This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Unmasking Anomalies in Road-Scene Segmentation

Shyam Nandan Rai[1], Fabio Cermelli[1,2], Dario Fontanel[1], Carlo Masone[1], Barbara Caputo[1]
[1]Politecnico di Torino, [2]Italian Institute of Technology
first.last@polito.it

Abstract

Anomaly segmentation is a critical task for driving applications, and it is approached traditionally as a per-pixel classification problem. However, reasoning individually about each pixel without considering their contextual semantics results in high uncertainty around the objects' boundaries and numerous false positives. We propose a paradigm change by shifting from a per-pixel classification to a mask classification. Our mask-based method, Mask2Anomaly, demonstrates the feasibility of integrating an anomaly detection method in a mask-classification architecture. Mask2Anomaly includes several technical novelties that are designed to improve the detection of anomalies in masks: i) a global masked attention module to focus individually on the foreground and background regions; ii) a mask contrastive learning that maximizes the margin between an anomaly and known classes; and iii) a mask refinement solution to reduce false positives. Mask2Anomaly achieves new state-of-the-art results across a range of benchmarks, both in the per-pixel and component-level evaluations. In particular, Mask2Anomaly reduces the average false positives rate by 60% w.r.t. the previous state-of-the-art.

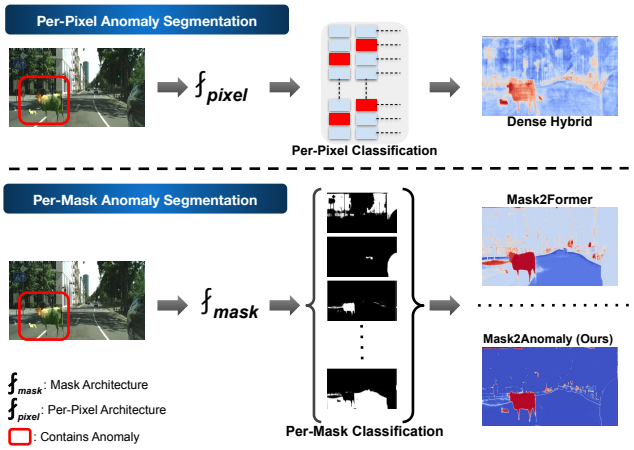

Figure 1: **Per-pixel vs per-mask Anomaly Segmentation:** Dense Hybrid [22], the state-of-the-art method for AS based on per-pixel classification can detect the anomalies, but it produces many false positives. Anomaly segmentation can be cast as a mask classification problem, but naively using MSP [25] on top of Mask2Former [12] does not produce good results. Our Mask2Anomaly exploits mask-transformers properties to refine the classification of anomalies, drastically reducing false positives. f_{pixel} and f_{mask} denotes per-pixel, and per-mask architecture. Anomalies in the output image are represented in red.

1. Introduction

Semantic segmentation [14, 41, 50, 48, 42] plays a pivotal role in self-driving cars, being used to obtain a detailed understanding of the surroundings of a vehicle. Generally, semantic segmentation models are trained to recognize a pre-defined set of semantic classes (*e.g.* car, pedestrian, road, etc.); however, in real-world applications, they may encounter objects not belonging to such categories (*e.g.* animals or cargo dropped on the road). Therefore, it is essential for these models to identify objects in a scene that are not present during training *i.e. anomalies*, both to avoid potential dangers and to enable continual learning [37, 8, 17, 7] and open-world solutions [6].

Anomaly segmentation (AS) [3, 47, 20, 27] addresses this problem, *i.e.* it aims to segment objects from classes that were absent during training. Existing AS methods are built upon the idea of individually classifying the pixels and assigning to each of them an anomaly score. This score may be given by a pixel-level discriminative method [1, 27, 22, 43], by estimating the uncertainty of the individual pixel predictions [39], or by comparing the per-pixel discrepancy between the original image and a synthetic image generated from the semantic predictions [34, 45, 46]. However, reasoning on the pixels individually produces noisy anomaly scores, thus leading to a high number of false positives and poorly localized anomalies (see Fig. 1).

In this paper, we address this problem by casting AS as a mask classification task rather than a pixel classification. This idea stems from the recent advances in mask-

transformer architectures [12, 13], which demonstrated that it is possible to achieve remarkable performance across various segmentation tasks by classifying masks, rather than pixels. We hypothesize that mask-transformer architectures are better suited to detect anomalies than per-pixel architectures [11, 26], because masks encourage objectness and thus can capture anomalies as whole entities, leading to more congruent anomaly scores and reduced false positives. To enable the segmentation of anomalies at the mask level, we revisit the Maximum Softmax Probability (MSP) [25], a classic method used in per-pixel AS, and apply it to the masks produced by a mask-transformer model. However, the effectiveness of such an approach hinges on the model's capability to output masks that capture well anomalies and we found that naively using MSP on top of the best mask-transformer architecture [12] does not yield good results (see Fig. 1). Hence, we propose several technical contributions to improve the capability of mask-transformer architectures to capture anomalies and reject false positives in driving scenes (see Fig. 1):

- At the **architectural** level, we propose a global masked-attention mechanism that allows the model to focus on both the foreground objects and on the background while retaining the efficiency of the original masked-attention [12].
- At the **training** level, we have developed a mask contrastive learning framework that utilizes outlier masks from additional out-of-distribution data to maximize the separation between anomalies and known classes.
- At the **inference** level, we propose a mask-based refinement solution that reduces false positives by filtering masks based on the panoptic principle [28] that distinguishes between "things" and "stuff".

