIYC [7] of all methods that are designed around mask-based segmentation networks. All methods use Mask2Former [12] as a backbone. The best method is marked in bold, and the second best is underlined. We separate methods that use auxiliary data during training.

Method	Aux. data	retrain	Anomaly Track				
			AP↑	FPR95↓	sIoU gt↑	PPV↑	mean F_1↑
EAM [25]	✓	✗	93.8	**4.1**	**67.1**	**53.8**	**60.9**
RbA [40]	✓	✓	**94.5**	4.6	64.9	47.5	51.9
EAM [25]	✗	✗	76.3	93.9	-	-	-
RbA [40]	✗	✓	<u>86.1</u>	<u>15.9</u>	**56.3**	<u>41.4</u>	<u>42.0</u>
Maskomaly (ours)	✗	✗	**93.4**	**6.9**	<u>55.4</u>	**51.5**	**49.9**

slightly. Maskomaly (Id 6) has a smoother curve and its graph lies completely above all individual component combinations (Id 2-5). Handling borders (Id 4) in addition to only rejecting areas (Id 3), only separate in the higher recall range. In that range, the borderless version loses by a small margin.

Method: To show the impact of our algorithm decoupled from its backbone we compare it to state-of-the-art methods paired with Mask2Former [12]. There are two other methods that are build around mask-based segmentation outputs. For a comparison against other methods which were fine-tuned on Mask2Former, see the discussion of Nayal et al. [40]. [2] In Table 5, we show the results of EAM [25], RbA [40] and our method on the SMIYC test set. We can clearly see that our method performs by far best when only considering the methods that did not resort to auxiliary data. Even though EAM and RbA gain much by training with anomalous data, our method can keep up in the two most important metrics AP and FPR95 as well as PPV. Especially with regard to AP, our method is only 1.1% behind the current best method.

5.4 Hyperparameter Study

Mask Count: In Section 5.2, we set T_{IoU} and computed $\mathcal{A}$ in advance for evaluating on the test set to ensure a fair comparison. In this section, we again rank the masks by IoU on the SMIYC validation set and evaluate how the test set scores change when we take the first n masks for $n = 1,...,16$. We stop at 16 because there is no relevant change anymore. Fig. 6 (left) shows that

---

[2]We did not include the results of RbA [40] in the other tables, as the results that RbA [40] reports for other methods deviate from what Grcic et al. [25] and we report.
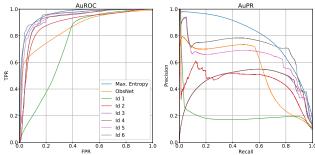
Figure 5: AuROC and AuPR curves of our ablated methods, ObsNet [1] and Max. Entropy [8] on RoadAnomaly [36]. Colors are consistent in the two plots and the Ids are the same as in Table 4.
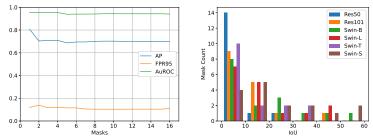

Figure 6: Left: AP, FPR95, and AuROC of Maskomaly for different mask counts on Road-Anomaly [36], right: mask counts within different ranges of IoU on SMIYC [7] validation.

for RoadAnomaly, the optimal configuration for AP would have included only one mask with 80.8% AP, and the optimal configuration for FPR95 would have included 9 with 10.37% FPR95. Nevertheless, it also shows that the method is robust with regard to the number of masks chosen, as the AP does not drop significantly.

Backbone: Although mask-based segmentation networks such as Mask2Former perform well on inlier classes of Cityscapes [12], we need to verify that they generally learn anomaly-predicting masks implicitly. To show this, we compute the IoU of the masks generated by all available pre-trained semantic segmentation backbones of Mask2Former with anomalies from a control set. We use the SMIYC validation set as the control set as it includes a diverse set of anomalies. Fig. 6 (right) shows the number of masks that fall into fixed ranges of IoU for all backbones. As in the previous investigation, we cut off our classification after 16 masks. Res101, Swin-S, Swin-B, and Swin-L have masks with a high IoU of more than 40%. This provides evidence that a variety of semantic segmentation backbones with large enough capacity implicitly learn masks for anomalies.

