Mask2Anomaly: Mask Transformer for Universal Open-set Segmentation

Mask2Anomaly: Mask Transformer for Universal Open-Set Segmentation

Shyam Nandan Rai, Fabio Cermelli, Barbara Caputo, and Carlo Masone, *Member, IEEE*

***Abstract*—Segmenting unknown or anomalous object instances is a critical task in autonomous 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 a mask-classification architecture to jointly address anomaly segmentation, open-set semantic segmentation, and open-set panoptic segmentation. Mask2Anomaly includes several technical novelties that are designed to improve the detection of anomalies/unknown objects: 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; iii) a mask refinement solution to reduce false positives; and iv) a novel approach to mine unknown instances based on the mask- architecture properties. By comprehensive qualitative and qualitative evaluation, we show Mask2Anomaly achieves new state-of-the-art results across the benchmarks of anomaly segmentation, open-set semantic segmentation, and open-set panoptic segmentation.**

***Index Terms*—Anomaly segmentation, open-set semantic segmentation, open-set panoptic segmentation, mask architecture.**

I. Introduction

IMAGE segmentation [16], [54], [56], [61], [64] plays a significant role in self-driving cars, being instrumental in achieving a detailed understanding of the vehicle's surroundings. Generally, 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 [9], [10], [20], [46] and open-world solutions [8]. The segmentation of unseen object categories can be performed at three levels of increasing semantic output information (see Fig. 1):

- *Anomaly segmentation* (AS) [6], [23], [35], [60] focuses on segmenting objects from classes that were absent during training, generating an output map that identifies the anomalous image pixels.
- *Open-set semantic segmentation* (OSS) [27] evaluates a segmentation model's performance on both anomalies and known classes. OSS ensures that when training an anomaly segmentation model, its performance on known classes remains unaffected.
- *Open-set panoptic segmentation* (OPS) [34] simultaneously segments distinct instances of unknown objects and performs panoptic segmentation [36] for the known classes.

In the literature, AS, OSS and OPS are typically addressed separately using specialized networks for each task. These networks rely on per-pixel classification architectures that individually classify the pixels and assign to each of them an anomaly score. However, reasoning on the pixels individually without any spatial correlation produces noisy anomaly scores, thus leading to a high number of false positives and poorly localized anomalies or unknown objects (see Fig. 2).

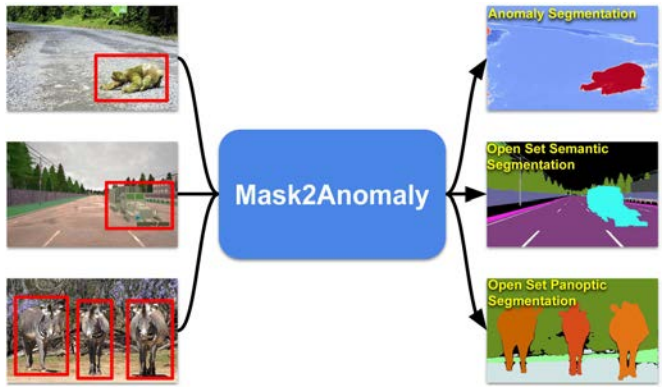

Fig. 1. *Mask2Anomaly*: We present a mask-based architecture that can jointly perform open-set semantic segmentation, open-set panoptic segmentation, and anomaly segmentation. In the figure, the objects enclosed in red boxes are anomaly/unknown.

Manuscript received 19 August 2023; revised 9 May 2024; accepted 15 June 2024. Date of publication 27 June 2024; date of current version 5 November 2024. This work was supported in part by European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)–MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D.1555 11/10/2022, PE00000013 - CUP: E13C22001800001) through Project FAIR - Future Artificial Intelligence Research and in part by European Lighthouse on Secure and Safe AI–ELSA, HORIZON EU under Grant 101070617. Recommended for acceptance by C. Fowlkes. *(Corresponding authors: Shyam Nandan Rai.)*

Shyam Nandan Rai, Barbara Caputo, and Carlo Masone are with the Politecnico di Torino, 10129 Torino, Italy (e-mail: shyam.rai@polito.it; barbara.caputo@polito.it; carlo.masone@polito.it).

Fabio Cermelli is with Focoos AI, 10125 Turin, Italy.

The code and pre-trained models are available: https://github.com/shyam671/Mask2Anomaly-Unmasking-Anomalies-in-Road-Scene-Segmentation/tree/main.

Digital Object Identifier 10.1109/TPAMI.2024.3419055

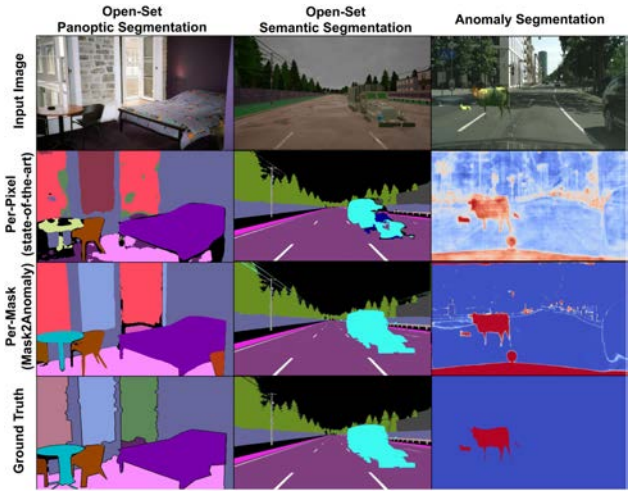

Fig. 2. *Per-pixel versus per-mask architecture:* We show significant shortcomings in the performance of state-of-the-art methods employing per-pixel architectures for anomaly segmentation or open-set segmentation tasks. These methods prediction have significant false positives and noisy outcomes. Mask2Anomaly(ours), an architecture based on mask-transformer properties that effectively addresses both anomaly segmentation and open-set segmentation tasks, leading to a substantial reduction in false positives and enhancing overall prediction quality.

In this paper, we propose to jointly address AS, OSS, and OPS with a single architecture (with minor changes during inference) by casting them as a mask classification task rather than a pixel classification task (see Fig. 1). The idea of employing mask-based architecture stems from the recent advances in mask-transformer architectures [14], [15], 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 because masks encourage objectness and thus can capture anomalies as whole entities, leading to more congruent anomaly scores and reduced false positives. However, the effectiveness of mask-transformer architectures hinges on the capability to output masks that captures anomalies well. Hence, we propose several technical contributions to improve the capability of mask-transformer architectures to capture anomalies or unknown objects and minimize false positives:

- 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 [14].
- 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, for anomaly segmentation, we propose a mask-based refinement solution that reduces false positives by filtering masks based on the panoptic segmentation that distinguishes between "things" and "stuff" and for open-set panoptic segmentation, we developed an approach to mine unknown instances based on mask-architecture properties.

We integrate these contributions on top of the mask architecture [14] and term this solution *Mask2Anomaly*. A few concurrent works have proposed methods that segment anomalies at the mask level by relying on the Mask2Former architecture [1], [28], [49], [52] but, to the best of our knowledge, Mask2Anomaly is the first universal architecture that jointly addresses AS, OSS, and OPS. We tested Mask2Anomaly on standard anomaly segmentation benchmarks (Road Anomaly [42], Fishyscapes [6], Segment Me If You Can [11], Lost&Found [51]), open-set semantic segmentation benchmark (Streethazard [31]), and open-set panoptic MS-COCO [34] dataset, achieving the best results among all methods for all task by a significant margin. All the code and pre-trained models are available here.

This work is an extension of our previous paper [52] that was accepted to ICCV 2023 (Oral) with the following contributions:

- We extend Mask2Anomaly to open-set segmentation tasks, namely open-set semantic segmentation and open-set panoptic segmentation.
- For the open-set panoptic segmentation task, we developed a novel approach to mine unknown instances based on the properties of the mask-architecture and provide related ablation studies to show its efficacy.
- Extensive qualitative and quantitative experiments demonstrate that Mask2Anomaly is an effective approach to address open-set segmentation tasks. Notably, Mask2 Anomaly gives a significant gain of 30% on Open-IoU metrics w.r.t best existing method.
- We extend Mask2Anomaly experimentation for the anomaly segmentation task by showing results on the Lost&Found dataset. Also, we show global mask attention can positively impact semantic segmentation by investigating its generalizability to other datasets.

II. Related Work

Mask-based semantic segmentation. Traditionally, semantic segmentation methods [13], [40], [44], [65], [66] have adopted fully-convolutional encoder-decoder architectures [2], [44] to address 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 [7] and instance segmentation [29]. In particular, Cheng et al. [15] proposed a mask-transformer architecture called MaskFormer 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 MaskFormer [15], the same authors later introduced a new transformer decoder called Mask2Former [14] adopting a novel masked-attention module and feeding the transformer decoder with one pixel-decoder high-resolution feature at a time.

While these mask-transformers have been originally considered exclusively in a closed-set setting, i.e, when there are

no unknown categories at test time, a few recent works have demonstrated their application also for AS in driving scenes [1], [28], [49], [52], thus empowering them with the capability to recognize anomalies in real-world setting.

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 [32] 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 [35], [39], [57], by calibrating the model over-confident outputs. Along this line of works, Zhang et al. [63] demonstrate that the auxiliary data may be combined with style transfer techniques. Alternatively, Lee et al. [38] learns a confidence score by using the Mahalanobis distance, and Chan et al. [12] introduces an entropy-based classifier to discover out-of-distribution classes. Recently, discriminative methods tailored for semantic segmentation [6] directly segment anomalies in embedding space. *Generative Methods* provides an alternative paradigm to segment anomalies based on generative models [18], [42], [59], [60]. 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 [23]. Grcić et al. [27] combine aspects of both generative and discriminative methods in a hybrid algorithm. Finally, *uncertainty-based* methods segment anomalies by leveraging uncertainty estimates via Bayesian neural networks [48].

Mask-based anomaly segmentation. The literature of anomaly segmentation has been dominated for a long time by methods that rely on per-pixel classification architectures to individually classify the pixels and assign to each of them an anomaly score [12], [18], [27], [35], [57], [59]. Few recent methods have challenged this structure by re-formulating the process of anomaly scores at the mask level rather than the level of individual pixels. This new methodology, enabled by universal segmentation architecture like MaskFormer [15] and Mask2Former [14], has led to solutions that achieve better anomaly localization and reduce the number of false positives, setting new standards over all the anomaly segmentation benchmarks. Originally, the concept of mask-based anomaly segmentation was introduced in Mask2Anomaly [52], a method that leverages Mask2Former to score anomalies on masks. Mask2Anomaly, further introduces a global-masked attention mechanism to attend to anomalies both in the background and foregound, a mask-refinement strategy tailored for driving scenes and a mask-contrastive loss to improve AS capabilities. Concurrently, Nayal et al. [49] leverages the same Mask2Former architecture and introduces a novel per-mask outlier scoring function, called RbA, based on the observation that the object queries in mask classification tend to behave like one versus all classifiers. Both Mask2Anomaly and RbA leverage outlier exposure to improve predictions. Alternatively, Maskomaly [1] uses a simple inference-time post-processing step in mask architecture to segment anomalies without any outlier exposure. This post-processing strategy combines two separate predictions, one that assigns low scores within inlier masks or across their shared borders, and another that assigns high probability to pixels included in masks which were found to predict anomalies on a validation set. Finally, Grcić et al. [28] experimentally show the advantages of using plain mask-based architecture for AS, and propose a formulation that combines the uncertainties of pixel-level mask assignment and mask-level recognition. This formulation leads to a new scoring function, denoted as EAM, that is given by the ensemble over all the anomaly scores (obtained for example via a max-logit detector) of mask-wide predictions. This method also uses negative examples to enhance performance.