We integrate these contributions on top of the mask architecture [12] and term this solution **Mask2Anomaly**. To the best of our knowledge, Mask2Anomaly is the first demonstration of an AS method that detects anomalies at the mask level. We tested Mask2Anomaly on standard anomaly segmentation benchmarks for road scenes (Road Anomaly [34], Fishyscapes [4], Segment Me If You Can [9]), achieving the best results among all AS methods by a significant margin. In particular, Mask2Anomaly reduces on average the false positives rate by more than half w.r.t. the previous state-of-the-art. Code and pre-trained models will be made publicly available upon acceptance.

2. Related Work

Mask-based semantic segmentation. Traditionally, semantic segmentation methods [35, 11, 52, 32, 51] have adopted fully-convolutional encoder-decoder architectures [35, 2] and addressed the task as a dense classification problem. However, transformer architectures have recently caused us to question this paradigm due to their outstanding performance in closely related tasks such as object detection [5] and instance segmentation [23]. In particular, [13] proposed a mask-transformer architecture that addresses segmentation as a mask classification problem. It adopts a transformer and a per-pixel decoder on top of the feature extraction. The generated per-pixel and mask embeddings are combined to produce the segmentation output. Building upon [13], [12] introduced a new transformer decoder adopting a novel masked-attention module and feeding the transformer decoder with one pixel-decoder high-resolution feature at a time.

So far, all these mask-transformers have been considered exclusively in a closed set setting, i.e, there are no unknown categories at test time. To the best of our knowledge, Mask2Anomaly is the first method that performs AS directly with mask-transformers, thus empowering these approaches with the capability to recognize anomalies in real-world settings.

Anomaly segmentation methods can be broadly divided into three categories: (a) Discriminative, (b) Generative, and (c) Uncertainty-based methods. *Discriminative Methods* are based on the classification of the model outputs. Hendrycks and Gimpel [25] established the initial AS discriminative baseline by applying a threshold over the maximum softmax probability (MSP) that distinguishes between in-distribution and out-of-distribution data. Other approaches use auxiliary datasets to improve performance [31, 27, 43] by calibrating the model over-confident outputs. Alternatively, [30] learns a confidence score by using the Mahalanobis distance, and [10] introduces an entropy-based classifier to discover out-of-distribution classes. Recently, discriminative methods tailored for semantic segmentation [4] directly segment anomalies in embedding space. In contrast, [22] proposes a hybrid approach that combines the known class posterior, dataset posterior, and an un-normalized data likelihood to estimate anomalies. *Generative Methods* provides an alternative paradigm to segment anomalies based on generative models [34, 16, 46, 45]. These approaches train generative networks to reconstruct anomaly-free training data and then use the generation discrepancy to detect an anomaly at test time. All the generative-based methods heavily rely on the generation quality and thus experience performance degradation due to image artifacts [20]. Finally, *Uncertainty based* methods segment anomalies by leveraging uncertainty estimates via Bayesian neural networks [39].

All the methods discussed above are based on per-pixel classification architectures and score the pixels individually without considering local semantics, leading to noisy anomaly predictions and many false positives. Mask2Anomaly overcomes this limitation by segmenting anomalies as semantically clustered masks, encouraging the

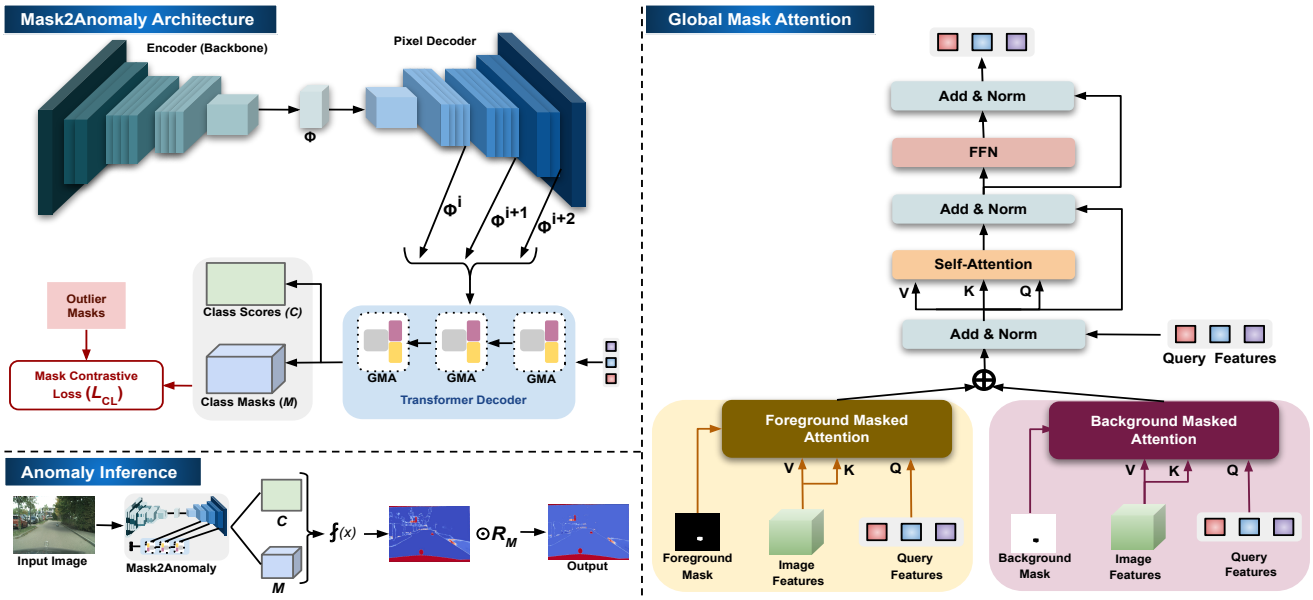

Figure 2: **Mask2Anomaly Overview.** Mask2Anomaly meta-architecture consists of an encoder, a pixel decoder, and a transformer decoder. We propose global mask attention (Sec. 3.2) that independently distributes the attention between foreground and background. V, K, and Q are Value, Key, and Query. ϕ is image features. $\phi^i, \phi^{i+1}, \phi^{i+2}$ are upsampled image features at multiple scales. Mask contrastive Loss L_{CL} (Sec. 3.3) utilizes outlier masks to maximize the separation between anomalies and known classes. During anomaly inference, we utilize refinement mask R_M (Sec. 3.4) to minimize false positives.

objectness of the predictions. To the best of our knowledge, this is the first work to use masks to score anomalies.

