Real-Time Anomaly Segmentation for Road Scenes

Gabriele Lucca, Matteo Martini and Andrea Scaturro

Abstract

Real-Time Anomaly Segmentation for Road Scenes is a crucial task to ensure safety and reliability for autonomous driving applications and various branches of computer vision problems such as continual learning and open-world scenarios. However, the presence of anomalous objects not included during training poses a significant challenge for existing deep neural networks. In this paper, we present a thorough analysis of methodologies for real-time anomaly segmentation, focusing on the utilization of lightweight deep learning models. Using the Cityscapes dataset for training and various test datasets including Road Anomaly, Road Obstacle, and Fishyscapes, we assess the effectiveness of models such as ERFNet, ENet, and BiSeNet.Furthermore, we explore the impact of temperature scaling for model confidence calibration. Our experimental results demonstrate that the proposed models can effectively identify anomalies in real-time road scenes, making a way for new opportunities in autonomous driving and computer vision.

1. Introduction

In the era of automation and autonomous driving, road safety has become a fundamental priority. Anomaly segmentation in road scenes plays a critical role in ensuring the safety and efficiency of autonomous vehicles. Anomaly analysis, particularly anomaly segmentation, is an essential task in many areas of computer vision.

However, the definition and identification of anomalies can vary significantly depending on the context and goals of the application. Generally, an anomaly is an observation that deviates significantly from the norm or expected behavior. In the field of computer vision, anomaly analysis focuses on detecting and segmenting objects or patterns that do not match the surrounding context or the model's expectations.

Deep neural networks have demonstrated exceptional results in numerous computer vision tasks, including semantic segmentation. However, when deployed in real-world scenarios, these networks often perform poorly in addressing anomalous or out-of-distribution (OOD) objects. So this work aims to address this challenge by implementing real-time anomaly segmentation models using lightweight deep learning approaches. These models are designed to fit into embedded devices with limited memory resources and processing capacity.

Our goal is to provide an effective and efficient solution for anomaly segmentation in road scenes, while ensuring the practicality and scalability of our models. Through the use of the Cityscapes dataset for training and a series of test datasets, described later in the paper.

We will evaluate the performance of our proposed models and analyze their capabilities to identify and segment anomalous objects in real-time. [4]

2. Related Work

In this section we first review previous datasets for anomaly detection, with some of them being designed for road anomaly segmentation. Then we briefly describe some of the methods on anomaly and obstacle segmentation.

2.1. Datasets and Benchmarks

Cityscapes is a large dataset commonly used for training artificial vision models, especially for semantic segmentation and object detection in urban contexts. It contains a vast set of high-resolution images acquired in various cities, annotated with semantic labels for roads, sidewalks, buildings, vehicles, people, and other urban objects. It is a valuable resource for the development and evaluation of algorithms for urban scene analysis.

RoadObstacle dataset is specifically designed for detecting obstacles on roads. It contains images of various road and environmental conditions, with a particular emphasis on the presence of obstacles such as debris, abandoned vehicles, construction work, and other potential hazards. The images are annotated with detailed information about the obstacles present, enabling the development and evaluation of algorithms for detecting and classifying road obstacles.

RoadAnomaly is a dataset focused on identifying anomalies and non-standard conditions on roads. This dataset contains a wide range of images representing atypical situations such as potholes, cracks, damaged road signs, and other irregularities that can affect traffic safety and flow.

Detailed annotations provide information about the type and severity of anomalies present, supporting the development of algorithms for detecting and classifying anomalous road conditions.