Although these concurrent approaches leverage the same Mask2Former architecture, they implement subtly different engineering choices (e.g., in the encoder size or training data) that make their results not directly comparable. In this work, we establish a fair comparison among these new solutions, by using a common architectural and training data setup. Moreover, we fully exploit the universal capabilities of the Mask2Former architecture by extending Mask2Anomaly to perform AS, OSS, and OPS. To the best of our knowledge, this is the first mask-based method to do so.

Open-set segmentation is the task of segmenting both the anomalies and in-distribution classes for a given image. Anomaly segmentation methods [33], [60] can be adapted to perform open-set semantic segmentation by fusing the in-distribution segmentation results. However, these methods show poor performance in open-set metrics because their in-distribution class segmentation capabilities degrade after training for anomaly segmentation. Bevandic et al. [3] formally introduces the problem of open-set semantic segmentation that uses multi-task model segment anomaly and predicts semantic segmentation maps. Later, Bevandic et al. [4] improved the prior method using noisy outlier labels. Recently, Grcic et al. [27] proposed a hybrid approach that combines the known class posterior, dataset posterior, and an un-normalized data likelihood to estimate anomalies and in-distribution classes simultaneously. Another challenging problem in the space of open-set segmentation is open-set panoptic segmentation [34]. In open-set panoptic segmentation, the goal is to simultaneously segment distinct instances of unknown objects and perform panoptic segmentation for in-distribution classes. Hwang et al. [34] proposed an exemplar-based open-set panoptic segmentation network (EOPSN) that is based on exemplar theory and utilizes Panoptic FPN [36] which is a per-pixel architecture to perform open-set panoptic segmentation.

All the methods discussed so far for anomaly and open-set segmentation rely on per-pixel classification and evaluate individual pixels without considering local semantics. This approach often leads to noisy anomaly predictions, resulting in significant false positives and reduced in-distribution class segmentation performance. Mask2Anomaly overcomes this limitation by segmenting anomalies and in-distribution classes as semantically clustered masks, encouraging the objectness of the predictions. To the best of our knowledge, this is the first work to use masks both to segment anomalies and for open-set semantic and panoptic segmentation.

III. Preliminaries

Notations: Let us denote $\mathcal{X} \subset \mathbb{R}^{3\times H\times W}$ the space of RGB images, where H and W are the height and width, respectively, and with $\mathcal{Y} \subset \mathbb{N}^{Z\times H\times W}$ the space of semantic labels that associate each pixel in an image to a semantic category from a predefined set $\mathcal{Z}$, with $|\mathcal{Z}| = Z$. 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 having pixel-wise semantic class labels. Alternatively, $\mathcal{Y}$ can also be described as the semantic partition of the image into Z regions that are represented as a set of binary masks M^{gt}, where the ground-truth labels of x_i can be represented as $M^{gt} = \{m_i | m_i \in [0,1]^{H\times W}\}_{i=1}^{Z}$. From the perspective of panoptic segmentation [36], if $m_i \in things$ category, it can be further divided as $m_i = \bigcup_{j=1}^{J} m_i^j$. Where, J represents the total number of instances present in m_i.

Mask architectures: The prototypical mask architecture consists of three meta 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* $C \in \mathbb{R}^{N\times Z}$. 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 obtained from the pixel-decoder.

During training, we use the Hungarian algorithm to match ground truth masks M^{gt} with the predicted masks M. Since the Hungarian algorithm requires one-to-one correspondences and $M \geq M^{gt}$, we pad the ground truth mask M^{gt} with "no object" masks, which we indicate as ϕ. The cost function for matching M and M^{gt} is given by

$$L_{\text{masks}} = \lambda_{\text{bce}} L_{\text{bce}} + \lambda_{\text{dice}} L_{\text{dice}}, \tag{1}$$

where L_{bce} and L_{dice} are, respectively, the binary cross entropy loss and the dice loss calculated between the matched masks. The weights λ_{bce} and λ_{dice} are both set to 5.0. Additionally, we also train the model on cross-entropy loss L_{ce} to learn the semantic class of each mask that is denoted by C. The total training loss is given by

$$L = L_{\text{masks}} + \lambda_{ce} L_{ce}, \tag{2}$$

with λ_{ce} set to 2.0 for the prediction that is matched with the ground truth and 0.1 for ϕ, i.e., for no object. At inference time, the segmentation output $g(x)$ is inferred by marginalization over the softmax of C and sigmoid of M. Formally, the pixel-wise class scores $S(x) \in [0,1]^{Z\times H\times W}$ for the input image x are given by

$$S(x) = \text{softmax}(C)^T \cdot \text{sigmoid}(M). \tag{3}$$

The tensor of class scores can be represented as a concatenation of $(H \times W)$-dimensional slices associated to the Z semantic categories in $\mathcal{Z}$, i.e., $S(x) = [S_{\mathcal{Z}_1}, \ldots, S_{\mathcal{Z}_Z}]$. Therefore, the segmentation output at a pixel (h, w) is given by

$$g(x)|_{h,w} = \underset{z\in\mathcal{Z}}{\text{argmax}}\ S_z(x)|_{h,w}\,, \tag{4}$$

where the operator $\cdot|_{h,w}$ indicates that the function is taken at the spatial coordinates (h, w). Hereinafter, we will combine (3)–(4) with the following shorthand notation

$$g(x) = \underset{\mathcal{Z}}{\text{argmax}} \left(\text{softmax}(C)^T \cdot \text{sigmoid}(M)\right). \tag{5}$$

For more details about the mask architecture, please refer to the Mask2Former paper [14]. In the subsequent sections, we will address the tasks of anomaly segmentation (Section IV), open-set semantic segmentation (Section V), and open-set panoptic segmentation (Section VI) using our proposed Mask2Anomaly architecture and delve into its novel elements.

IV. Anomaly Segmentation

A. Problem Setting

Anomaly segmentation can be achieved in per-pixel semantic segmentation architectures [13] by applying the *Maximum Softmax Probability* (MSP) [32] on top of the per-pixel classifier. Formally, given the pixel-wise class scores $S(x) \in [0,1]^{Z\times H\times W}$ obtained by segmenting the image x with a per-pixel architecture, we can compute the anomaly score $f(x)$ as

$$f(x) = 1 - \max_{\mathcal{Z}}(S(x)), \tag{6}$$

where we used the same shorthand notation already adopted in (5) to indicate the max operations along the first dimension of the tensor. In this paper, we propose to adapt this framework based on MSP for mask-transformer segmentation architectures. Given such a mask-transformer architecture, we calculate the anomaly scores for an input x as

$$f(x) = 1 - \max_{\mathcal{Z}} \left(\text{softmax}(C)^T \cdot \text{sigmoid}(M)\right). \tag{7}$$

Here, $f(x)$ utilizes the same marginalization strategy of class and mask pairs as [15] to get anomaly scores. Without loss of generality, we implement the anomaly scoring (7) on top of the Mask2Former [14] architecture. However, this strategy hinges on the fact that the masks predicted by the segmentation architecture can capture anomalies well. We found that simply applying the MSP on top of Mask2Former as in (7) does not yield good results (see Fig. 1 and the results in Section VII-E). 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. 3. Now, we will discuss the proposed novel components in Mask2Anomaly.

B. Global Masked Attention

One of the key ingredients to Mask2Former [14] 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 lth masked-attention layer can

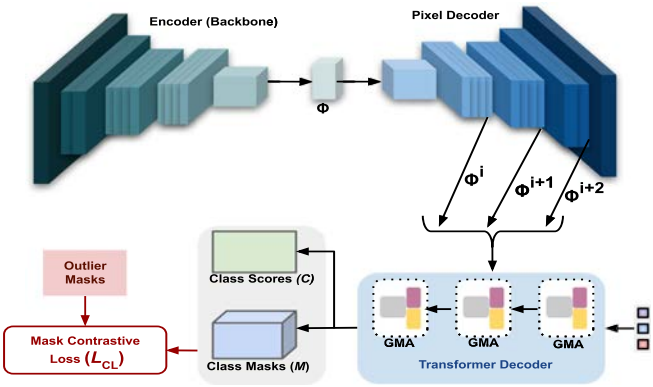

Fig. 3. *Mask2Anomaly Overview.* Mask2Anomaly meta-architecture consists of an encoder, a pixel decoder, and a transformer decoder. We propose GMA: Global Mask Attention that is discussed in Section IV-B and Fig. 4. ϕ is image features. $\phi^i, \phi^{i+1}, \phi^{i+2}$ are upsampled image features at multiple scales. Mask contrastive Loss L_{CL} (Section IV-C) utilizes outlier masks to maximize the separation between anomalies and known classes. During anomaly inference, we utilize refinement mask R_M (Section IV-D) to minimize false positives.

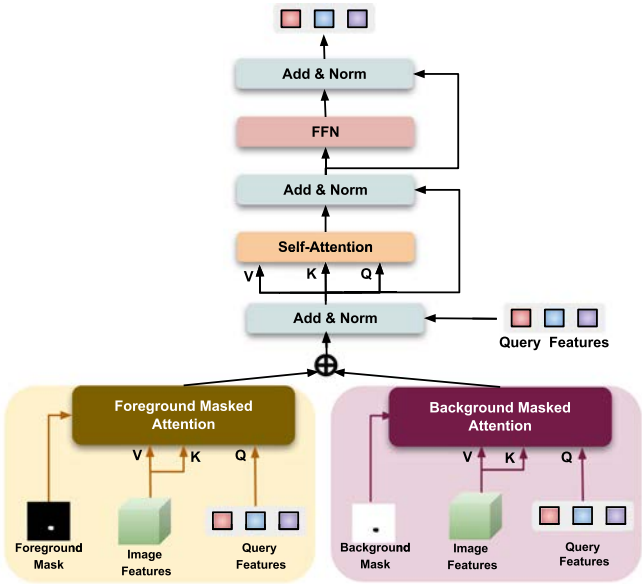

Fig. 4. *Global Mask Attention:* independently distributes the attention between foreground and background. V, K, and Q are Value, Key, and Query.

Fig. 5. *Limitation of Mask-Attention:* Masked-attention [14] selectively attends to foreground regions resulting in low attention scores (dark regions) for anomalies. Anomalies are in red. Best viewed with Zoom.

be formulated as

$$\text{softmax}(\mathcal{M}_l^F + QK^T)V + X_{in}, \tag{8}$$

where $X_{in} \in \mathbb{R}^{N \times C}$ are the N C-dimensional query features from the previous decoder layer. The 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 $f_k(\cdot)$ and $f_v(\cdot)$. Finally, $\mathcal{M}_l^F$ is the predicted foreground attention mask that at each pixel location (i, j) is defined as

$$\mathcal{M}_l^F(i,j) = \begin{cases} 0 & \text{if } M_{l-1}(i,j) \geq 0.5 \\ -\infty & \text{otherwise,} \end{cases} \tag{9}$$

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. 5. 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. 4). We call this a *global masked-attention* (GMA) formally expressed as

$$\begin{aligned} X_{out} = \text{softmax}(\mathcal{M}_l^F + QK^T)V \\ + \text{softmax}(\mathcal{M}_l^B + QK^T)V + X_{in}, \end{aligned} \tag{10}$$

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

$$\mathcal{M}_l^B(i,j) = \begin{cases} 0 & \text{if } M_{l-1}(i,j) < 0.5 \\ -\infty & \text{otherwise.} \end{cases} \tag{11}$$

The global masked-attention in (10) 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.