3. Method

In this section, we begin by introducing the problem-setting, followed by describing a generic mask-transformer architecture for anomaly segmentation. Next, we delve into our Mask2Anomaly architecture and its novel elements.

3.1. Preliminaries

Let us denote with $\mathcal{X} \subset \mathbb{R}^{3\times H\times W}$ the space of RGB images, where H and W are the height and width, respectively, and with $\mathcal{Y} \subset \mathbb{N}^{K\times H\times W}$ the space of semantic labels that associate each pixel in an image to a semantic category from a predefined set $\mathcal{K}$, with $|\mathcal{K}| = K$. At training time we assume to have a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{D}$, where $x_i \in \mathcal{X}$ is an image and $y_i \in \mathcal{Y}$ is its ground truth semantic mask. The goal for an anomaly segmentation model is to learn a function f that maps the image space to an anomaly score space, *i.e.* $f : \mathcal{X} \mapsto \mathbb{R}^{H\times W}$. For traditional semantic segmentation architectures based on per-pixel classification [11], the function f can be obtained in various ways, for example, applying the *Maximum Softmax Probability* (MSP) [25] on top of the per-pixel classifier. Formally, given the pixel-wise class scores $S(x) \in [0, 1]^{K\times H\times W}$ obtained by segmenting the image x with a per-pixel architecture, we compute the anomaly score as:

$$f(x) = 1 - \max_{k=1}^{K}(S(x)). \tag{1}$$

In this paper, we propose to adapt this framework based on MSP to mask-transformer segmentation architectures. We recall that the mask classification problem is formulated as a direct set prediction task with the goal of producing a fixed-size set of N predictions [5]. Based on this idea, the mask classification meta-architecture for semantic segmentation consists of three parts: a) a *backbone* that acts as feature extractor, b) a *pixel-decoder* that upsamples the low-resolution features extracted from the backbone to produce high-resolution *per-pixel embeddings*, and c) a *transformer decoder*, made of L transformer layers, that takes the image features to output a fixed number of object queries consisting of *mask embeddings* and their associated *class scores* $C \in \mathbb{R}^{N\times K}$. The final *class masks* $M \in \mathbb{R}^{N\times(H\times W)}$ are obtained by multiplying the mask embeddings with the per-pixel embeddings. The mask-transformer is trained using a combination of binary cross-entropy loss and dice loss [38] for the class masks and cross-entropy loss for the class scores, unlike per-pixel architecture that is trained only on cross-entropy loss (more details on these losses are given in the supplementary material).

Given such a mask-transformer architecture, we propose

Input Image Attention Map Ground Truth

Figure 3: **Limitation of Mask-Attention:** Masked-attention [12] selectively attends to foreground regions resulting in low attention scores (dark regions) for anomalies. Anomalies are in red. Best viewed with zoom.

to calculate the anomaly scores for an input x as

$$f(x) = 1 - \max_{k=1}^{K} \left(\text{softmax}(C)^T \cdot \text{sigmoid}(M)\right). \quad (2)$$

Here, $f(x)$ utilizes the same marginalization strategy of class and mask pairs as [13] to get anomaly scores. Without loss of generality, we implement the anomaly scoring (Eq. (2)) on top of the Mask2Former [12] architecture. However, this strategy hinges on the ability of the masks predicted by the segmentation architecture to capture anomalies well. We found that simply applying the MSP on top of Mask2Former as in Eq. (2) does not yield good results (see Fig. 1 and the results in Sec. 4.2). To overcome this problem, we introduce improvements in the architecture, training procedure, and anomaly inference mechanism. We name our method Mask2Anomaly, and its overview is shown in Fig. 2 (left). In the rest of the sections, we will discuss in detail the technical novelties of Mask2Anomaly.

3.2. Global Masked Attention

One of the key ingredients to Mask2Former [12] state-of-the-art segmentation results is the replacement of the *cross-attention* (CA) layer in the transformer decoder with a *masked-attention* (MA). The masked-attention attends only to pixels within the foreground region of the predicted mask for each query, under the hypothesis that local features are enough to update the query object features. The output of the l-th masked-attention layer can be formulated as

$$\text{softmax}(\mathcal{M}_l^F + QK^T)V + X_{in} \quad (3)$$

where $X_{in} \in \mathbb{R}^{N \times C}$ are the N C-dimensional query features from the previous decoder layer. The input queries $Q \in \mathbb{R}^{N \times C}$ are obtained by linearly transforming the query features with a learnable transformation whereas the keys and values K, V are the image features under learnable linear transformations $f_k(.)$ and $f_v()$. Finally, $\mathcal{M}_l^F$ is the predicted foreground attention mask that at each pixel location (i, j) is defined as

$$\mathcal{M}_l^F(i,j) = \begin{cases} 0 & \text{if } M_{l-1}(i,j) \geq 0.5 \\ -\infty & \text{otherwise,} \end{cases} \quad (4)$$

where M_{l-1} is the output mask of the previous layer.