Fishyscapes benchmark for anomaly segmentation consists of cut-and-paste anomalies from out-of-distribution domains. This is problematic, because the anomalies stand out as clearly unnatural in context. For instance, the orientation of anomalous objects is unnatural, and the lighting of the cut-and-paste patch differs from the lighting in the original image, providing an unnatural cue to anomaly detectors that would not exist for real anomalies. These anomalies are integrated into the scene with proper lighting and orientation, mimicking real-world anomalies and making them significantly more difficult to detect. [8]

2.2. Networks

ERFNet (Efficient Residual Factorized Network)

This architecture is a ConvNet for real-time and accurate semantic segmentation. The core element of our architecture is a novel layer design that leverages skip connections and convolutions with 1D kernels. This proposed block is stacked sequentially to build our encoder-decoder architecture, which produces semantic segmentation end-to-end in the same resolution as the input. [] It uses factorized convolutions to reduce the number of parameters and computational complexity, making it suitable for implementations on devices with limited resources. It also incorporates residual blocks that allow the network to learn deep representations without excessively increasing parameters. It stands out for its computational efficiency. [5]

ENet (Efficient Neural Network)

This architecture is designed to perform large-scale computations in a much faster and more efficient manner, which might lead to significant savings.[paper enet] It uses a combination of separable deep convolutions and traditional deep convolutions to reduce the number of parameters while maintaining learning capacity. However, its asymmetric architecture may limit its ability to learn deep representations compared to deeper architectures, especially in complex scenarios. [2]

BiSeNet v1 (Bilateral Segmentation Network)

This architecture is proposed in this paper to improve the speed and accuracy of real-time semantic segmentation simultaneously. Our proposed BiSeNet contains two paths: Spatial Path (SP) and Context Path (CP). The Spatial Path is designed to preserve the spatial information from original images. And the Context Path utilizes the lightweight model and global average pooling to obtain sizeable receptive field rapidly.[1] Thanks to its hierarchical structure, it can acquire contextual information at various scales, improving the network's ability to understand and segment complex scenes. However, it may require more computational resources compared to more efficient architectures like ERFNet and ENet, making it less suitable for implementations on embedded devices or in real-time scenarios. [1]

2.3. Methods

In this section, we describe the methods utilized for semantic image segmentation in our study. Semantic segmentation is a crucial task in computer vision, aiming to assign semantic labels to each pixel in an image, thereby partitioning the image into meaningful regions. We present three distinct methods: Maximum Softmax Probability (MSP), Max Logit, and Max Entropy, each offering unique approaches to pixel-wise class prediction.

2.3.1 Introductionally formulas

Before delving into the detailed explanations of each method, let us define some key formulas:

1. **Softmax Function**:

$$\text{softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$$

2. **Entropy**:

$$H(p) = -\sum_{i=1}^{C} p_i \log(p_i)$$

Where z represents the logits produced by the neural network, C is the number of classes, and p is the probability distribution outputted by the softmax function.

2.3.2 MSP (Maximum Softmax Probability)

MSP is a method employed in semantic image segmentation to determine the predicted class for each pixel based on the maximum probability obtained through the softmax function. Mathematically, for each pixel in the image, the class with the highest probability predicted by the neural network is selected.

$$\text{MSP}(x, y) = \arg\max_{c} \left[\text{softmax}(z(x, y))_c\right]$$

Where x and y represent the pixel coordinates, $z(x, y)$ is the vector of logits for pixel (x, y), and c denotes the class index. [7]

2.3.3 Max Logit

Max Logit entails selecting the predicted class for each pixel based on the maximum value among the outputs of the neural network, without the application of the softmax function. Essentially, the class with the highest logit (i.e., the output of the neural network before softmax) is directly chosen for each pixel.

$$\text{Max Logit}(x, y) = \arg\max_c \left[z(x, y)_c\right]$$

Where x and y represent the pixel coordinates, $z(x, y)$ is the vector of logits for pixel (x, y), and c denotes the class index. [7]

2.3.4 Max Entropy

Max Entropy relies on the entropy of the probability distribution obtained from the softmax function. For each pixel in the image, the class that maximizes the entropy of the probability distribution is selected, indicating the class with the highest uncertainty in prediction.

$$\text{Max Entropy}(x, y) = \arg\max_c \left[H(\text{softmax}(z(x, y)))\right]$$

Where x and y represent the pixel coordinates, $z(x, y)$ is the vector of logits for pixel (x, y), H denotes the entropy function, and c denotes the class index.