C. 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 [6], [26], [27].

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 them 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

Fig. 6. *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.

the negative likelihood of in-distribution classes using the class scores C and class masks M as

$$l_N = -\max_{\mathcal{Z}} \left(\text{softmax}(C)^T \cdot \text{sigmoid}(M)\right). \tag{12}$$

Ideally, for pixels corresponding to in-distribution classes l_N should be -1 since the value of $\text{softmax}(C)^T$ and $\text{sigmoid}(M)$ would be close to 1. On the other hand, for the anomalous pixels, l_N should be 0 as the likelihood of these pixels belonging to any in-distribution classes is 0 resulting $\text{softmax}(C)^T$ to be 0. Using l_N, we define our contrastive loss as

$$L_{CL} = \frac{1}{2}(l_{CL}^2),$$
$$l_{CL} = \begin{cases} l_N & \text{if } M_{\text{OOD}} = 0 \\ \max(0, m - l_N) & \text{otherwise,} \end{cases} \tag{13}$$

where the margin m is a hyperparameter that decides the minimum distance between the out-of-distribution and in-distribution classes. During mask contrastive training, we also preserve the in-distribution accuracy by training on L_{masks} and L_{ce} which formulates our total training loss as

$$L_{ood} = L_{CL} + L_{\text{masks}} + \lambda_{ce} L_{ce}. \tag{14}$$

D. Refinement Mask

False positives are one of the main problems in anomaly segmentation, particularly around object boundaries. Handcrafted methods such as iterative boundary suppression [35] 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 [14] to perform all segmentation tasks. Our method stems from the panoptic perspective [36] 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. 6). Formally, for an image x the refined anomaly scores f^r is computed as

$$f^r(x) = R_M \odot f(x), \tag{15}$$

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 $\bar{M} = \text{sigmoid}(M) > 0.5$ and the class filter $\bar{C}$ is equal to 1 only where the highest class score of $\text{softmax}(C)$ belongs to "things" or "road" classes and is greater than 0.95.

Inference: During inference, we pass the input image through Mask2Anomaly to get anomaly scores (7). Then, we refine the anomaly scores via refinement mask (15).

V. Open-Set Semantic Segmentation

A. Problem Setting

Anomaly segmentation methods solely focus on segmenting road scene anomalies. However, a strong performance for in-distribution classes is equally important. For instance, an anomaly segmentation model deployed in an autonomous vehicle that fails to identify a person crossing the road can result in a fatal accident. Hence, it is crucial that while recognizing anomalies, the performance of the model on in-distribution classes remains preserved. Open-set semantic segmentation addresses this problem by jointly assessing the model's performance on in-distribution and out-of-distribution classes. We utilize the mask properties of Mask2Anomaly to perform open-set semantic segmentation by only modifying its inference process with respect to the Anomaly Segmentation task.

Inference: Our open-set semantic segmentation network has identical mask architecture as anomaly segmentation that contains global mask attention. During the inference, we first threshold the anomaly scores obtained from (7) at a true positive rate of 95%, similar to [27]. We denote the thresholded anomaly scores by $\hat{f}(x)$. Next, we calculate the in-distribution class performance $g(x)$ by (5). Finally, we formulate the open-set semantic segmentation f_{oss} prediction of an image x as

$$f_{oss}(x) = \arg\max(\text{concat}(g(x), \hat{f}(x))). \tag{16}$$

VI. Open-Set Panoptic Segmentation

A. Problem Overview

Panoptic segmentation [36] jointly addresses the dense prediction task of semantic segmentation and instance segmentation. In this task, we divide an image into two broad categories: i) *stuff*, i.e., amorphous areas of an image that have homogeneous texture, such as grass and sky, and ii) *things*, i.e., countable objects such as pedestrians. Every pixel belonging to a *things* category is assigned a semantic label and a unique instance id, whereas, for *stuff* regions, only semantic labels are given, and the instance id is ignored. However, constructing and annotating large-scale panoptic segmentation datasets is expensive and requires significant human effort. Hwang et al. [34] address this problem by formulating it as an open-set panoptic segmentation

(OPS) problem where a model can perform panoptic segmentation on a pre-defined set of classes and identify unknown objects. This ability of the OPS model could accelerate the process of constructing large-scale panoptic segmentation datasets from existing ones.

B. Problem Setting

The key difference between panoptic and open-set panoptic segmentation is the presence of unknown objects while testing. However, handling the classification of unknown object in OPS is quite challenging. First, in comparison to open-set image classification, OPS requires the classification of unknown objects at the pixel level. Second, the absence of semantic information about unknown objects means that they are generically labeled as background during training. In order to make the problem tractable, we follow Hwang et al. [34] and make three assumptions:

1) we categorize all the unknowns into things categories (i.e., the unknowns are countable objects);
2) elements of known categories cannot be classified as unknown classes;
3) the unknown objects are always found in the background/void regions. This avoids confusion between known and unknown class regions.

We address open-set panoptic segmentation by utilizing the mask properties of Mask2Anomaly and leveraging its global mask attention. We first mask out the known *stuff* and *things* regions of an image, and then within the remaining background area, we mine the instances of the unknown objects. We will now formally discuss the method in more detail. For an input image x, Mask2Anomaly outputs a set of masks M and its corresponding class scores C. Among these, we denote the joint set of known *stuff* and *things* class masks as $M_k \in [0,1]^{N_k \times H \times W}$ and its corresponding class scores as $C_k \in \mathbb{R}^{N_k \times Z}$. Finally, we denote the number of known class masks as N_k. We obtain the background region $\mathcal{B}$ of x by using the weighted combination of M_k and C_k given by

$$\mathcal{B} = 1 - \max_{N_k}(\max_{Z}(\text{softmax}(C_k)) \cdot \text{sigmoid}(M_k)). \quad (17)$$

In light of our assumptions, $\mathcal{B}$ consists of background *stuff* classes and unknown *things* classes.

C. Mining Unknown Instances

Generally in panoptic segmentation datasets such as MS-COCO [41] the background class consists of only background *stuff* classes. However, in open-set panoptic segmentation, the background class consists of background *stuff* classes and unknown *things* classes. So, we mine the unknown instances from background $\mathcal{B}$ obtained from (17) using the following steps:

1) In the first step, we employ the connected component algorithm [5] to cluster and identify unique segments in $\mathcal{B}$.
2) Next, we calculate each connected component's overlap with the individual masks of M. Intersection over union is used for calculating the overlap.
3) If there is a significant overlap between a connected component and a mask $M^i \in M$, we calculate the average stuff class entropy $\mathcal{E}_S$ and average things class entropy $\mathcal{E}_T$ using the corresponding class scores $C^i \in C$.
4) Finally, if $\mathcal{E}_S > \mathcal{E}_T$ we can conclude that the connected component is more likely to belong to the *things* class. Hence, we classify the connected component to be an unknown instance.

Inference: During the inference, we first calculate $\mathcal{B}$ from (17). Then, we identify the unknown instances in $\mathcal{B}$ by following the above described steps of mining unknown instances.

VII. Experimentation

A. Datasets

Anomaly Segmentation: We train Mask2Anomaly on the Cityscapes [16] dataset, which consists of 2,975 training and 500 validation images. To evaluate anomaly segmentation, we use Road Anomaly [42], Lost & Found [51], Fishyscapes [6], and Segment Me If You Can (SMIYC) benchmarks [11].

Road Anomaly: is a collection of 60 web images with anomalous objects on or near the road.

Lost & Found: has 1068 test images with small obstacles for road scenes.

Fishyscapes (FS): consists of two datasets, Fishyscape static (FS static) and Fishyscapes lost & found (FS lost & found). Fishyscapes static is built by blending Pascal VOC [22] 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 [51], 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 to segment road anomalies and has 30 validation and 327 test images.

Open-set panoptic segmentation: We perform all the open-set panoptic segmentation experiments on the panoptic segmentation dataset of MS-COCO [41]. The dataset consists of 118 thousand training images and 5 thousand validation images having 80 *thing* classes and 53 *stuff* classes. We construct open-set panoptic segmentation dataset by removing the labels of a small set of known *things* classes from the train set of panoptic segmentation dataset. The removed set of *things* classes are treated as unknown classes. We construct three different training dataset split with increasing order of difficulty with (5%, 10%, 20%) of unknown classes. The removed classes in each split that are removed cumulatively is given as: 5%: {car, cow, pizza, toilet}, 10%: {boat, tie, zebra, stop sign }, 20%: {dining table, banana, bicycle, cake, sink, cat, keyboard, bear}.

Open-set semantic segmentation: We use StreetHazards [31], a synthetic dataset for open-set semantic segmentation. StreetHazards is created using the CARLA simulator [19] and leveraging the Unreal Engine to render realistic road scene images in which diverse anomalous objects are inserted. The dataset consists of 5,125 training images and 1,031 validation images

having 12 classes. The test set has 1,500 images along with an additional anomaly class.

B. Evaluation Metrics

Anomaly Segmentation: We evaluate all the anomaly segmentation methods at pixel and component levels that are described next.

Pixel-Level: For pixel-wise evaluation, $Y \in \{Y_a, Y_{na}\}$ is the pixel level annotated ground truth labels for an image x containing anomalies. Y_a and Y_{na} represents the anomalous and non-anomalous labels in the ground-truth, respectively. Assume that $\hat{Y}(\gamma)$ is the model prediction obtained by thresholding at γ. Then, we can write the precision and recall equations as

$$\text{precision}(\gamma) = \frac{|Y_a \cap \hat{Y}_a(\gamma)|}{|\hat{Y}_a(\gamma)|} \tag{18}$$

$$\text{recall}(\gamma) = \frac{|Y_a \cap \hat{Y}_a(\gamma)|}{|Y_a|}, \tag{19}$$

and the AuPRC can be approximated as

$$\text{AuPRC} = \int_{\gamma} \text{precision}(\gamma)\text{recall}(\gamma). \tag{20}$$

The AuPRC works well for unbalanced datasets making it particularly suitable for anomaly segmentation since all the datasets are significantly skewed. Next, we consider the False Positive Rate at a true positive rate of 95% (FPR_{95}), an important criterion for safety-critical applications that is calculated as

$$\text{FPR}_{95} = \frac{|\hat{Y}_a(\gamma^*) \cap Y_{na}|}{|Y_{na}|}, \tag{21}$$

where γ^* is a threshold when the true positive rate is 95%.

Component-Level: SMIYC [11] introduced component-level evaluation metrics that solely focus on detecting anomalous objects regardless of their size. These metrics are important to be considered because pixel-level metrics may not penalize a model for missing a small anomaly, even though such a small anomaly may be important to be detected. In order to have a component-level assessment of the detected anomalies, the quantities to be considered are the component-wise true-positives (TP), false-negatives (FN), and false-positives (FP). These component-wise quantities can be measured by considering the anomalies as the positive class. From these quantities, we can use three metrics to evaluate the component-wise segmentation of anomalies: sIoU, PPV, and F1^*. Here we provide the details of how these metrics are computed, using the notation $\mathcal{K}$ to denote the set of ground truth components, and $\hat{\mathcal{K}}$ to denote the set of predicted components.