By focusing only on the foreground objects, masked-attention grants faster convergence and better semantic segmentation performance than cross-attention. However, focusing only on the foreground region constitutes a problem for anomaly segmentation because anomalies may also appear in the background regions. Removing background information leads to failure cases in which the anomalies in the background are entirely missed, as shown in the example in Fig. 3. To ameliorate the detection of anomalies in these corner cases, we extend the masked attention with an additional term focusing on the background region (see Fig. 2, right). We call this a *global masked-attention* (GMA) formally expressed as

$$\begin{aligned} X_{out} =& \text{softmax}(\mathcal{M}_l^F + QK^T)V \\ &+ \text{softmax}(\mathcal{M}_l^B + QK^T)V + X_{in} \end{aligned} \quad (5)$$

where $\mathcal{M}_l^B$ is the additional background attention mask that complements the foreground mask $\mathcal{M}_l^F$, and it is defined at the pixel coordinates (i, j) as

$$\mathcal{M}_l^B(i,j) = \begin{cases} 0 & \text{if } M_{l-1}(i,j) < 0.5 \\ -\infty & \text{otherwise.} \end{cases} \quad (6)$$

The global masked-attention in Eq. (5) differs from the masked-attention by additionally attending to the background mask region, yet it retains the benefits of faster convergence w.r.t. the cross-attention.

3.3. Mask Contrastive Learning

The ideal characteristic of an anomaly segmentation model is to predict high anomaly scores for out-of-distribution (OOD) objects and low anomaly scores for in-distribution (ID) regions. Namely, we would like to have a significant margin between the likelihood of known classes being predicted at anomalous regions and vice-versa. A common strategy used to improve this separation is to fine-tune the model with auxiliary out-of-distribution (anomalous) data as supervision [21, 22, 4].

Here we propose a contrastive learning approach to encourage the model to have a significant margin between the anomaly scores for in-distribution and out-of-distribution classes. Our mask-based framework allows us to straightforwardly implement this contrastive strategy by using as supervision outlier images generated by cutting anomalous objects from the auxiliary OOD data and pasting it on top of the training data. For each outlier image, we can then generate a binary outlier mask M_{OOD} that is 1 for out-of-distribution pixels and 0 for in-distribution class pixels. With this setting, we first calculate the negative likelihood of in-distribution classes using the class scores C and class masks M as:

$$l_N = -\max_{k=1}^{K} \left(\text{softmax}(C)^T \cdot \text{sigmoid}(M)\right) \quad (7)$$

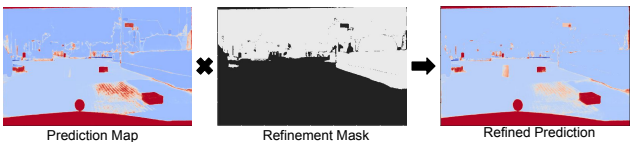

Figure 4: **Mask Refinement Illustration:** To obtain the refined prediction, we multiply the prediction map with a refinement mask that is built by assigning zero anomaly scores for pixels that are categorized as "stuff", except for the "road". The refinement eliminates many false positives at the boundary of objects and in the background. The region to be masked is white in the refinement mask.

Ideally, for pixels corresponding to in-distribution classes l_N should be -1 since the value of softmax$(C)^T$ and sigmoid(M) would be close to 1. On the other hand, for an anomalous pixel, sigmoid(M) is ideally 0 as M contains only inlier classes mask that results in l_N to be 0. Using l_N, we define our contrastive loss as:

$$L_{CL} = \frac{1}{2}(l_{CL}^2),$$
$$l_{CL} = \begin{cases} l_N & \text{if} M_{OOD} = 0 \\ max(0, m - l_N) & \text{otherwise,} \end{cases} \tag{8}$$

where the margin m is a hyperparameter that decides the minimum distance between the out-of-distribution and in-distribution classes.

3.4. Refinement Mask

False positives are one of the main problems in anomaly segmentation, particularly around object boundaries. Handcrafted methods such as iterative boundary suppression [27] or dilated smoothing have been proposed to minimize the false positives at boundaries or globally, however, they require tuning for each specific dataset. Instead, we propose a general refinement technique that leverages the capability of mask transformers [12] to perform all segmentation tasks. Our method stems from the panoptic perspective [28] that the elements in the scene can be categorized as *things*, *i.e.* countable objects, and *stuff*, *i.e.* amorphous regions. With this distinction in mind, we observe that in driving scenes, i) unknown objects are classified as things, and ii) they are often present on the road. Thus, we can proceed to remove most false positives by filtering out all the masks corresponding to "stuff", except the "road" category. We implement this removal mechanism in the form of a binary refinement mask $R_M \in [0,1]^{H \times W}$, which contains zeros in the segments corresponding to the unwanted "stuff" masks and one otherwise. Thus, by multiplying R_M with the predicted anomaly scores f we filter out all the unwanted "stuff" masks and eliminate a large portion of the false positives (see Fig. 4). Formally, for an image x the refined anomaly scores f^r is computed as:

$$f^r(x) = R_M \odot f(x), \tag{9}$$

where $\odot$ is the Hadamard product.

R_M is the dot product between the binarized output mask $\bar{M} \in \{0,1\}^{N \times (H \times W)}$ and the class filter $\bar{C} \in \{0,1\}^{1 \times N}$, *i.e.* $R_M = \bar{C} \cdot \bar{M}$. We define $\bar{M} =$ sigmoid$(M) > 0.5$ and the class filter $\bar{C}$ is equal to 1 only where the highest class score of softmax(C) belongs to "things" or "road" classes and is greater than 0.95.