These methods serve as crucial components in our semantic segmentation pipeline, each offering distinct advantages and considerations in pixel-wise class prediction. [6]

3. Loss Functions

In the training process of our neural network model, the choice of an appropriate loss function plays a crucial role in optimizing the network's parameters. In this section, we present the loss functions employed in our experiments.

3.1. Cross Entropy Loss

Cross entropy loss, also known as negative log likelihood loss, is a commonly used loss function for classification tasks. It measures the dissimilarity between the predicted probability distribution and the actual distribution of the target classes. Mathematically, it is defined as:

$$\text{CrossEntropyLoss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(p_{ij}),$$

where N is the number of samples, C is the number of classes, y_{ij} is an indicator function (1 if sample i belongs to class j, 0 otherwise), and p_{ij} is the predicted probability of sample i belonging to class j.

3.2. Logit Normalization Loss

The Logit Normalization Loss is designed to address the issue of class imbalance in classification tasks. It normalizes the logits of each class to mitigate the impact of heavily weighted classes on the training process. The loss function is defined as:

$$\text{LogitNormalizationLoss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\frac{e^{z_{ij}}}{\sum_{k=1}^{C} e^{z_{ik}}},$$

where z_{ij} represents the logit for class j of sample i.

3.3. Focal Loss

Focal Loss is particularly effective for addressing class imbalance and the issue of easy and hard examples in classification tasks. It introduces a modulating factor to down-weight the loss assigned to well-classified examples, thereby focusing more on the difficult examples. The formula for Focal Loss is given by:

$$\text{FocalLoss} = -\frac{1}{N}\sum_{i=1}^{N}(1 - p_{i,y_i})^{\gamma}\log(p_{i,y_i}),$$

where p_{i,y_i} is the predicted probability of the correct class for sample i, and γ is a focusing parameter that controls the rate at which easy examples are down-weighted.

3.4. Enhanced Isotropy Maximization Loss

Enhanced Isotropy Maximization Loss is a novel loss function designed to enhance the isotropy of feature representations learned by neural networks. By encouraging feature vectors to have similar magnitudes, this loss function aims to improve the generalization performance of the model. The Enhanced Isotropy Maximization Loss is defined as:

$$\text{EIMLoss} = \frac{1}{N}\sum_{i=1}^{N}(1 - \frac{z_i^T z_i}{\|z_i\|^2}),$$

where z_i is the feature vector of sample i.

These loss functions were employed in our experiments to train and optimize our neural network models for the given tasks.

4. Experiments

On pre-trained ERF-Net model

We leveraged a pre-trained ERF-Net model, trained on the Cityscapes dataset. This model served as a baseline for our experiments. We have conduct various anomaly inferences using this pre-trained model alongside the provided anomaly segmentation test dataset. Additionally, code for the Maximum Softmax Probability (MSP) method was provided, while resources for implementing Max-Entropy and Max-Logit methods were referenced [3] [4].

		SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
Method	**mIoU**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**
MSP	72.20%	14.5854	95.0901	0.7207	94.7685	0.2678	95.8130	1.9816	95.2585	9.4270	95.3010
MaxLogit	72.20%	13.1934	97.0150	1.1526	86.8155	0.2139	96.4409	1.6453	96.4615	8.7080	93.7640
MaxEntropy	72.20%	14.3098	96.7160	0.8318	94.0838	0.2374	96.8160	1.9527	94.0524	9.1070	95.3126