The *sIoU* metric used in SMIYC [11] is a modified version of the component-wise intersection over union proposed in [53], which considers the ground-truth components in the computation of the TP and FN. Namely, it is computed as

$$\text{sIoU}(k) = \frac{|k \cap \hat{K}(k)|}{|k \cap \hat{K}(k) \backslash \mathcal{A}(k)|}, \qquad \hat{K}(k) = \bigcup_{\hat{k} \in \hat{\mathcal{K}},\, \hat{k} \cap k \neq \emptyset} \hat{k}, \tag{22}$$

where $\mathcal{A}(k)$ is an adjustment term that excludes from the union those pixels that correctly intersect with another ground-truth component different from k. Given a threshold $\tau \in [0, 1]$, a target $k \in \mathcal{K}$ is considered a TP if $sIoU(k) > \tau$, and a FN otherwise.

The positive predictive value (*PPV*) is a metric that measures the FP for a predicted component $\hat{k} \in \hat{\mathcal{K}}$, and it is computed as

$$\text{PPV}(\hat{k}) = \frac{|\hat{k} \cap \hat{K}(k)|}{|\hat{k}|}. \tag{23}$$

A predicted component $\hat{k} \in \hat{\mathcal{K}}$ is considered a FP if $PPV(\hat{k}) \leq \tau$. Finally, the $F1^*$ summarizes all the component-wise TP, FN, and FP quantities by the following formula:

$$\text{F1}^*(\tau) = \frac{2\text{TP}(\tau)}{2\text{TP}(\tau) + \text{FN}(\tau) + \text{FP}(\tau)}. \tag{24}$$

Open-set semantic segmentation: We use open-IoU [27] to evaluate open-set semantic segmentation. Unlike, IoU, open-IoU takes into account the false positives (FP^{OOD}) and false negatives (FN^{OOD}) of an anomaly segmentation model. To measure open-IoU, we first threshold the output of the anomaly segmentation model at a true positive rate of 95% and then re-calculate the classification scores of in-distribution classes according to the anomaly threshold. Now, FP^{OOD} and FN^{OOD} for a class α can be calculated as

$$\text{FP}_{\alpha}^{\text{OOD}} = \sum_{i=1, i\neq\alpha}^{Z+1} \text{FP}_{\alpha}^{i}, \; \text{FN}_{\alpha}^{\text{OOD}} = \sum_{i=1, i\neq\alpha}^{Z+1} \text{FN}_{\alpha}^{i}. \tag{25}$$

Using $\text{FP}_{\alpha}^{\text{OOD}}$ and $\text{FN}_{\alpha}^{\text{OOD}}$, we can calculate the open-IoU for class α as

$$\text{open-IoU}_{\alpha} = \frac{\text{TP}_{\alpha}}{\text{TP}_{\alpha} + \text{FP}_{\alpha}^{\text{OOD}} + \text{FN}_{\alpha}^{\text{OOD}}}. \tag{26}$$

TP_{α} denotes the true-positive of class α. An ideal open-set model will have open-IoU to be equal to IoU.

Open-set panoptic segmentation: We measure the panoptic segmentation quality of known and unknown classes by using the panoptic quality (PQ) metric [36]. For each class, PQ is calculated individually and averaged over all the classes making PQ independent of class imbalance. Every class has predicted segments p and its corresponding ground truths g that is divided into three parts: true positives (TP): matched pair of segments, false positives (FP): unmatched predicted segments, and false negatives (FN): unmatched ground truth segments. Given the three sets, PQ can be formulated as

$$\text{PQ} = \underbrace{\frac{\sum_{(p,g)\in \text{TP}} \text{IoU}(p, g)}{|\text{TP}|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|\text{TP}|}{|\text{TP}| + \frac{1}{2}|\text{FP}| + \frac{1}{2}|\text{FN}|}}_{\text{recognition quality (RQ)}}. \tag{27}$$

From the above equation, we can see PQ as the product of a segmentation quality (SQ) and a recognition quality (RQ). RQ can be inferred as an F1 score that gives the estimation of segmentation quality. SQ is the average IoU of matched segments.

C. Implementation Details

Anomaly Segmentation: Our implementation is derived from [14], [15]. We use a ResNet-50 [30] encoder, and its weights are initialized from a model that is pre-trained with barlow-twins [62] self-supervision on ImageNet [17]. We freeze the encoder weights during training, saving memory and training time. We use a multi-scale deformable attention Transformer (MSDeformAttn) [68] 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 Cheng et al. [14] and consists of 9 layers with 100 queries. We train Mask2Anomaly using a combination of binary cross-entropy loss and the dice loss [47] 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 [45] with a weight decay of 0.05. We use an image crop of 380×760 with large-scale jittering [21] 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 [57] where we cut an object from MS-COCO [41] 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 4,000 iterations with m as 0.75, a learning rate of 1e-5, and batch size 8, keeping all the other hyper-parameters the same as above. The probability of choosing an outlier in a training batch is kept at 0.2. The existing mask-transformers methods – EAM [28], AEM [28], Maskomaly [1], and RbA [49] – report in their papers results obtained using different training data and backbones. So, for a fair comparison, we train all the methods on the same data and using the same ResNet-50 backbone.

Open-set semantic segmentation: We the use Streethazard [31] dataset to train Mask2Anomaly with a Swin-Base backbone. The model was trained for 50 thousand iterations keeping all the parameters the same as anomaly segmentation. Next, we train Mask2Anomaly on outlier images in a contrastive settings. The outlier image was created by AnomalyMix [57] using MS-COCO [41] and Streethazard image. We train the network for 5000 iterations keeping the Swin-Base backbone frozen. The image crop size was kept at 380×760, the rest all the other hyper-parameters were the same as for anomaly segmentation. We adapt EAM [28], AEM [28], Maskomaly [1], and RbA [49] for open-set semantic segmentation task by keeping the same set of hyper-parameters and architectural details as ours.

Open-set panoptic segmentation: We train Mask2Anomaly having ResNet-50 backbone for 370 thousand iterations. Our training approach employs a batch size of 8, incorporating cropped input images sized at 640×640. We keep the remaining hyperparameters the same as specified in the anomaly segmentation. Across all the three training datasets, which contain 5%, 10%, and 15% of unknown classes, the number of connected components were 2, 2, and 3 respectively. The number of iteration for connected component algorithm was kept at 500 for each training dataset.

D. Main Results

Anomaly Segmentation: Table I shows the pixel-level anomaly segmentation results achieved by Mask2Anomaly and recent SOTA methods on Fishyscapes, SMIYC, and Road Anomaly datasets. We observed that Mask2Anomaly significantly improves average AuPRC and FPR_{95} compared with per-pixel and mask-transformer methods. Another observation is 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 yields excellent results on all the datasets. Similarly, such generalization issues can also be found in mask-transformers based methods, such as Maskomaly [1]. Next, Table II demonstrates that Mask2Anomaly outperforms all the baselined methods on component-level evaluation metrics. Interestingly, RbA shows a sharp decline in the component-level anomaly segmentation for Lost & Found dataset. We attribute this decrease to the behavior of the tanh function present in the inference function of RbA. We test our assumption on the Lost & Found dataset by replacing the RbA inference function with ours. We found that $sIoU$, PPV, and $F1^*$ significantly improved to 47.85, 38.26, and 40.67, respectively. To conclude, Mask2Anomaly yields state-of-the-art anomaly segmentation performance both in pixel and component metrics. To get a better understanding of the visual results, in Fig. 8, we qualitatively compare the anomaly scores predicted by Mask2Anomaly and its closest competitors: Dense Hybrid [27], Maximized Entropy [12], and RbA [49]. The results from RbA, 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 the precise segmentation of anomalies while at the same time having minimal false positives.

Another critical characteristic of any anomaly segmentation method is that it should not disturb the in-distribution classification performance, or else it would make the semantic segmentation model unusable. We show in Table V(c) that Mask2Anomaly using only GMA achieves a mIoU of 80.45. However, when adding also the mask contrastive training, Mask2Anomaly in-distribution accuracy on the Cityscapes validation dataset drops slightly to 78.88 mIoU, but is still 1.46 points higher than the vanilla Mask2Former. Moreover, it is important to note that both Mask2Anomaly and Mask2Former are trained for 90 k iterations, indicating that, although Mask2Anomaly additionally attends to the background mask region, it shows convergence similar to Mask2Former. Fig. 9 qualitatively shows that Mask2Anomaly's semantic segmentation results are almost identical to Mask2Former.

Open-set semantic segmentation: Table III illustrates the open-set semantic segmentation performance of Mask2Anomaly on the StreetHazards test set. In terms of anomaly segmentation performance, we observe that Mask2Anomaly gives a significant gain of 90% compared to DenseHybrid in AuPRC with minimal increase in false positives. Notably, Mask2Anomaly also gives the second best closed set performance, indicating its ability to improve in-distribution

TABLE I

PIXEL-LEVEL EVALUATION: ON AVERAGE, MASK2ANOMALY SHOWS SIGNIFICANT IMPROVEMENT OVER BASELINED PER-PIXEL AND MASK-TRANSFORMER BASED METHODS

Methods	SMIYC RA-21		SMIYC RO-21		FS L&F		FS Static		Road Anomaly		Lost & Found		Average	
	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95}↓	AuPRC ↑	FPR_{95}↓
Max Softmax [32](ICLR'17)	27.97	72.02	15.72	16.6	1.77	44.85	12.88	39.83	15.72	71.38	30.14	33.20	17.36	46.31
Entropy [32](ICLR'17)	-	-	-	-	2.93	44.83	15.4	39.75	16.97	71.1	-	-	11.76	51.89
Mahalanobis [38](NeurIPS'18)	20.04	86.99	20.9	13.08	-	-	-	-	14.37	81.09	54.97	12.89	27.57	48.51
Image Resynthesis [42](ICCV'19)	52.28	25.93	37.71	4.7	5.7	48.05	29.6	27.13	-	-	57.08	8.82	36.47	22.92
Learning Embedding [6](IJCV'21)	37.52	70.76	0.82	46.38	4.65	24.36	57.16	13.39	-	-	61.70	10.36	32.37	33.05
Void Classifier [6](IJCV'21)	36.61	63.49	10.44	41.54	10.29	22.11	4.5	19.4	-	-	4.81	47.02	13.33	38.71
JSRNet [59](ICCV'21)	33.64	43.85	28.09	28.86	-	-	-	-	**94.4**	**9.2**	74.17	6.59	57.57	22.12
SML [35](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 [18](CVPR'21)	56.44	61.86	71.34	3.15	43.22	15.79	72.59	18.75	38.21	64.75	81.71	4.64	60.58	28.15
Maximized Entropy [12](ICCV'21)	85.47	15.00	85.07	0.75	29.96	35.14	86.55	8.55	48.85	31.77	77.90	9.70	68.96	16.81
Dense Hybrid [27](ECCV'22)	77.96	**9.81**	87.08	0.24	**47.06**	**3.97**	80.23	5.95	31.39	63.97	78.67	**2.12**	67.06	14.34
PEBEL [57](ECCV'22)	49.14	40.82	4.98	12.68	44.17	7.58	92.38	1.73	45.10	44.58	57.89	4.73	48.94	18.68
Mask2Former [14]	10.72	89.82	53.57	82.07	43.50	89.01	16.20	40.10	30.99	75.27	22.19	67.09	29.52	73.89
EAM [28](CVPRw'23)	31.30	48.20	0.70	45.20	41.93	80.08	59.42	22.87	29.89	54.98	27.82	76.01	31.84	54.55
AEM [28](CVPRw'23)	31.00	47.90	0.60	45.60	42.16	81.20	57.68	33.57	27.87	87.72	26.80	79.89	31.01	62.64
Maskomaly [1](BMVC'23)	35.00	81.30	0.60	43.50	1.77†	44.85†	12.88†	39.83†	16.33	73.17	1.79	49.53	11.39	55.38
RbA [49](ICCV'23)	70.70	20.60	46.00	95.20	21.66	39.60	83.26	4.22	52.97	32.91	53.49	39.20	54.68	38.62
Mask2Anomaly (Ours)	**88.70**	14.60	**93.30**	**0.20**	46.04	4.36	**95.20**	**0.82**	79.70	13.45	**86.59**	5.75	**81.59**	**6.53**

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, and the line in-between the table divides the per-pixel architecture and mask-transformer based methods.† are the results on validation dataset.