6 Conclusion

We have proposed Maskomaly, a simple yet effective anomaly segmentation method that builds upon pre-trained mask-based semantic segmentation networks. The method is easy to implement, requires no additional training, and sets the new state of the art among methods that do not train with auxiliary data. We have shown the merit of our method against state-of-the-art approaches, ablated its components, and showed the generality of our findings and method across different backbones. We have also introduced a novel metric that promotes robust anomaly segmentation methods in view of real-world scenarios. Our results demonstrate the potential of mask-based semantic segmentation networks for anomaly segmentation and provide insights for developing more practical and effective methods for anomaly segmentation in safety-critical systems such as autonomous cars.

References

[1] Victor Besnier, Andrei Bursuc, David Picard, and Alexandre 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*, pages 15701–15710, 2021.

[2] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In *Pattern Recognition: 41st DAGM German Conference, DAGM GCPR 2019, Dortmund, Germany, September 10–13, 2019, Proceedings 41*, pages 33–47. Springer, 2019.

[3] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Dense outlier detection and open-set recognition based on training with noisy negative images. *arXiv preprint arXiv:2101.09193*, 2021.

[4] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, 129:3119–3135, 2021.

[5] Jasmin Breitenstein, Andreas Bär, Daniel Lipinski, and Tim Fingscheidt. Detection of collective anomalies in images for automated driving using an earth mover's deviation (emdev) measure. In *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, pages 90–97. IEEE, 2021.

[6] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15333–15342, 2021.

[7] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[8] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5128–5137, 2021.

[9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[11] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pages 17864–17875, 2021.

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

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPR Workshop on the Future of Datasets in Vision*, volume 2. sn, 2015.

[14] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16918–16927, 2021.

[15] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–308, 2009.

[17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[18] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *Advances in Neural Information Processing Systems*, 2022.

[19] Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Séverine Dubuisson, and Isabelle Bloch. One versus all for deep neural network incertitude (ovnni) quantification. *arXiv preprint arXiv:2006.00954*, 2020.

[20] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[21] Matej Grcić and Siniša Šegvić. Hybrid open-set segmentation with synthetic negative data. *arXiv preprint arXiv:2301.08555*, 2023.

[22] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Dense open-set recognition with synthetic outliers generated by real nvp. *arXiv preprint arXiv:2011.11094*, 2020.

[23] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Dense anomaly detection by robust learning on synthetic negative data. *arXiv preprint arXiv:2112.12833*, 2021.

[24] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pages 500–517. Springer, 2022.

[25] Matej Grcić, Josip Šarić, and Siniša Šegvić. 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.

[26] Douglas M Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

[27] Florian Heidecker, Abdul Hannan, Maarten Bieshaar, and Bernhard Sick. Towards corner case detection by modeling the uncertainty of instance segmentation networks. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*, pages 361–374. Springer, 2021.

[28] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations, ICLR*, 2017.

[29] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations, ICLR*, 2018.

[30] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *International Conference on Machine Learning (ICML)*, 2022.

[31] Ivan Krešo, Josip Krapac, and Siniša Šegvić. Efficient ladder-style densenets for semantic segmentation of large images. *IEEE Transactions on Intelligent Transportation Systems*, 22(8):4951–4961, 2020.

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

[33] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *6th International Conference on Learning Representations, ICLR*, 2018.

[34] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Gmmseg: Gaussian mixture based generative semantic segmentation models. In *Advances in Neural Information Processing Systems*, 2022.

[35] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR*, 2018.

[36] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2152–2161, 2019.

[37] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[39] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations, ICLR*, 2019.

[40] Nazir Nayal, Mısra Yavuz, João F Henriques, and Fatma Güney. Pixels together strong: Segmenting unknown regions rejected by all. *arXiv preprint arXiv:2211.14293*, 2022.

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

[42] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations, ICLR*, 2020.

[43] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 246–263. Springer, 2022.

[44] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15651–15660, 2021.

[45] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[46] Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, pages 12427–12436. PMLR, 2021.

[47] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[48] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8865, 2019.