Table 1. Evaluation on pre-trained ERF-Net model

		SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
Method	**mIoU**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**
MSP	72.20%	14.5854	95.0901	0.7208	94.7685	0.2678	95.8130	1.9816	95.2585	9.4271	95.3010
MSP(t = 0.5)	72.20%	14.6749	95.0522	0.6988	94.8856	0.2793	95.3960	2.0169	95.1832	9.6064	95.1719
MSP(t = 0.75)	72.20%	14.6263	95.0723	0.7100	94.8268	0.2733	95.6029	1.9968	95.2266	9.5094	95.2439
MSP(t = 1.1)	72.20%	14.5705	95.0976	0.7249	94.7454	0.2658	95.8959	1.9765	95.2674	9.3979	95.3194
MSP(best t)	72.20%	14.5705 (t=e-04)	95.0976 (t=e-04)	0.7249 (t=0.01)	94.7454 (t=0.01)	0.2658 (t=e-04)	95.8959 (t=e-04)	1.9765 (t=e-06)	95.2674 (t=e-06)	9.3979 (t=e-05)	95.3194 (t=e-05)

Table 2. Studying the effect of temperature scaling

		SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
Method	**mIoU**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**
ENet	58.60%	14.5938	95.1338	0.6718	95.1184	0.2750	95.2100	2.0936	94.8344	9.7372	95.0344
ERF-Net	71.62%	14.7575	94.4175	0.8005	94.3740	0.2533	96.0658	2.4871	92.0669	9.0139	94.9432
BiSeNet	43.76%	23.4244	95.5065	2.7489	99.8951	0.3649	95.1364	1.9310	89.9655	12.6710	96.0435

Table 3. Evaluation on different nets

Studying the effect of temperature scaling

Temperature scaling, a method for confidence calibration, was employed to potentially enhance the anomaly segmentation capabilities of the network. Our objective in this stage was to determine the optimal temperature value that yields the most effective anomaly segmentation results during inference. [3]

Implementing the void classifier

The Cityscapes dataset encompasses 19 known category classes along with a void (background) class. We treated the void class as an anomaly and proceeded to train ENet and BiSeNet networks accordingly. Subsequently, during anomaly inference, only the output of the void class was considered, serving as another baseline for our experimentation. We fine-tuned using pre-trained model classes with 20 epochs.

Studying the effect of training loss function

In this stage, we investigated the impact of specific loss functions tailored for anomaly detection on the training process. We wanted to analyze the effects of the Enhanced Isotropy Maximization Loss and Logit Normalization Loss when employed individually and in conjunction with focal loss and cross-entropy loss. Unfortunally, the Enhanced Isotropy Maximization Loss was not suitable for our problem (due to the fact that it applies to a linear block not present in the ERF-Net network) so we discarded it. For each loss function and available method, we conducted 50 epochs.

Studying the effect of pruning and quantization

Exploring avenues to reduce model size and latency, we explored pruning and quantization techniques applied to the trained models. The experiment results encompassed metrics such as mean Intersection over Union (mIOU), Floating Point Operations per Second (FLOPS), the number of trainable model parameters, alongside the performance of anomaly segmentation. We trained for 50 epochs, and every 10 epochs, we zeroed out the least influential 10% of the weights.

5. Results

From the results obtained, we can say that BiSeNet is not an optimal network for this task because it requires more computational resources compared to the other networks, making it less suitable for implementations on embedded devices or in real-time scenarios. ENet has been specifically designed for embedded systems and prefers to maintain excellent speed over precision. ERF-Net is the best network because it is an excellent compromise between precision and efficiency in solving the problem (Table 3).

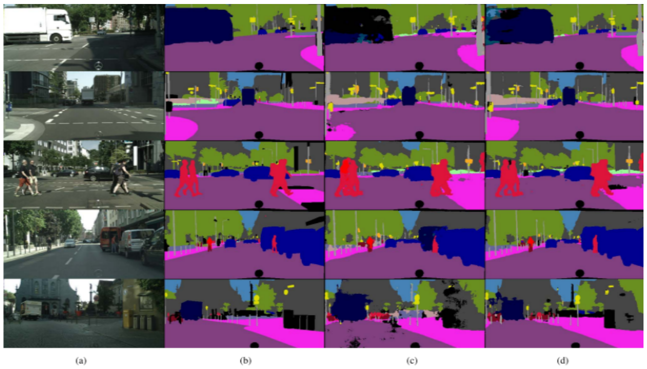

Figure 1. Qualitative examples of the segmentation produced by our architecture ERFNet (d) compared to the ground truth labels (b) and ENet. (a) Input image. (b) Ground truth. (c) ENet. (d) ERFNet. [5]

		SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
Method	**mIoU**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**
MSP	69.33%	13.9849	95.6219	0.5752	96.6891	0.2520	96.1628	2.2168	92.6406	10.2030	94.7492
MaxEntropy	69.33%	13.3582	99.0708	0.4758	99.0867	0.2489	95.4395	2.1685	91.8139	10.4629	92.2233
MaxLogit	69.33%	15.5002	99.5639	0.4730	99.4214	0.2567	97.3257	1.8476	93.8357	10.3386	94.4091
Void	69.33%	20.1531	91.8372	0.8406	95.2570	0.5042	94.0883	1.8079	93.7700	10.3610	94.4803

Table 4. Evaluation on Focal Loss

		SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
Method	**mIoU**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**
MSP	66.90%	15.8091	94.6036	0.6797	95.1398	0.3146	94.3897	2.2872	93.8782	11.0029	94.2432
MaxEntropy	66.90%	16.3835	96.4486	0.6523	98.9810	0.3467	92.5962	2.3397	93.03570	11.6699	91.0048
MaxLogit	66.90%	15.2272	99.2131	0.8284	99.5484	0.8284	99.5484	3.06358	97.4154	12.4759	89.9208
Void	66.90%	21.5682	93.8824	6.8911	90.6028	0.4673	94.0769	1.7123	96.2121	13.4112	93.8581

Table 5. Evaluation on Logit Normalization Loss

		SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
Method	**mIoU**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**
MSP	59.47%	14.8609	94.9472	0.7203	94.7894	0.2722	95.4418	3.0833	92.3422	10.7781	94.5128
MaxEntropy	59.47%	15.9636	93.6252	0.8186	94.6547	0.2643	92.8552	3.4454	88.4774	12.5977	91.3643
MaxLogit	59.47%	16.1298	97.5778	1.0886	91.9565	0.2458	88.0840	3.6319	88.9562	12.6721	83.7857
Void	59.47%	15.3866	94.5524	0.8424	94.5999	0.2467	95.3784	2.6237	91.9282	11.8665	93.9225

Table 6. Evaluation on Cross Entropy Loss, Focal Loss and Logit Normalization Loss joint togheter

		SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly	
Method	**mIoU**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**	**AuPRC**	**FPR95**
MSP	41.73%	16.6829	94.1869	0.6766	94.9483	0.3299	94.5960	2.0814	90.3079	9.6481	91.5028
MaxEntropy	41.73%	17.8747	92.0277	0.8583	84.4113	0.3398	90.4530	2.0185	85.1242	12.9739	72.6023
MaxLogit	41.73%	18.3697	96.3537	0.8337	83.3651	0.4480	90.0608	1.5364	89.4773	13.9256	73.5770
Void	41.73%	18.2554	93.5158	0.7617	94.0725	0.3549	91.3122	1.3890	96.3515	12.3362	84.9841

Table 7. Evaluation on the pruned ERF-Net

With MaxEntropy, we expect to achieve the best results given its probabilistic nature. However, due to the small size of the proposed networks, the obtained result falls short of expectations. In contrast, MSP shows a slightly better fit for the proposed problem, likely due to its simplicity and suitability for small-scale architectures. Moreover, despite expectations, MaxLogit yields the poorest results, underscoring the importance of considering the characteristics of the task and the model's capacity(Table 1). Additionally, the application of temperature scaling does not significantly improve the performance, indicating the complexity of enhancing the network's predictive capabilities within the constraints of its architecture(Table 2).