TABLE II

ANOMALY SEGMENTATION COMPONENT-LEVEL EVALUATION: MASK2ANOMALY ACHIEVES LARGE IMPROVEMENT ON COMPONENT LEVEL EVALUATION METRICS AMONG THE BASELINED METHODS

Methods	SMIYC RA-21			SMIYC RO-21			Lost & Found			Average		
	sIoU ↑	PPV ↑	$F1^*$↑	sIoU ↑	PPV ↑	$F1^*$↑	sIoU ↑	PPV ↑	$F1^*$↑	sIoU ↑	PPV ↑	$F1^*$↑
Max Softmax [32](ICLR'17)	15.48	15.29	5.37	19.72	15.93	6.25	14.20	62.23	10.32	16.47	31.15	7.31
Ensemble [37](NurIPS'17)	16.44	20.77	3.39	8.63	4.71	1.28	6.66	7.64	2.68	10.58	11.04	2.45
Mahalanobis [38](NeurIPS'18)	14.82	10.22	2.68	13.52	21.79	4.70	33.83	31.71	22.09	20.72	21.24	9.82
Image Resynthesis [42](ICCV'19)	39.68	10.95	12.51	16.61	20.48	8.38	27.16	30.69	19.17	27.82	20.71	13.35
MC Dropout [48](CVPR'20)	20.49	17.26	4.26	5.49	5.77	1.05	17.35	34.71	12.99	14.44	19.25	6.10
Learning Embedding [6](IJCV'21)	33.86	20.54	7.90	35.64	2.87	2.31	27.16	30.69	19.17	32.22	18.03	9.79
SML [35](ICCV'21)	26.00	24.70	12.20	5.10	13.30	3.00	32.14	27.57	26.93	21.08	21.86	14.04
SynBoost [18](CVPR'21)	34.68	17.81	9.99	44.28	41.75	37.57	36.83	**72.32**	48.72	38.60	43.96	32.09
Maximized Entropy [12](ICCV'21)	49.21	39.51	28.72	47.87	62.64	48.51	45.90	63.06	49.92	47.66	55.07	42.38
JSRNet [59](ICCV'21)	20.20	29.27	13.66	18.55	24.46	11.02	34.28	45.89	35.97	24.34	33.21	20.22
Void Classifier [6](IJCV'21)	21.14	22.13	6.49	6.34	20.27	5.41	1.76	35.08	1.87	9.75	25.83	4.59
Dense Hybrid [27](ECCV'22)	54.17	24.13	31.08	45.74	50.10	50.72	46.90	52.14	52.33	48.94	42.12	44.71
PEBEL [57](ECCV'22)	38.88	27.20	14.48	29.91	7.55	5.54	33.47	35.92	27.11	34.09	23.56	15.71
Mask2Former [14]	25.20	18.20	15.30	5.00	21.90	4.80	17.88	18.09	9.77	16.03	19.40	9.96
EAM [28](CVPRw'23)	25.90	12.70	12.80	29.60	3.00	3.60	36.21	40.39	35.77	30.57	18.69	17.39
AEM [28](CVPRw'23)	25.20	12.40	12.10	29.50	2.80	3.40	36.64	41.07	36.21	30.44	18.75	17.23
Maskomaly [32](BMVC'23)	25.50	23.50	12.00	28.20	9.80	12.00	12.31	26.13	17.45	22.00	19.81	13.81
RbA [32](ICCV'23)	44.10	22.00	18.70	26.00	24.70	15.40	1.44	7.30	0.43	23.84	18.00	11.51
Mask2Anomaly (Ours)	**60.40**	**45.70**	**48.60**	**61.40**	**70.30**	**69.80**	**56.07**	63.41	**62.78**	**59.29**	**59.80**	**60.39**

Higher values of sIoU, PPV, and F1* are better. The line in-between the table divides the per-pixel architecture and mask-transformer based methods. The best and second best results are bold and underlined, respectively.

while giving state-of-the-art anomaly segmentation results. Furthermore, we measure open-set semantic segmentation using Open-IoU metrics, which allows us to measure anomalous and in-distribution class performance jointly.

The StreetHazard test dataset consists of two sets: t5 and t6. So, to calculate Open-IoU on t5: Open-IoU^{t5}, we select the anomaly threshold from t6 at a true positive rate of 95% and then re-calculate the classification scores of in-distribution classes t5. We repeat the same steps to get Open-IoU^{t6}. To get the overall Open-IoU on the StreetHazard test, we calculate the weighted average of Open-IoU on t5 and t6 according to the number of images in each set. In Table III, we can observe Mask2Anomaly outperforms other baselined methods by a significant margin of 30% on Open-IoU metrics. It is also important to note that methods such as OOD-Head achieve good close-set performance but show low Open-IoU. On the other hand, Outlier Exposure has a relatively better Open-IoU but loses close set performance. Similarly, mask-transformers such as RbA and AEM show high close-set performance but fail to perform on open-set and anomaly segmentation metrics. Mask2Anomaly does not suffer from such shortcomings and yields the best open-set performance while maintaining a strong closed-set performance. From Fig. 7 we can visually infer that Mask2Anomaly is able to segment the anomalous/open-set objects more accurately than the baseline methods.

Open-set panoptic segmentation: Table IV summarises the open-set panoptic segmentation performance of per-pixel architecture based methods: Void-train and EOPSN, and

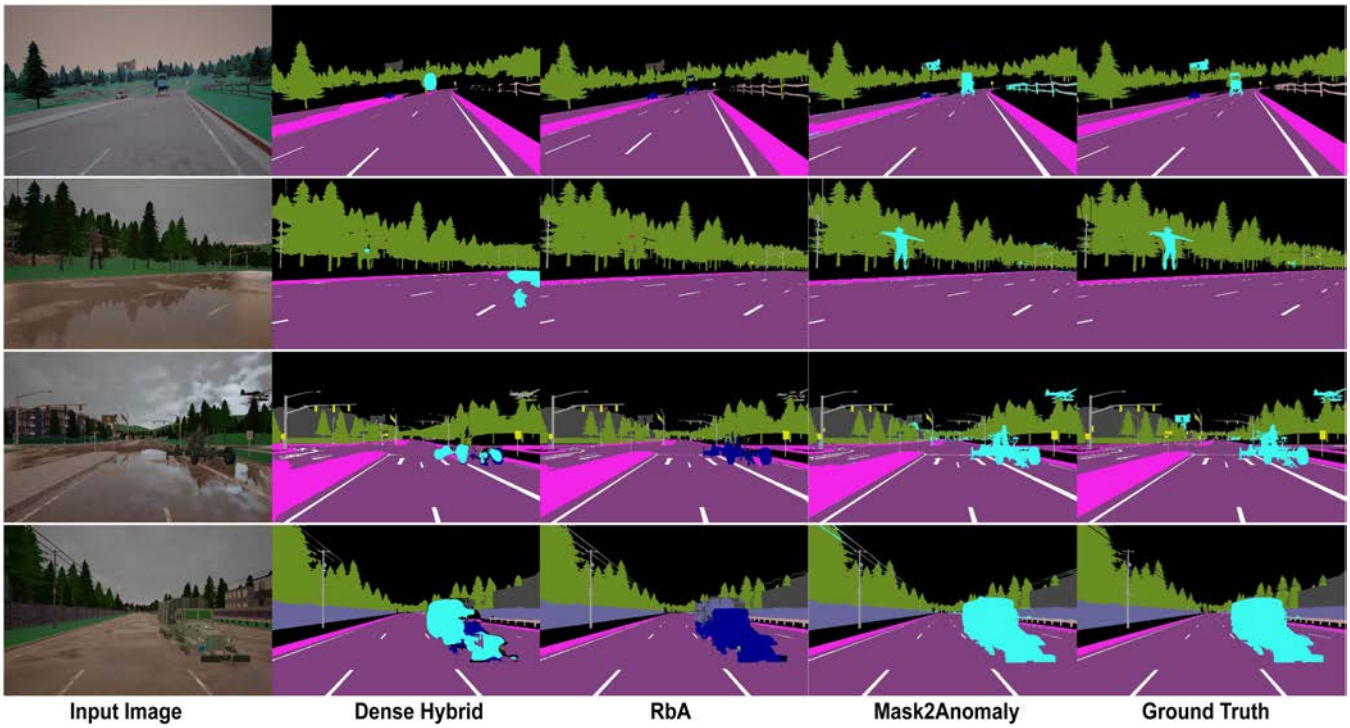

Fig. 7. *Qualitative results of open-set semantic segmentation*: We can observe that the Mask2Anomaly gives precise boundaries for open-set objects compared to best-performing per-pixel architecture and mask-transformer architectures.

TABLE III
OPEN-SET SEMANTIC SEGMENTATION QUANTITATIVE EVALUATION: WE OBSERVE THAT MASK2ANOMALY ACHIEVES THE BEST PERFORMANCE ON OPEN-SET SEGMENTATION MATRICES

	Anomaly Segmentation		Closed Set Performance	Open Set Performance		
Methods	AuPRC ↑	FPR_{95} ↓	mIoU↑	Open-IoU^{t5} ↑	Open-IoU^{t6} ↑	Open-IoU↑
MSP [32] (ICLR'17)	7.5	27.9	65.0	32.7	40.2	35.1
ODIN [39] (ICLR'18)	7.0	28.7	65.0	26.4	33.9	28.8
Outlier Exposure [33] (ICLR'19)	14.6	17.7	61.7	43.7	44.1	43.8
OOD-Head [3] (GCPR'19)	19.7	56.2	66.6	33.7	34.3	33.9
MC Dropout [48] (CVPR'20)	7.5	79.4	-	-	-	-
SynthCP [60] (ECCV'20)	9.3	28.4	-	-	-	-
TRADI [25] (ECCV'20)	7.2	25.3	-	-	-	-
OVNNI [24] (CoRR'20)	12.6	22.2	54.6	-	-	-
Energy [43] (NurIPS'20)	12.9	18.2	63.3	41.7	44.9	42.7
PAnS [23] (CVPRW'21)	8.8	23.2	-	-	-	-
SO+H [26] (VISIGRAPP'21)	12.7	22.2	59.7	-	-	-
DML [8] (ICCV'21)	14.7	17.3	-	-	-	-
ReAct [55] (NurIPS'21)	10.9	21.2	62.7	33.0	36.2	34.0
OH*MSP [4] (CoRR'21)	18.8	30.9	66.6	43.3	44.2	43.6
ML [31] (ICML'22)	11.6	22.5	65.0	39.6	44.5	41.2
DenseHybrid [27] (ECCV'22)	30.2	**13.0**	63.0	46.1	45.3	45.8
AEM [28] (CVPRw'23)	30.7	99.7	71.3	35.3	54.6	41.4
EAM [28] (CVPRw'23)	31.2	20.3	-	-	-	-
Maskomaly [1] (BMVC'23)	27.1	43.1	-	-	-	-
RbA [49] (ICCV'23)	50.1	96.9	**73.2**	10.4	12.1	10.9
Mask2Anomaly	**58.1**	14.9	72.3	**59.9**	**59.7**	**59.8**