4. Experiments

Dataset: We train Mask2Anomaly on Cityscapes [14] and for evaluation we use Road Anomaly [34], Fishyscapes [3] and Segment Me If You Can (SMIYC) benchmarks [9].
Road Anomaly: is a collection of 60 web images having anomalous objects located on or near the road.
Fishyscapes (FS): consists of two datasets, Fishyscape static (FS static) and Fishyscapes lost & found (FS lost & found). Fishyscape static is built by blending Pascal VOC [19] objects on Cityscapes images containing 30 validation and 1000 test images. Fishyscapes lost & found is based on a subset of the Lost and Found dataset [40], with 100 validation and 275 test images.
SMIYC: consists of two datasets, RoadAnomaly21 (SMIYC-RA21) and RoadObstacle21 (SMIYC-RO21). The SMIYC-RA21 contains 10 validation and 100 test images with diverse anomalies. The SMIYC-RO21 is collected with a focus on segmenting road anomalies and has 30 validation and 327 test images.

Evaluation Metrics: We evaluate all the anomaly segmentation methods at pixel and component levels. For pixel-wise evaluation, we use Area under the Precision-Recall Curve (AuPRC) and False Positive Rate at a true positive rate of 95% (FPR_{95}). Since pixel-level evaluation metrics can neglect small anomalies and be biased towards anomalies with large sizes, we also include component-level evaluations using the averaged component-wise F1 ($F1^*$), the positive predictive value (PPV), and the component-wise intersection over union (sIoU). Further, details of all the metrics can be found in the supplementary material.

Implementation Details: Our implementation is derived from [13, 12]. We use a ResNet-50 [24] encoder, and its weights are initialized from a model that is pre-trained with barlow-twins [49] self-supervision on ImageNet [15]. We freeze the encoder weights during training, saving memory and training time. We use a multi-scale deformable attention Transformer (MSDeformAttn) [53] as the pixel decoder. The MSDeformAttn gives features maps at $1/8, 1/16$, and $1/32$ resolution, providing image features to the transformer decoder layers. Our transformer decoder is adopted from [12] and consists of 9 layers with

Methods	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly		Average	
	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95}↓
Max Softmax [25](ICLR'17)	27.97	72.02	15.72	16.6	1.77	44.85	12.88	39.83	15.72	71.38	14.81	48.93
Entropy [25](ICLR'17)	-	-	-	-	2.93	44.83	15.4	39.75	16.97	71.1	11.66	51.89
Mahalanobis [30](NeurIPS'18)	20.04	86.99	20.9	13.08	-	-	-	-	14.37	81.09	18.42	60.38
Image Resynthesis [34](ICCV'19)	52.28	25.93	37.71	4.7	5.7	48.05	29.6	27.13	-	-	31.32	26.45
Learning Embedding [4](IJCV'21)	37.52	70.76	0.82	46.38	4.65	24.36	57.16	13.39	-	-	26.18	45.43
Void Classifier [4](IJCV'21)	36.61	63.49	10.44	41.54	10.29	22.11	4.5	19.4	-	-	15.46	36.63
JSRNet [45](ICCV'21)	33.64	43.85	28.09	28.86	-	-	-	-	**94.4**	**9.2**	52.04	47.3
SML [27](ICCV'21)	46.8	39.5	3.4	36.8	31.67	21.9	52.05	20.5	17.52	70.7	30.28	37.88
SynBoost [16](CVPR'21)	56.44	61.86	71.34	3.15	43.22	15.79	72.59	18.75	38.21	64.75	56.36	32.86
Maximized Entropy [10](ICCV'21)	85.47	15.00	85.07	0.75	29.96	35.14	86.55	8.55	48.85	31.77	67.18	18.24
Dense Hybrid [22](ECCV'22)	77.96	**9.81**	87.08	0.24	**47.06**	**3.97**	80.23	5.95	31.39	63.97	64.74	16.79
PEBEL [43](ECCV'22)	49.14	40.82	4.98	12.68	44.17	7.58	92.38	1.73	45.1	44.58	47.15	31.47
Mask2Anomaly (Ours)	**88.7**	14.60	**93.3**	**0.20**	46.04	4.36	**95.20**	**0.82**	79.70	13.45	**80.59**	**6.68**

Table 1: **Pixel level evaluation:** On average, Mask2Anomaly shows significant improvement among the compared methods. Higher values for AuPRC are better, whereas for FPR_{95} lower values are better. The best and second best results are **bold** and underlined, respectively. '-' indicates the unavailability of benchmark results.

Methods	SMIYC RA-21			SMIYC RO-21		
	sIoU ↑	PPV ↑	$F1^*$↑	sIoU ↑	PPV ↑	$F1^*$↑
Max Softmax [25](ICLR'17)	15.48	15.29	5.37	19.72	15.93	6.25
Ensemble [29](NurIPS'17)	16.44	20.77	3.39	8.63	4.71	1.28
Mahalanobis [30](NeurIPS'18)	14.82	10.22	2.68	13.52	21.79	4.70
Image Resynthesis [34](ICCV'19)	39.68	10.95	12.51	16.61	20.48	8.38
MC Dropout [39](CVPR'20)	20.49	17.26	4.26	5.49	5.77	1.05
Learning Embedding [4](IJCV'21)	33.86	20.54	7.90	35.64	2.87	2.31
SML [27](ICCV'21)	26.00	24.70	12.20	5.10	13.30	3.00
SynBoost [16](CVPR'21)	34.68	17.81	9.99	44.28	41.75	37.57
Maximized Entropy [10](ICCV'21)	49.21	39.51	28.72	47.87	62.64	48.51
JSRNet [45](ICCV'21)	20.20	29.27	13.66	18.55	24.46	11.02
Void Classifier [4](IJCV'21)	21.14	22.13	6.49	6.34	20.27	5.41
Dense Hybrid [22](ECCV'22)	54.17	24.13	31.08	45.74	50.10	50.72
PEBEL [43](ECCV'22)	38.88	27.20	14.48	29.91	7.55	5.54
Mask2Former [12]	25.20	18.20	15.30	5.00	21.90	4.80
Mask2Anomaly (Ours)	**60.40**	**45.70**	**48.60**	**61.40**	**70.30**	**69.80**

Table 2: **Component level evaluation:** Mask2Anomaly achieves large improvement on component level evaluation metrics among the baselined methods. Higher values of sIoU, PPV, and $F1^*$ are better. The best and second best results are **bold** and underlined, respectively.

100 queries. We train Mask2Anomaly using a combination of binary cross-entropy loss and the dice loss [38] for class masks and cross-entropy loss for class scores. The network is trained with an initial learning rate of 1e-4 and batch size of 16 for 90 thousand iterations on AdamW [36] with a weight decay of 0.05. We use an image crop of 380×760 with large-scale jittering [18] along with a random scale ranging from 0.1 to 2.0.

Next, we train the Mask2Anomaly in a contrastive setting. We generate the outlier image using AnomalyMix [43] where we cut an object from MS-COCO [33] dataset image and paste them on the Cityscapes image. The corresponding binary mask for an outlier image is created by assigning 1 to the MS-COCO image area and 0 to the Cityscapes image area. We randomly sample 300 images from the MS-COCO dataset during training to generate outliers. We train the network for 4000 iterations with m as 0.75, a learning rate of 1e-5, and batch size 8, keeping all the other hyperparameters the same as above. The probability of choosing an outlier in a training batch is kept at 0.2.

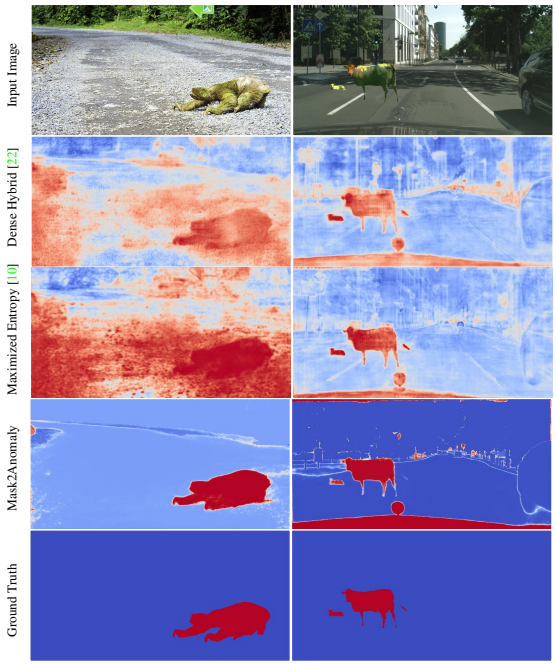

Figure 5: **Qualitative Results**: We observe that per-pixel classification architectures: Dense Hybrid [22] and Maximized Entropy [10] suffer from large false positives, whereas Mask2Anomaly, which is a mask-transformer, shows accurate pixel-wise anomaly segmentation results.

4.1. Main Results

Table 1 shows the pixel-level anomaly segmentation results achieved by Mask2Anomaly and recent SOTA methods on Fishyscapes, SMIYC, and Road Anomaly datasets. We can observe that Mask2Anomaly significantly improves the average AuPRC by 20% and the FPR_{95} by 60% compared to the second-best method. We observe that anomaly segmentation methods based on per-pixel architecture, such as JSRNet, perform exceptionally well on the Road Anomaly dataset. However, JSRNet does not generalize well on other datasets. On the other hand, Mask2Anomaly

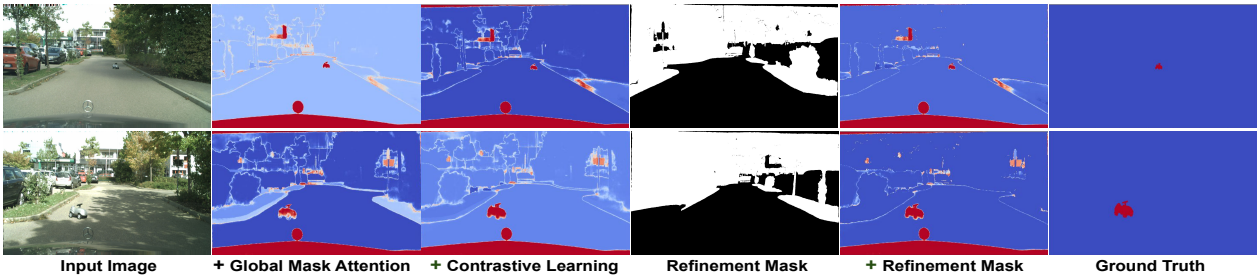

Figure 6: **Mask2Anomaly Qualitative Ablation**: demonstrates the performance gain by progressively adding (left to right) proposed components. Masked-out regions by refinement mask are shown in white. Anomalies are represented in red.

yields excellent results on all the datasets. Moreover, the property of our mask architecture to encourage objectness, rather than individual pixel anomalies, not only reduces the false positive but also improves the localization of whole anomalies. Indeed, Tab. 2 demonstrates that Mask2Anomaly outperforms all the baselined methods on component-level evaluation metrics. To conclude, Mask2Anomaly yields state-of-the-art anomaly segmentation performance both in pixel and component metrics.

Qualitative results: To get a better understanding of the visual results, in Fig. 5 we visually compare the anomaly scores predicted by Mask2Anomaly and its closest competitors: Dense Hybrid [22] and Maximized Entropy [10]. The results from both: Dense Hybrid and Maximized Entropy exhibit a strong presence of false positives across the scene, particularly on the boundaries of objects ("things") and regions ("stuff"). On the other hand, Mask2Anomaly demonstrates precise segmentation of anomalies while at the same time having minimal false positives. Additional qualitative results are in the supplementary material.

Segmentation results: Another critical characteristic of any anomaly segmentation method is that it should not disrupt the in-distribution classification performance, or else it would make the semantic segmentation model unusable. We find that adding only GMA to the base model boosts in-distribution accuracy to 80.45 on the validation set of Cityscapes. The final Mask2Anomaly model maintains an in-distribution accuracy of 78.88 mIoU, which is still 1.46 points higher than the vanilla Mask2Former. Moreover, it is important to note that both Mask2Anomaly and Mask2Former are trained for 90k iterations, indicating that, although Mask2Anomaly additionally attends to the background mask region, it shows convergence similar to Mask2Former. Extended quantitative and qualitative segmentation results with both Mask2Anomaly and Mask2Former are presented in the supplementary material.

4.2. Ablations

All the results reported in this section are from the Fishyscapes lost and found validation dataset.

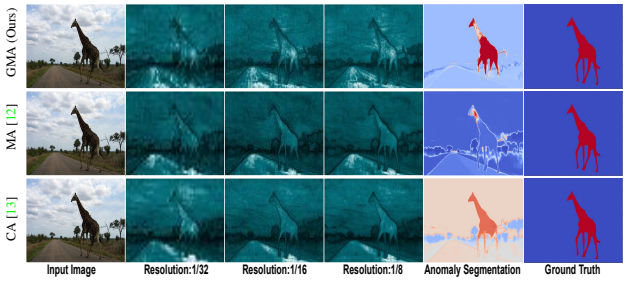

Figure 7: **Visualization of negative attention maps and results:** Global mask attention gives high attention scores to anomalous regions across all resolutions showing the best anomaly segmentation results among the compared attention mechanisms. Cross-attention performs better than mask-attention but has high false positives and low confidence prediction for the anomalous region. Darker regions represent low attention values. Details to calculate negative attention are given in Section:4.2.

Mask2Anomaly: Table 3(a) presents the results of a component-wise ablation of the technical novelties included in Mask2Anomaly. We use Mask2Former as the baseline. As shown in the table, removing any individual component from Mask2Anomaly drastically reduces the results, thus proving that their individual benefits are complimentary. In particular, we observe that the global masked attention has a big impact on the AuPRC and the contrastive learning is very important for the FPR_{95}. The mask refinement brings further improvements to both. Figure 6 visually demonstrates the positive effect of all the components.

Global Mask Attention: To better understand the effect of the global masked attention (GMA), in Tab. 3(c), we compare it to the masked-attention (MA) [12] and cross-attention (CA) [44]. We can observe that although the MA increases the mIoU w.r.t. the CA, it degrades all the metrics for anomaly segmentation, thus confirming our preliminary experiment shown in Fig. 3. On the other hand, the GMA provides improvements across all the metrics. This is confirmed visually in Fig. 7, where we show the negative attention maps for the three methods at different resolutions. The

GMA	CL	RM	AuPRC↑	FPR_{95}↓
			10.60	*89.35*
✓		✓	35.05	87.11
	✓	✓	57.23	31.93
✓	✓		68.95	24.07
✓	✓	✓	**69.41**	**9.46**

(a)

margin(m)	AuPRC↑	FPR_{95}↓
1	65.37	11.61
0.95	65.40	12.20
0.90	66.05	13.49
0.80	66.20	14.89
0.75	**69.41**	**9.46**
0.50	62.07	13.26

(b)

	mIoU↑	AuPRC↑	FPR_{95}↓
CA [13]	76.43	20.30	89.35
MA [12]	77.42	10.60	89.39
GMA	**80.45**	**32.35**	**25.95**

(c)

	AuPRC↑	FPR_{95}↓
w/o Refinement Mask	68.95	24.07
$L_{\{things \setminus road\}}$	67.04	39.11
$L_{\{stuff \setminus road\}}$	**69.41**	**9.46**

(d)

Batch Outlier Probability	AuPRC↑	FPR_{95}↓
0.1	63.01	14.66
0.2	**69.41**	**9.46**
0.5	69.20	11.03
1	68.77	10.53

(e)

Table 3: **Mask2Anomaly Ablation tables:** **(a)** Component-wise ablation of Mask2Anomaly. Results in *italics* show Mask2Former results. GMA: Global Mask Attention, CL: Contrastive Learning, and RM: Refinement Mask. **(b)** Shows the behavior of L_{CL} on different margin(m) values. We empirically find the best results when m is 0.75. **(c)** Global masked attention (GMA) performs best among various attention mechanisms: Cross-Attention (CA) and Masked-Attention (MA). **(d)** We show the performance gain by using a refinement mask that masks the $\{stuff \setminus road\}$ regions as anomalies are categorized as $things$ class. **(e)** Batch outlier probability is the likelihood of selecting an outlier image for a batch during contrastive training. Best result is achieved at 0.2 probability. (*All the results reported on FS Lost & Found validation set*).

negative attention is calculated by averaging all the queries (since there is no reference known object) and then subtracting one. Note that the GMA has a high response on the anomaly (the giraffe) across all resolutions.

Refinement Mask: Table 3(d) shows the performance gains due to the refinement mask. We observe that filtering out the {"stuff" \ "road"} regions of the prediction map improves the FPR_{95} by 14.61 along with marginal improvement in AuPRC. On the other hand, removing the {"things" \ "road"} regions degrades the results, confirming our hypothesis that anomalies are likely to belong to the "things" category. Figure 6 qualitatively shows the improvement achieved with the refinement mask. Also, refinement mask adds a small overhead of 1.12 GFlops compared to Mask2Anomaly 258 GFlops inference cost.

Mask Contrastive Learning: We tested the effect of the margin in the contrastive loss L_{CL}, and we report these results in Tab. 3(b). We find that the best results are achieved by setting m to 0.75, but the performance is competitive for any value of m in the table. Similarly, we tested the effect of the batch outlier probability, which is the likelihood of selecting an outlier image in a batch. The results shown in Tab. 3(e) indicate that the best performance is achieved at 0.2, but the results remain stable for higher values of the batch outlier probability.

Effect of bigger backbones: We demonstrate the efficacy of Mask2Anomaly by comparing it to the vanilla Mask2Former but using larger backbones. The results in Tab. 4 show that despite the disadvantage, Mask2Anomaly with a ResNet-50 still performs better than Mask2Former using large transformer-based backbones. It is also important to note that the number of training parameters for Mask2Anomaly can be reduced to $23M$ by using a frozen self-supervised pre-trained encoder, which is significantly less than all the Mask2Former variations.

Method	Backbone	AuPRC↑	FPR_{95} ↓	FLOPs↓	Training ↓ Parameters
Mask2Former [12]	ResNet-50	10.60	89.35	**226G**	44M
	ResNet-101	9.11	45.83	293G	63M
	Swin-T	24.54	37.98	232G	42M
	Swin-S	30.96	36.78	313G	69M
Mask2Anomaly‡	ResNet-50	**32.35**	**25.95**	258G	**23M**

Table 4: **Architectural Efficiency of Mask2Anomaly:** Mask2Anomaly outperforms best Mask2Former architecture having Swin-S backbone with only 30% trainable parameters. Mask2Anomaly‡ only uses global mask attention.

5. Conclusion

In this work, we present Mask2Anomaly, a novel anomaly segmentation architecture established on masked architecture. Mask2Anomaly contains global mask attention specifically designed to improve the attention mechanism for anomaly segmentation tasks. Next, we develop a mask contrastive learning framework that utilizes outlier masks to maximize the separation between anomalies and known classes. Finally, we introduced mask refinement that reduces false positives and improves the overall performance. We show the efficacy of Mask2Anomaly and its components through extensive qualitative and quantitative results. We hope Mask2Anomaly will open doors for new anomaly segmentation methods based on mask architecture.

References

[1] Matt Angus, Krzysztof Czarnecki, and Rick Salay. Efficacy of pixel-level ood detection for semantic segmentation. *arXiv preprint arXiv:1911.02897*, 2019. 1

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12):2481–2495, 2017. 2

[3] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 5

[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(11):3119–3135, 2021. 2, 4, 6

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 3

[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. 1

[7] Fabio Cermelli, Dario Fontanel, Antonio Tavera, Marco Ciccone, and Barbara Caputo. Incremental learning in semantic segmentation from image labels. In *CVPR*, pages 4371–4381, 2022. 1

[8] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *CVPR*, 2020. 1

[9] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Mathieu Salzmann, Pascal Fua, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. *arXiv preprint arXiv:2104.14812*, 2021. 2, 5

[10] 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. 2, 6, 7

[11] 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. 2, 3

[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. 1, 2, 4, 5, 6, 7, 8

[13] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2, 4, 5, 7, 8

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 5

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[16] 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. 2, 6

[17] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *CVPR*, 2021. 1

[18] Xianzhi Du, Barret Zoph, Wei-Chih Hung, and Tsung-Yi Lin. Simple training strategies and model scaling for object detection. *arXiv preprint arXiv:2107.00057*, 2021. 6

[19] 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(2):303–338, 2010. 5

[20] Dario Fontanel, Fabio Cermelli, Massimiliano Mancini, and Barbara Caputo. Detecting anomalies in semantic segmentation with prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–121, 2021. 1, 2

[21] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Dense open-set recognition with synthetic outliers generated by real nvp. *arXiv preprint arXiv:2011.11094*, 2020. 4

[22] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *European Conference on Computer Vision*, pages 500–517. Springer, 2022. 1, 2, 4, 6, 7

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[25] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 1, 2, 3, 6

[26] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[27] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in

urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15425–15434, 2021. 1, 2, 5, 6

[28] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 2, 5

[29] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 6

[30] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 2, 6

[31] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 2

[32] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 2

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[34] 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. 1, 2, 5, 6

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[37] Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1114–1124, 2021. 1

[38] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 3, 6

[39] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 1, 2, 6

[40] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1099–1106. IEEE, 2016. 5

[41] Ruoqi Sun, Xinge Zhu, Chongruo Wu, Chen Huang, Jianping Shi, and Lizhuang Ma. Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4355–4364, 2019. 1

[42] Antonio Tavera, Fabio Cermelli, Carlo Masone, and Barbara Caputo. Pixel-by-pixel cross-domain alignment for few-shot semantic segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1959–1968, 2022. 1

[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 *European Conference on Computer Vision*, pages 246–263. Springer, 2022. 1, 2, 6

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 7

[45] 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. 1, 2, 6

[46] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[47] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L. Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *Computer Vision – ECCV 2020*, pages 145–161, 2020. 1

[48] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4084–4094, 2020. 1

[49] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 5

[50] Junyi Zhang, Ziliang Chen, Junying Huang, Liang Lin, and Dongyu Zhang. Few-shot structured domain adaptation for virtual-to-real scene parsing. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 9–17, 2019. 1

[51] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *ECCV*, 2018. 2

[52] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2

[53] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5