Despite the use of a limited number of training epochs, additional loss optimization techniques yielded results comparable to those obtained using the ERF-Net network with Cross Entropy. This suggests that increased computational power and longer training durations could lead to superior performance compared to the default loss network (Tables 4-5-6).

However, it should be noted that while pruning is an option for reducing network complexity, it is discouraged in this context as the network itself is characterized by a limited number of parameters. Additionally, although post-training quantization can improve network latency, it is important to consider that this technique results in decreased precision, which is why it is applied after training (Table 7).

6. Conclusion

The segmentation of anomalies in road scenes is a crucial task to ensure the safety and reliability of autonomous driving applications and various computer vision problems. In this work, we presented a thorough analysis of methodologies for real-time anomaly segmentation, focusing on the use of lightweight deep learning models. Using the Cityscapes dataset for training and various test datasets including Road Anomaly, Road Obstacle, and Fishyscapes, we evaluated the effectiveness of models such as ERFNet, ENet, and BiSeNet. Additionally, we explored the impact of temperature scaling for model confidence calibration. Our experimental results demonstrate that the proposed models can effectively identify anomalies in road scenes in real-time, opening new opportunities for autonomous driving and computer vision.

Despite using a small number of training epochs, implementing additional loss optimization techniques produced comparable results to those achieved by the erfnet network with crossentropy. This implies that increasing computational resources and extending training durations could enhance performance beyond that of the default loss network.

However, it's worth noting that pruning to reduce network complexity is discouraged here due to the network's inherently limited parameter count. Additionally, post-training quantization can enhance network latency but at the cost of reduced precision, hence it's typically applied after training.

Unfortunately, despite the efforts made, the results obtained with the erfnet network were not satisfactory due to the combination of a limited dataset and a small number of available training epochs. Nevertheless, this study provides a solid foundation for future developments, suggesting that further analysis with a larger number of epochs and a broader dataset could lead to significant improvements in network performance.

To view the code, please check our GitHub repository: https://github.com/GabrieleLucca/AnomalySegmentation_Project4.

References

[1] Chao Peng Changxin Gao Gang Yu Nong Sang Changqian Yu, Jingbo Wang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. 2018. arXiv:1808.00897 [cs.CV]. 2

[2] Chao Peng Changxin Gao Gang Yu Nong Sang Changqian Yu, Jingbo Wang. Enet: A deep neural network architecture for real-time semantic segmentation. 2018. 10.1109/TITS.2017.2750080. 2

[3] Yu Sun Kilian Q. Weinberger Chuan Guo, Geoff Pleiss. On calibration of modern neural networks. 2017. arXiv:1706.04599v2 [cs.LG]. 3, 4

[4] Mantas Mazeika Andy Zou Joe Kwon Mohammadreza Mostajabi Jacob Steinhardt Dawn Song Dan Hendrycks, Steven Basart. Scaling out-of-distribution detection for real-world settings. 2022. arXiv:1911.11132v4 [cs.CV]. 1, 3

[5] Luis M. Bergasa Roberto Arroyo Eduardo Romera, José M. Álvarez. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. 2011. arXiv:1808.00897 [cs.CV]. 2, 5

[6] Roland Siegwart Giancarlo Di Biase, Hermann Blum and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. 2021. arXiv:2103.05445v1 [cs.CV]. 3

[7] Matthias Rottmann Robin Chan, Faculty of Mathematics Hanno Gottschalk IZMD, and University of Wuppertal Natural Sciences. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. 2021. arXiv:2012.06575v2 [cs.CV]. 2, 3

[8] Svenja Uhlemeyer Hermann Blum Sina Honari Roland Siegwart Pascal Fua Mathieu Salzmann Matthias Rottmann Robin Chan, Krzysztof Lis. Segmentmeifyoucan: A benchmark for anomaly segmentation. 2021. arXiv:2104.14812v2 [cs.CV]. 2