'-' indicates the unavailability of benchmarked results and the line in-between the table divides the per-pixel architecture and mask-transformer based methods.

mask-transformers: EAM, AEM, and Mask2Anomaly. Void-train is a baseline method in which we train the void regions of an image by treating it as a new class. We can observe Mask2Anomaly shows the best open-set panoptic segmentation results among all the baselined methods on different proportions of unknown classes. Additionally, it also shows strong results on in-distribution classes that are indicated by various panoptic evaluation metrics. Fig. 10 illustrates the qualitative comparison of Mask2Anomaly with baselined methods on most challenging dataset having 20% unknown classes. In Fig. 10 (Row:

TABLE IV
OPEN-SET PANOPTIC SEGMENTATION QUANTITATIVE RESULTS: WE SHOW QUANTITATIVE RESULTS ON THE COCO VAL SET BY ALL THE METHODS ON VARYING KNOWN-UNKNOWN SPLITS

		Known Classes									Unknown Classes		
K(%)	Methods	$PQ_{Th}\uparrow$	$SQ_{Th}\uparrow$	$RQ_{Th}\uparrow$	$PQ_{St}\uparrow$	$SQ_{St}\uparrow$	$RQ_{St}\uparrow$	PQ ↑	SQ ↑	RQ ↑	$PQ_{unk}\uparrow$	$SQ_{unk}\uparrow$	$RQ_{unk}\uparrow$
5	Void-train	43.6	80.4	52.8	28.2	71.5	36.0	37.3	76.7	45.9	8.6	72.7	11.8
	EOPSN [34]	44.8	80.5	54.2	28.3	71.9	36.2	38.0	76.9	46.8	23.1	74.7	30.9
	Mask2Anomaly	**53.0**	**82.8**	**63.4**	**41.3**	**80.8**	**50.1**	**48.2**	**81.9**	**57.9**	**24.3**	**78.2**	**32.1**
10	Void-train	43.7	80.1	53.1	28.1	73.0	35.9	37.1	77.1	45.8	8.1	72.6	11.2
	EOPSN [34]	44.5	80.6	53.8	28.4	71.8	36.2	37.7	76.8	46.3	17.9	76.8	23.3
	Mask2Anomaly	**51.8**	**82.7**	**62.0**	**41.6**	**80.5**	**50.5**	**47.4**	**81.8**	**57.1**	**19.7**	**77.0**	**25.7**
20	Void-train	44.1	80.1	53.5	27.9	71.6	35.6	36.8	76.3	45.4	7.5	72.9	10.3
	EOPSN [34]	45.0	80.3	54.5	28.2	71.2	36.2	37.4	76.2	46.2	11.3	73.8	15.3
	EAM [28]	-	-	-	-	-	-	43.5	**82.0**	52.2	11.3	73.3	15.3
	AEM [28]	-	-	-	-	-	-	43.5	**82.0**	52.2	13.2	73.4	18.0
	Mask2Anomaly	**50.8**	**81.6**	**60.8**	**40.4**	**80.8**	**49.0**	**46.1**	81.2	**55.5**	**14.6**	**76.2**	**19.1**

K denoted the % of unknown classes present in the dataset. The best results are highlighted in bold.

TABLE V
MASK2ANOMALY ABLATION TABLES: **(A)** COMPONENT-WISE ABLATION OF MASK2ANOMALY

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

(a)

margin(m)	AuPRC↑	$FPR_{95}\downarrow$
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↑	$FPR_{95}\downarrow$
CA [15]	76.43	20.30	89.35
MA [14]	77.42	10.60	89.39
GMA	**80.45**	**32.35**	**25.95**

(c)

	AuPRC↑	$FPR_{95}\downarrow$
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↑	$FPR_{95}\downarrow$
0.1	63.01	14.66
0.2	**69.41**	**9.46**
0.5	69.20	11.03
1	68.77	10.53

(e)

Results in italics show Mask2Former results. GMA: global mask attention, CL: contrastive learning, and RM: refinement mask. (b) shows the behavior of LCL by choosing different margin (M) Values. We empirically find the best results when M is 0.75. (C) global masked attention (GMA) performs the best among various attention mechanisms: cross-attention (CA) and masked-attention (MA). It is also important to note that the derived results do not have any additional proposed components of Mask2Anomaly apart from GMA. (d) we show the performance gain by using a refinement mask that masks the {stuff \ 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. The best result is achieved at 0.2 probability. (All the results reported on FS lost & found validation set).

1-3), we can see Mask2Anomaly can better perform panoptic segmentation on unknown instances compared with baselined methods. Fig. 10 (Row: 4), shows the panoptic segmentation on known classes where we can observe Mask2Anomaly outputs are precise with minimal false positives.

E. Ablations

All the results reported in this section are based on the FS L&F validation dataset.

Mask2Anomaly: Table V(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 contrastive learning is very important for the FPR_{95}. The mask refinement brings further improvements to both. Fig. 11 visually demonstrates the positive effect of all the components.

Global Mask Attention: To better understand the effect of the global masked attention (GMA), in Table V(c), we compare it to the masked-attention (MA) [14] and cross-attention (CA) [58]. 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. 5. On the other hand, the GMA provides improvements across all the metrics. This is confirmed visually in Fig. 12, where we show the negative attention maps for the three methods at different resolutions. The 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 V(d) shows the performance gains due to the refinement mask. We observe that filtering out the $\{''\text{stuff}'' \setminus ''\text{road}''\}$ regions of the prediction map improves the FPR_{95} by 14.61 along with marginal improvement in AuPRC. On the other hand, removing the $\{''\text{things}'' \setminus ''\text{road}''\}$ regions degrades the results, confirming our hypothesis that anomalies are likely to belong to the "things" category. Fig. 11 qualitatively shows the improvement achieved with the refinement mask.

Mask Contrastive Learning: We tested the effect of the margin in the contrastive loss L_{CL}, and we report these results in Table V(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

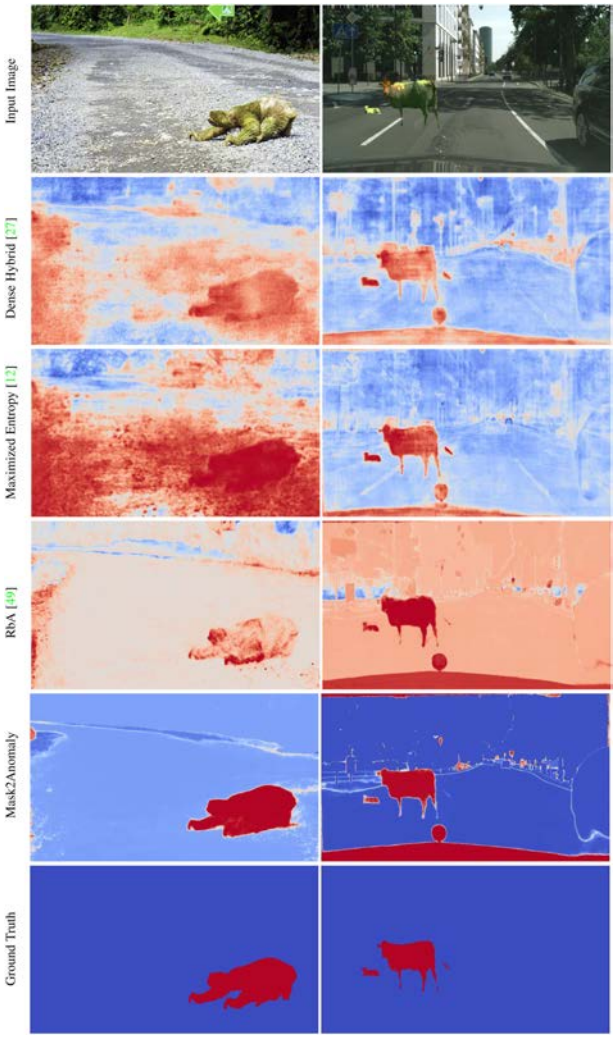

Fig. 8. *Anomaly segmentation qualitative results*: We observe that Dense Hybrid [27], RbA [49], and Maximized Entropy [12] suffer from large false positives, whereas Mask2Anomaly shows accurate pixel-wise anomaly segmentation results.

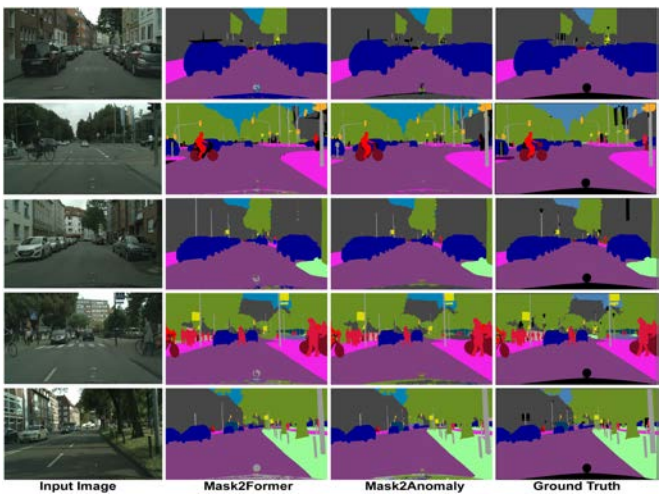

Fig. 9. *Semantic segmentation results:* We can visually infer that Mask2Anomaly shows similar segmentation results when compared with Mask2Former [14].

TABLE VI
Connected Component Training Iteration: We Show the Panoptic Segmentation Performance of Unknown Classes With the Increasing Number of Iterations

Number of Iterations	PQ ↑	SQ ↑	RQ ↑
100	11.4	77.2	14.8
200	12.5	**77.8**	16.0
500	**14.6**	76.2	**19.1**
1000	10.9	76.1	14.9

We find the best performance at 500 iterations. Best results are shown in bold.

TABLE VII
Number of Connected Components: Shows the Panoptic Segmentation Performance of Unknown Classes With the Increasing Number of Connected Components

Number of Connected Components	PQ ↑	SQ ↑	RQ ↑
1	10.9	76.1	14.9
2	12.4	**76.5**	16.3
3	**14.6**	76.2	**19.1**
5	14.0	78.2	17.9

We find the best performance at 3. Best results are shown in bold.

probability, which is the likelihood of selecting an outlier image in a batch. The results shown in Table V(e) indicate that the best performance is achieved at 0.2, but the results remain stable for higher values of the batch outlier probability.

Mining Unknowns Instances: We quantitatively summarise the impact of mining unknown instances in panoptic segmentation of unknown instances shown in Table VIII. We can clearly observe that removing the mining of unknown instances from Mask2Anomaly drastically reduces the performance across all the metrics. Also, the absence of global mask attention further degrades performance.

Connected Components: Tables VI and VII shows the impact of connected components hyperparameters on open-set panoptic segmentation of unknown classes. In both tables, we train the model on dataset split having 20 % of unknown classes. In Table VI, we can observe that Mask2Anomaly shows the best performance at 500 iterations, whereas in Table VII we achieve the best performance when the number of connected components is set to 3.

Architectural Efficacy of Mask2Anomaly: We demonstrate the efficacy of Mask2Anomaly by comparing it to the vanilla Mask2Former but using larger backbones. The results in Table X show that despite the disadvantage, Mask2Anomaly with a ResNet-50 still performs better than Mask2Former using large transformer-based backbones like Swin-S. It is also important to note that the number of training parameters for Mask2Anomaly can be reduced to 23 M as we use a frozen self-supervised pre-trained encoder during the entire training, which is significantly less than all the Mask2Former variations.

VIII. Discussion

Performance stability: Employing an outlier set to train an anomaly segmentation model presents a challenge because the model's performance can vary significantly across different sets of outliers. Here, we show that Mask2Anomaly performs similarly when trained on different outlier sets. We randomly

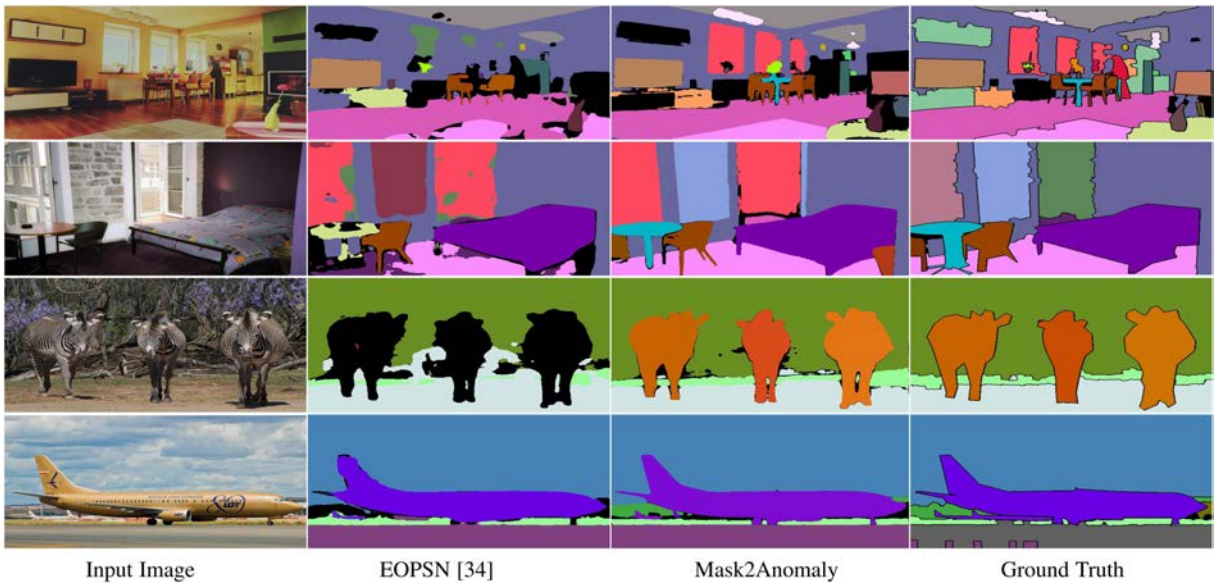

Fig. 10. *Open-set panoptic segmentation qualitative results*: Row 1-3: We can observe that Mask2Anomaly is better able to segment the different instances of unknown objects compared with the baselined method. Row 4: Shows that Mask2Anomaly gives better panoptic segmentation with precise boundaries on known classes.

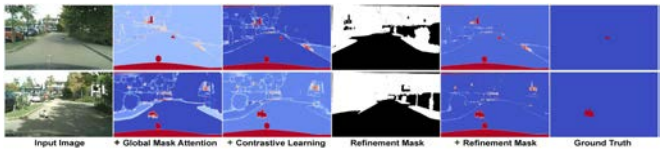

Fig. 11. *Mask2Anomaly Qualitative Ablation*: demonstrates the performance gain by progressively adding (left to right) proposed components. Masked-out regions by refinement mask are shown in white. Anomalies are represented in red.

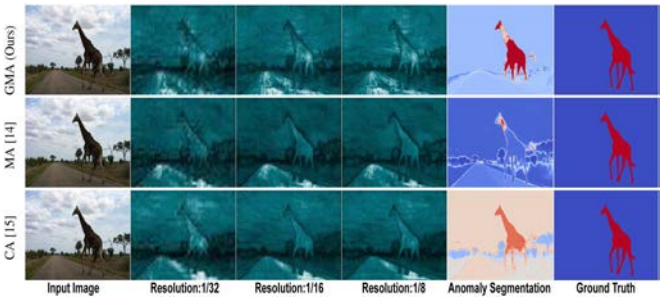

Fig. 12. *Visualization of negative attention maps and results:* Global mask attention gives high attention scores to anomalous regions across all resolutions, showing the best anomaly segmentation results among the compared attention mechanisms. Cross-attention performs better than mask-attention but 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:VII-E.

TABLE VIII
MINING UNKNOWN INSTANCES FOR OPS: WE SHOW THE NEGATIVE IMPACT ON PANOPTIC SEGMENTATION PERFORMANCE OF UNKNOWN CLASSES BY PROGRESSIVELY SUBTRACTING GLOBAL MASK ATTENTION AND MINING UNKNOWN INSTANCES COMPONENTS

	PQ ↑	SQ ↑	RQ ↑
ours	14.6	76.2	19.1
- mining unknown instances	9.1	72.9	12.8
- global mask attention	9.0	72.3	11.3

chose two subsets of 300 MS-COCO images (S1, S2) as our outlier dataset for training Mask2Anomaly and DenseHybrid. Table IX shows the performance of Mask2Anomaly and Dense Hybrid trained on S1 and S2 outlier sets, along with the standard deviation(σ) in the performance. We can observe that the variation in performance for the dense hybrid is significantly higher than Mask2Anomaly. Specifically, in dense hybrid, the average deviation in AuPRC is greater than 300%, and the average variation in FPR_{95} is more than 200% compared to Mask2Anomaly.

Reducing the supervision gap: In our previous discussion, we show models that are trained with outlier supervision have varying performance across different sets of outliers. So, we extend the previous discussion by demonstrating the performance of Mask2Anomaly without reliance on outlier supervision. We evaluate the performance of all the baselined method average over the validation dataset of FS static, FS L&F, SMIYC-RA21 and SMIYC-RO21. Fig. 13 shows the performance of Mask2Anomaly with or without outlier supervision names as Mask2Anomaly *(w OS)* and Mask2Anomaly *(w/o OS)*, respectively. In the plot, we can see unequivocally that Mask2Anomaly *(w/o OS)* significantly reduces the anomaly segmentation performance gap between the methods with outlier supervision and notably outperforms methods that do not use outlier supervision.

Outlier Loss: In this discussion, we will examine the efficacy of mask contrastive loss in anomaly segmentation. We empirically demonstrate why mask contrastive loss, a margin-based loss, performs better at anomaly segmentation by comparing it with binary cross-entropy loss as an outlier loss. So, we train Mask2Anomaly with M_{OOD} using binary-cross entropy which equates the outlier loss as

$$L_{\text{BCE}} = M_{\text{OOD}} \log(l_N) + (1 - M_{\text{OOD}}) \log(1 - l_N), \quad (28)$$

and, the new total loss at the outlier learning stage becomes

$$L_{\text{ood}} = L_{\text{BCE}} + L_{\text{masks}} + \lambda_{\text{ce}} L_{\text{ce}}, \quad (29)$$

TABLE IX
PERFORMANCE STABILITY OF MASK2FORMER: WE CAN OBSERVE THAT THE AVERAGE PERFORMANCE DEVIATION IN DENSE HYBRID IS SIGNIFICANTLY HIGHER THAN MASK2ANOMALY

Methods	SMIYC-RA21		SMIYC-RO21		FS L&F		FS Static		Average σ	
	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC ↑	FPR_{95} ↓	AuPRC	FPR_{95}
Mask2Anomaly-S1	95.48	2.41	92.89	0.15	69.41	9.46	90.54	1.98	-	-
Mask2Anomaly-S2	92.03	3.22	92.3	0.27	69.19	13.47	85.63	5.06	-	-
σ(Mask2Anomaly)	± 2.44	±0.57	±0.42	±0.08	±0.16	±2.84	±3.47	±2.18	±1.62	±1.41
Dense Hybrid-S1	52.99	38.87	66.91	1.91	56.89	8.92	52.58	6.03	-	-
Dense Hybrid-S2	60.59	32.14	79.64	1.01	47.97	18.35	54.22	5.24	-	-
σ(Dense Hybrid)	±5.37	±4.76	±9.00	±0.64	±6.31	±6.67	±1.16	±0.56	±5.46	±3.15

σ denotes the standard deviation.

TABLE X
ARCHITECTURAL EFFICIENCY OF MASK2ANOMALY: MASK2ANOMALY OUTPERFORMS THE BEST PERFORMING MASK2FORMER ARCHITECTURE WITH SWIN-S AS BACKBONE BY USING ALMOST 30% TRAINABLE PARAMETERS

Method	Backbone	AuPRC↑	FPR_{95} ↓	FLOPs↓	Training ↓ Parameters
Mask2Former [14]	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**

Mask2Anomaly‡ only uses global mask attention.

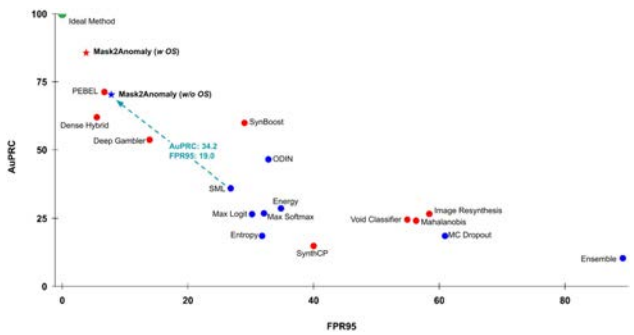

Fig. 13. *Bridging the supervision gap:* In this figure, we represent methods that utilize outlier supervision in red, and those without outlier supervision are in blue. We can observe Mask2Anomaly *(w/o OS):* Mask2Anomaly without using outlier supervision, shows significant performance gain among anomaly segmentation methods that do not use any extra supervision. Also, displays a similar performance to PEBEL, which is the best per-pixel method that utilizes additional supervision).

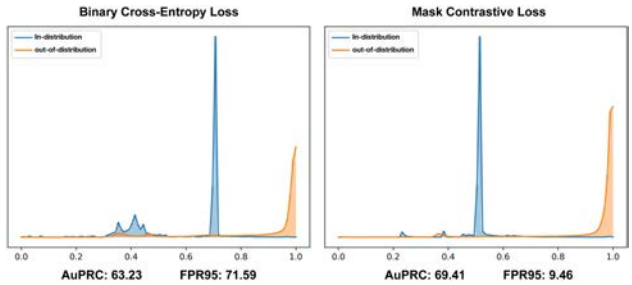

Fig. 14. *Outlier Loss Comparision:* During the training Mask2Anomaly, on the outlier set, we find that incorporating a mask contrastive loss, which is a margin-based loss function, resulted in better performance compared to the conventional binary cross-entropy loss. These experiments were conducted on the FS L&F validation set.

Fig. 15. *Failure Cases:* (a, b) Mask2Anomaly fail to segment trailer or carriage as they have a similar appearance as car or bus. Mask2Anomaly struggle to perform well in poor illumination (c) and weather conditions (d). The white region represents the anomalies that are enclosed in red boxes.

l_N is the negative likelihood of in-distribution classes calculated using the class scores C and class masks M. Fig. 14 illustrates the anomaly segmentation performance comparison on FS L&F validation dataset between the Mask2Anomaly when trained with the binary cross entropy loss and mask contrastive loss, respectively. We can observe that the mask contrastive loss achieves a wider margin between out-of-distribution(anomaly) and in-distribution prediction while maintaining significantly lower false positives.

Global Mask Attention: The application of global mask attention in semantic segmentation has shown a positive impact on performance, as demonstrated in Table V(c). So, we further investigate to assess the generalizability of this positive effect on ADE20K [67] and Vistas [50]. To evaluate the possible benefits of global mask attention, we trained the Mask2Former architecture using both masked attention and global masked attention for 40 thousand iterations. Mask2Former performed mIoU scores of 43.20 and 38.17 on masked attention, while global mask attention yields better mIoU scores of 43.80 (+0.6) and 38.92 (+0.75) on Ade20 K and Vistas, respectively.

Failure Cases: Fig. 15 illustrates the failure cases predicted by Mask2Anomaly. It is apparent that Mask2Anomaly faces difficulties when anomalies exhibit a resemblance to in-distribution classes like cars or buses, as shown in Fig. 15(a) and (b). In Fig. 15(c) shows increased false positives around anomalies when illumination conditions are poor. Weather conditions adversely effects Mask2Anomaly performance as seen in Fig. 15(d). We think that improving anomaly segmentation in such scenarios would be a promising avenue for future research.

IX. CONCLUSION

In this work, we introduce Mask2Anomaly, a universal architecture that is designed to jointly address anomaly and open-set segmentation utilizing a mask-based architecture. Mask2Anomaly incorporates a global mask attention

mechanism specifically to improve the attention mechanism for anomaly and open-set segmentation tasks. For the anomaly segmentation task, we propose a mask contrastive learning framework that leverages outlier masks to maximize the distance between anomalies and known classes. Furthermore, we introduce a mask refinement technique aimed at reducing false positives and improving overall performance. For the open-set segmentation task, we developed a novel approach to mine unknown instances based on mask-architecture properties. Through extensive qualitative and quantitative analysis, we demonstrate the effectiveness of Mask2Anomaly and its components. Our results highlight the promising performance and potential of Mask2Anomaly in the field of anomaly and open-set segmentation. We believe this work will open doors for a new development of novel anomaly and open-set segmentation approaches based on masked architecture, stimulating further advancements in the field.

ACKNOWLEDGMENT

The authors would like to thank Francesco Pappariello, Principal Engineer, Fabien Castanier, Program Manager, and Viviana D.alto, Artificial Intelligence Software & Tools Research Platform Director, of STMicroelectronics for their (guidance and) support throughout this work. The would also like to acknowledge that this manuscript reflects only the authors. views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

[1] J. Ackermann, C. Sakaridis, and F. Yu, "Maskomaly: Zero-shot mask anomaly segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2023, p. 329.
[2] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[3] P. Bevandić, I. Krešo, M. Oršić, and S. Šegvić, "Simultaneous semantic segmentation and outlier detection in presence of domain shift," in *Proc. German Conf. Pattern Recognit.*, 2019, pp. 33–47.
[4] P. Bevandić, I. Krešo, M. Oršić, and S. Šegvić, "Dense outlier detection and open-set recognition based on training with noisy negative images," 2021, *arXiv:2101.09193*.
[5] A. Bieniek and A. Moga, "A connected component approach to the watershed segmentation," in *Proc. 4th Int. Symp. Math. Morphol. Appl. Image Signal Process.*, 1998, pp. 215–222.
[6] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena, "The fishyscapes benchmark: Measuring blind spots in semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3119–3135, 2021.
[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
[8] J. Cen, P. Yun, J. Cai, M. Y. Wang, and M. Liu, "Deep metric learning for open world semantic segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 15313–15322.
[9] F. Cermelli, D. Fontanel, A. Tavera, M. Ciccone, and B. Caputo, "Incremental learning in semantic segmentation from image labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4361–4371.
[10] F. Cermelli, M. Mancini, S. R. Bulò, E. Ricci, and B. Caputo, "Modeling the background for incremental learning in semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9230–9239.
[11] R. Chan et al., "SegmentMeIfYouCan: A benchmark for anomaly segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021.
[12] R. Chan, M. Rottmann, and H. Gottschalk, "Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5108–5117.
[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.
[14] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1280–1289.
[15] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 17864–17875.
[16] M. Cordts et al., "The CityScapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[18] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-wise anomaly detection in complex driving scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16913–16922.
[19] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Conf. Robot Learn.*, 2017, pp. 1–16.
[20] A. Douillard, Y. Chen, A. Dapogny, and M. Cord, "PLOP: Learning without forgetting for continual semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4039–4049.
[21] X. Du, B. Zoph, W.-C. Hung, and T.-Y. Lin, "Simple training strategies and model scaling for object detection," 2021, *arXiv:2107.00057*.
[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[23] D. Fontanel, F. Cermelli, M. Mancini, and B. Caputo, "Detecting anomalies in semantic segmentation with prototypes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2021, pp. 113–121.
[24] G. Franchi, A. Bursuc, E. Aldea, S. Dubuisson, and I. Bloch, "One versus all for deep neural network incertitude (OVNNI) quantification," 2020, *arXiv: 2006.00954*.
[25] G. Franchi, A. Bursuc, E. Aldea, S. Dubuisson, and I. Bloch, "TRADI: Tracking deep neural network weight distributions," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 105–121.
[26] M. Grcić, P. Bevandić, and S. Šegvić, "Dense open-set recognition with synthetic outliers generated by real NVP," in *Proc. Int. Joint Conf. Comput. Vis. Imag. Comput. Graph. Theory Appl.*, 2021, pp. 133–143.
[27] M. Grcić, P. Bevandić, and S. Šegvić, "DenseHybrid: Hybrid anomaly detection for dense open-set recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 500–517.
[28] M. Grcic, J. Saric, and S. Segvic, "On advantages of mask-level recognition for outlier-aware segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2023, pp. 2937–2947.
[29] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[31] D. Hendrycks, S. Basart, M. Mazeika, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 8759–8773.
[32] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Representations*, 2017.
[33] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. Int. Conf. Learn. Representations*, 2019.
[34] J. Hwang, S. W. Oh, J.-Y. Lee, and B. Han, "Exemplar-based open-set panoptic segmentation network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1175–1184.
[35] S. Jung, J. Lee, D. Gwak, S. Choi, and J. Choo, "Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 15405–15414.
[36] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9396–9405.
[37] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6402–6413.
[38] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7167–7177.

[39] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *Proc. Int. Conf. Learn. Representations*, 2018.
[40] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5168–5177.
[41] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[42] K. Lis, K. Nakka, P. Fua, and M. Salzmann, "Detecting the unexpected via image resynthesis," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 2152–2161.
[43] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, Art. no. 1802.
[44] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[45] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019.
[46] U. Michieli and P. Zanuttigh, "Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1114–1124.
[47] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.
[48] J. Mukhoti and Y. Gal, "Evaluating Bayesian deep learning methods for semantic segmentation," 2018, *arXiv: 1811.12709*.
[49] N. Nayal, M. Yavuz, J. F. Henriques, and F. Güney, "RbA: Segmenting unknown regions rejected by all," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 711–722.
[50] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 5000–5009.
[51] P. Pinggera, S. Ramos, S. Gehrig, U. Franke, C. Rother, and R. Mester, "Lost and found: Detecting small road hazards for self-driving vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 1099–1106.
[52] S. N. Rai, F. Cermelli, D. Fontanel, C. Masone, and B. Caputo, "Unmasking anomalies in road-scene segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 4014–4023.
[53] M. Rottmann et al., "Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–9.
[54] R. Sun, X. Zhu, C. Wu, C. Huang, J. Shi, and L. Ma, "Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4355–4364.
[55] Y. Sun, C. Guo, and Y. Li, "ReAct: Out-of-distribution detection with rectified activations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 144–157.
[56] A. Tavera, F. Cermelli, C. Masone, and B. Caputo, "Pixel-by-pixel cross-domain alignment for few-shot semantic segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1959–1968.
[57] Y. Tian, Y. Liu, G. Pang, F. Liu, Y. Chen, and G. Carneiro, "Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 246–263.
[58] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[59] T. Vojir, T. Šipka, R. Aljundi, N. Chumerin, D. O. Reino, and J. Matas, "Road anomaly detection by partial image reconstruction with segmentation coupling," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 15631–15640.
[60] Y. Xia, Y. Zhang, F. Liu, W. Shen, and A. Yuille, "Synthesize then compare: Detecting failures and anomalies for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 145–161.
[61] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4084–4094.
[62] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.
[63] D. Zhang, K. Sakmann, W. Beluch, R. Hutmacher, and Y. Li, "Anomaly-aware semantic segmentation via style-aligned ood augmentation," in *Proc. Int. Conf. Comput. Vis. Workshop*, 2023, pp. 4067–4075.
[64] J. Zhang, Z. Chen, J. Huang, L. Lin, and D. Zhang, "Few-shot structured domain adaptation for virtual-to-real scene parsing," in *Proc. Int. Conf. Comput. Vis. Workshop*, 2019, pp. 9–17.
[65] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "ExFuse: Enhancing feature fusion for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 273–288.
[66] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[67] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5122–5130.
[68] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2021.

Shyam Nandan Rai received the master's degree in computer science from IIIT Hyderabad, in 2020. He is currently working toward the ELLIS PhD degree in computer and control engineering with the Politecnico di Torino, funded by Italian National PhD Program in Artificial Intelligence. He is a member of the Visual Learning and Multimodal Applications Laboratory (VANDAL), supervised by Prof. Barbara Caputo and Prof. Carlo Masone.

Fabio Cermelli received the PhD degree in computer and control engineering from the Politecnico di Torino, funded by the Italian Insitute of Technology (IIT). He is the CTO of Focoos AI. During that time, he was member of the Visual Learning and Multimodal Applications Laboratory (VANDAL), supervised by Prof. Barbara Caputo. During his first year, he was a visiting PhD student in the Technologies of Vision Laboratory with FBK.

Barbara Caputo received the PhD degree in computer science from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2005. From 2007 to 2013, she was a senior researcher with Idiap-EPFL (CH). Then, she moved to Sapienza Rome University thanks to a MUR professorship, and joined the Politecnico di Torino, in 2018. Since 2017, she has been a double affiliation with the Italian Institute of Technology (IIT). She is currently a full professor with the Politecnico of Torino, where she leads the Hub AI at PoliTo. She is one of the 30 experts who contributed to write the Italian Strategy on AI, and coordinator of the Italian National PhD on AI and Industry 4.0, sponsored by MUR. She is also an ERC Laureate and an ELLIS fellow. Since 2019, she serves on the ELLIS Board.

Carlo Masone (Member, IEEE) received the BS and MS degrees in control engineering from the Sapienza University, Rome, Italy, in 2006 and 2010 respectively, and the PhD degree in control engineering from the University of Stuttgart in collaboration with the Max Planck Institute for Biological Cybernetics (MPI-Kyb), Stuttgart, Germany, in 2014. He is currently an assistant professor with Politecnico di Torino and member of the Turin ELLIS Unit. From 2014 to 2017 he was a postdoctoral researcher with MPI-kyb, within the Autonomous Robotics & Human-Machine Systems group. From 2017 to 2020 he worked in industry on the development of self-driving cars. From 2020 to 2022 he was a senior researcher with the Visual and Multimodal Applied Learning, Politecnico di Torino.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement