Abstract

Semantic segmentation methods typically perform per-pixel classification by assuming a fixed set of semantic categories. While they perform well on the known set, the network fails to learn the concept of objectness, which is necessary for identifying unknown objects. In this paper, we explore the potential of query-based mask classification for unknown object segmentation. We discover that object queries specialize in predicting a certain class and behave like one vs. all classifiers, allowing us to detect unknowns by finding regions that are ignored by all the queries. Based on a detailed analysis of the model’s behavior, we propose a novel anomaly scoring function. We demonstrate that mask classification helps to preserve the objectness and the proposed scoring function eliminates irrelevant sources of uncertainty. Our method achieves consistent improvements in multiple benchmarks, even under high domain shift, without retraining or using outlier data. With modest supervision for outliers, we show that further improvements can be achieved without affecting the closed-set performance.

1. Introduction

We address the problem of semantic segmentation of unknown categories. Detecting novel objects, for example, in front of a self-driving vehicle is crucial for safety yet very challenging. The distribution of potential objects on the road has a long tail of unknowns such as wild animals, vehicle debris, litter, etc., manifesting in small quantities on the existing datasets [6,17,73]. The diversity of unknowns in terms of appearance, size, and location adds to the difficulty. In addition to the challenges of data, deep learning has evolved around the closed-set assumption. Most existing models for category prediction owe their success to curated datasets with a fixed set of semantic categories. These models fail in the open-set case with unknown objects by over-confidently assigning the labels of known classes to unknowns [32,58].

The existing approaches to segmenting unknown objects can be divided into two depending on whether they use supervision for unknown objects or not. In either case, the model has access to known classes during training, i.e. inlier or in-distribution, and the goal is to identify the pixels belonging to an unknown class, i.e. anomalous, outlier, or out-of-distribution (OoD). Earlier approaches resort to an ensemble of models [39] or Monte Carlo dropout [22] which require multiple forward passes, therefore costly in practice. More recent approaches use the maximum class probability [34] predicted by the model as a measure of its confidence. However, this approach requires the probability predictions to be calibrated, which is not guaranteed [26,37,54,58,64]. In the supervised case, the model can utilize outlier data to learn a discriminative representation, however, outlier data is limited. Typically, another dataset from a different domain is used for this purpose [11], or outlier objects are artificially added to driving images [25,65]. Despite the better performance [65], outlier supervision can bias the model to a particular distribution of outliers.

The common approach to segmentation is per-pixel classification by making a decision for each pixel independently. For spatial smoothness or objectness, the models can rely on...
Figure 2. Overview. This figure provides an overview of the training pipeline based on the Mask2Former architecture [14] and the inference of unknown objects. The model predicts region class probabilities $P$ and membership scores $M$. After matching regions to the ground truth via bipartite matching, we train the model with a semantics loss on $P$ and a mask loss on $M$. For unknown inference, we aggregate votes from object queries in logits $L$ as a result of multiplying $P$ and $M$. We compute the anomaly score as the negative sum over the class dimension $K$. For $\clubsuit\spadesuit$, see Section 3.2.

learning necessary inductive biases from data. However, unknown segmentation methods built on this paradigm fail to preserve the object boundaries or to uniformly assign a score to the whole object region as shown in Fig. 1. Preserving objectness is important for the following steps such as incremental learning in open-world settings [9]. In this work, we explore the potential of classifying regions instead of pixels to better learn the concept of objectness for unknown object segmentation. Region or mask classification has been initially proposed for instance segmentation [31] and recently extended to include semantic segmentation [14, 15].

We explore the potential of query-based mask classification to discover unknown objects. Specifically, we found that object queries in Mask2Former [14] behave like one vs. all region classifiers. Object queries specialize in predicting the region of a particular class and unknown regions are ignored by all the queries. Each object query casts a vote representing its confidence that a specific pixel belongs to a particular region. By interpreting these votes as logits, we show that simple baselines such as max logit [33] become considerably more reliable for detecting unknowns and even achieve state-of-the-art performance on some datasets.

The accuracy and false positive rate are commonly reported together in the literature because a good accuracy at detecting unknowns is typically achieved at the cost of a high false positive rate. Furthermore, different sources of uncertainty make predicting unknowns difficult, especially at the boundaries separating inliers or in ambiguous background regions. To reduce false positives, we examine the behavior of logits at each pixel and propose a novel anomaly scoring function called Negative Logit Sum (NLS). With mask classification and NLS, our method without supervision outperforms the existing methods that use outlier supervision on challenging datasets with high distribution shifts such as Road Anomaly [47] and SMIYC benchmark [10]. NLS does not require retraining, outlier supervision, or an external network. We show that regularizing NLS using outlier supervision further improves the performance without affecting the closed-set performance.

2. Related Work

Semantic Segmentation Paradigms: Since the appearance and success of Fully Convolutional Networks (FCN) [62], semantic segmentation architectures have revolved around the per-pixel classification paradigm. This paradigm has been extensively studied to increase the closed-set performance with various convolution and pooling operations [12, 13, 18, 71, 78], and to aggregate multi-scale contextual information [74, 75]. Recent work shifted towards transformer-based architectures [63, 70, 72, 76, 80] and attention mechanisms [21, 29, 30, 36, 41, 42, 79].

On the other hand, mask classification has been mainly adopted by instance segmentation and object detection models [7, 28, 31] since it allows pixels to belong to multiple proposals and provides the flexibility to detect a variable number of objects in the scene. Max-DeepLab [68] employs mask classification for panoptic segmentation but with many auxiliary losses. Although some earlier efforts have been made to apply mask classification to semantic segmentation [8, 28], they were quickly outperformed by the per-pixel methods until recently. MaskFormer variants [14, 15] apply query-based mask classification and attention to obtain a unified segmentation model which shows competitive performance with specialized semantic and instance segmentation architectures across benchmarks [17, 45, 57, 81].

Unsupervised Anomaly Segmentation: Unsupervised methods utilize their knowledge about inlier data to detect anomalies at inference time. Early work measures uncertainty based on the observation that anomaly samples typically result in low-confidence predictions. The uncertainty of a model can be estimated through maximum softmax probabilities [34, 44], ensembles [39], Bayesian approximation [55], Monte Carlo dropout [22], or by learning to estimate its confidence [19]. However, posterior probabilities of a closed-set model are not necessarily calibrated, leading to overconfident predictions on unseen categories [26, 37, 54, 58, 64]. Therefore, follow-up work focuses on making a clear distinction between inliers and outliers by using true class probabilities [16], unnormalized logits instead of softmax probabilities [33], standardized class-wise logits [38], and the distance to learned prototypes of known classes [9]. Overall, unsupervised approaches are typically efficient without any extra training but they are inherently limited to which extent they can separate inliers and outliers due to a lack of supervision with outlier data.
Deep generative models are also used for unsupervised anomaly segmentation. Early methods are primarily based on density estimation [40, 59] while subsequent works focus on reconstruction. Several works rely on the predicted segmentation maps to resynthesize [27, 47, 67, 69] or inpaint the inputs [46] and measure discrepancy with comparison networks. Others apply localized adversarial attacks [1], synthesize negatives using normalizing flows [24], or combine Gaussian mixture models with discriminative representation learning [43]. Generative methods are typically impractical for real-time safety-critical applications due to high computational costs and long inference times. Additional comparison modules and the change in input distributions require extra training. Moreover, synthesized unknowns often do not generalize well to real anomalies [24]. Several works [56, 61, 77] show that generative models tend to estimate high likelihoods on out-of-distribution samples.

**Anomaly Segmentation with Outlier Supervision:** Out-of-distribution data can be used to regularize the model’s feature space by learning a representation of unknowns. With the increase in the availability of wide-range datasets, initial approaches utilize generic datasets such as ImageNet [60] and 80 Million Tiny Images [66] for OoD. Given data, OoD detection can be simply treated as binary classification [2, 3]. Outlier data can also be used to estimate the distributional uncertainty of OoD samples [52] or to fine-tune parametrized OoD detectors [35]. The energy score has been proposed as a better alternative to softmax in terms of separation [49]. SynBoost [4] is a supervised image resynthesis method that treats void regions as anomalies to obtain an uncertainty signal. Recent work uses a subset of COCO [45] or ADE20K [81], either as entire images [11] or after cut-and-paste into the inlier scenes [25, 65]. Meta-OoD [11] maximizes the entropy on outliers, whereas PEBAL [65] learns adaptive energy-based penalties by abstention learning. Combining likelihood and posterior evaluation, DenseHybrid [25] achieves state-of-the-art results. However, different models are fine-tuned for each benchmark by using multiple datasets with high distribution shifts, resulting in a higher degree of supervision and variety. Our model, on the other hand, can achieve better performance by only using COCO [45] for fine-tuning.

### 3. Methodology

In open-set segmentation, the goal is to segment unknown objects without compromising closed-set performance. We approach this challenge by exploring mask classification and uncovering its innate ability to detect semantically unknown regions. Based on our analysis, we propose a novel anomaly score function enabled by the unique properties of mask classification. We further show that minimal fine-tuning using outliers improves unknown detection without affecting closed-set performance.
3.1. Mask Classification

We build our method on top of the Mask2Former architecture [14], which is an improved version of the initial MaskFormer [15]. We give only a brief overview to make the discussion self-contained; please refer to Cheng et al. [14] for details. Mask2Former consists of three main parts: the backbone, the pixel decoder, and the transformer decoder. The backbone processes the input image to extract features at multiple scales. Then, the pixel decoder further processes the multi-scale features to produce high-resolution per-pixel features $\mathbf{F} \in \mathbb{R}^{C_p \times H \times W}$. The transformer decoder takes the resulting multi-scale features $\{\mathbf{f}_i\}_{i=1}^D$ as well as $N$ learnable object queries $\mathbf{Q} \in \mathbb{R}^{N \times C_q}$, where $C_p$ and $C_q$ denote the embedding dimensions. At each layer of the transformer decoder, object queries are refined by interacting with each other and with one of the scales $\mathbf{f}_i$ in a round-robin order.

The refined object queries are first processed with a 3-layer MLP, resulting in $\mathbf{Q}_p \in \mathbb{R}^{N \times C_p}$ to predict $N$ regions. The binary masks for all regions are obtained by multiplying $\mathbf{Q}_p$ with pixel features $\mathbf{F}$ and applying a sigmoid $\sigma$ to the result:

$$\mathbf{M} = \sigma(\mathbf{Q}_p \mathbf{F})$$  \hspace{1cm} (1)

$\mathbf{M} \in \mathbb{R}^{N \times H \times W}$ represents the membership score of each pixel belonging to a region. In parallel, refined object queries are fed to a linear layer followed by softmax to produce posterior class probabilities $\mathbf{P} \in [0, 1]^{N \times K}$ of $K$ classes.

In contrast to per-pixel semantic segmentation, the ground truth masks are partitioned into multiple binary masks such that each mask contains all the pixels that belong to a class. Then, bipartite matching is used to match every ground truth mask to an object query in a way to minimize the losses. For region classification, the cross-entropy loss is applied to the matched object query class predictions. For region prediction, a weighted combination of dice loss [53] and binary cross-entropy loss is applied to the binary mask predictions. The training steps are highlighted with a red box in Fig. 2. In inference, the class scores or logits $\mathbf{L} \in \mathbb{R}^{K \times H \times W}$ are calculated as the multiplication of mask predictions with class predictions by broadcasting the class prediction to all the pixels within the region:

$$\mathbf{L} = \sum_{n=1}^N \mathbf{P}_n \mathbf{M}_n$$ \hspace{1cm} (2)

3.2. One vs. All Behavior of Object Queries

The logit term $\mathbf{L}$ as defined in Eq. 2 has a deeper interpretation because of its structure. In essence, $\mathbf{L}$ aggregates weighted votes over all object queries to decide whether the pixel belongs to a certain class. There is an important difference between $\mathbf{L}$ and logits in standard per-pixel segmentation. In the standard setting, the logit that belongs to the true class is directly optimized to be higher than the logits of other classes. In the case of $\mathbf{L}$ in Eq. 2, regions $\mathbf{M}$ predicted by the object queries are not constrained to be mutually exclusive or exhaustively cover the entire image. This allows a pixel to be assigned to more than one region or possibly to none. Independent region predictions lead to independent votes from object queries about the per-class membership of a pixel. This allows object queries to behave implicitly as one vs. all classifiers, where each object query predicts a certain class, leaving ignored regions to be interpreted as unknowns.

We empirically demonstrate this behavior on the BDD100K validation set [73]. We first identify which class each object query specializes in by counting how many times it predicts a certain class with high confidence, e.g. greater than 98%. As shown in Fig. 3, some object queries consistently focus on a single class. After identifying specialized object queries, we mask out other object queries and use only those to test our hypothesis on one vs. all behavior. We apply two types of masking: before the transformer decoder (♣ in Fig. 2), which enforces hard masking, or after the transformer decoder (♦ in Fig. 2), which allows them to still interact in the transformer, leading to soft masking. Fig. 4 shows per-class IoU using both strategies compared to the original model without masking. Due to the independence of object queries, the performance remains the same for frequent classes but drops significantly for rare classes, such as train, truck, or bus. The drop in performance indicates that when a class cannot be learned properly, e.g. due to data scarcity, the object query relies on interactions in the transformer decoder, which explains the slightly better performance of soft masking than hard masking.

3.3. Negative Logit Sum (NLS)

The term $\mathbf{L}$ in Eq. 2 aggregates the independent decisions of object queries about whether a pixel belongs to a certain class. Based on this behavior, we can identify several distinct modes of $\mathbf{L}$. We cluster the logits over classes at each pixel, i.e. $K$-dimensional vector, using k-means to characterize the modes, visualized in Fig. 5e. For an inlier pixel, only a single object query will vote for it with high confidence (Fig. 5b), whereas true outlier pixels do not receive any votes from any object query (Fig. 5a). Simply using max logit as an anomaly score already achieves great performance in detecting unknowns (Fig. 6 and Table 2), since the maximum will be high for inliers and low for outliers. However, there are pixels that disrupt the separability between the inliers and the outliers. For example, pixels on a boundary between two inlier classes (Fig. 5c) or ambiguous background pixels (Fig. 5d) end up with a max logit lower than the inliers, causing them to be mistaken as an outlier. To mitigate this, we propose to use the negative sum of the logits as an anomaly
score instead of max logits:

\[
\text{NLS} = - \sum_{k=1}^{K} L_k
\]

where \(L_k\) is the logit of class \(k\). Boundary and ambiguous regions are commonly characterized by having more than one weak vote from object queries. By summing these weak votes, NLS can assign a lower anomaly score than the max logit to these regions. This simple function brings a considerable decrease in false positive rate, especially for small OOD objects (Table 2). Fig. 6 highlights the differences between the anomaly maps predicted by max logit and NLS.

3.4. Outlier Supervision

Given minimal outlier supervision, NLS can be directly regularized leading to significant improvements in segmenting unknowns without affecting closed-set performance. Without the need to retrain the entire model, we fine-tune the MLP and the linear layers after the transformer decoder (see Supplementary), which constitutes only the total model parameters. We use Anomaly Mix as proposed in [65], where objects from COCO dataset [45] are cut and pasted on Cityscapes images [17]. Regularization is achieved by encouraging NLS to be higher than some threshold \(\alpha_{\text{out}}\) for outliers and below some threshold \(\alpha_{\text{in}}\) for inliers. The additional outlier loss function \(L_{\text{nls}}\) is defined as the average of the two following loss functions:

\[
L_{\text{in}} = \sum_{p \in \Omega_{\text{in}}} (\max(0, \text{NLS}(p) - \alpha_{\text{in}}))^2
\]

\[
L_{\text{out}} = \sum_{p \in \Omega_{\text{out}}} (\max(0, \alpha_{\text{out}} - \text{NLS}(p)))^2
\]

where \(\Omega_{\text{in}}\) and \(\Omega_{\text{out}}\) are the set of inlier and outlier pixels, respectively. The hyper-parameters \(\alpha_{\text{in}}\) and \(\alpha_{\text{out}}\) can be chosen without tuning by considering the properties of \(L\). The logits depicted in Fig. 5a and Fig. 5b correspond to the ideal form that inlier and outlier logits should take, respectively.

For inlier pixels, it is sufficient to be chosen by a single object query, therefore \(\alpha_{\text{in}} = -1\) suffices, and the same logic applies to \(\alpha_{\text{out}}\) for being close to zero.

4. Experiments

4.1. Datasets

We train the model on the Cityscapes dataset [17], which consists of 2975 training and 500 validation images. It contains 19 classes which are considered as inliers in anomaly segmentation benchmarks. The classes in the dataset can be seen in Fig. 3. For evaluation, we consider multiple datasets. First, Segment Me If You Can (SMIYC) Benchmark [10] contains two datasets: anomaly track and obstacle track. The anomaly track has 100 images that contain unknown objects of various sizes in diverse environments. The obstacle track contains 412 images with typically small unknown objects on the road, 85 of which are taken at night and in adverse weather conditions. Both datasets are characterized by a high domain shift compared to Cityscapes, making them particularly challenging. Road Anomaly [47] is an earlier and smaller version of SMIYC. It consists of 60 images with diverse objects in diverse environments. We report results on the Fishyscapes Lost&Found dataset [5], which has 100 validation and 275 test images. The domain of this dataset is similar to that of Cityscapes, and the anomalous objects are mostly small and less diverse compared to other datasets.

4.2. Experimental Setup

Implementation Details: We follow the setup of [14] for closed-set training on Cityscapes. We use the Swin-B [50] architecture as the backbone. Differently, we use only one decoder layer in the transformer decoder instead of nine (see Section 4.4). For outlier supervision, we fine-tune the MLP and the linear layers for 5K iterations with a batch size of 16 using the standard loss functions used in [14] in addition to the losses defined in Eq. 4. The previous work [65] samples 300 new images every epoch out of 40K COCO images with
Figure 6. Max Logs vs NLS. While max logit [33] performs well in segmenting unknowns, the proposed NLS eliminates falsely high anomaly scores at the boundaries separating inliers (marked in ellipses), thereby reducing the false positive rate considerably (see Table 2).

| Method                  | OoD Extra | Anomaly Track | Obstacle Track |
|-------------------------|-----------|---------------|----------------|
|                         | Data Net. | AP ↑ FPR@95 ↓ sIoU gt ↑ PPV ↑ mean F1 ↑ | AP ↑ FPR@95 ↓ sIoU gt ↑ PPV ↑ mean F1 ↑ |
| Emb. Density [5]        | ×         | 37.5 70.8 33.9 20.5 7.9                | 0.8 46.4 35.6 2.9 2.3                |
| JSRNet [67]             | ×         | 33.6 43.9 20.2 29.3 13.7               | 28.1 28.9 18.6 24.5 11.0               |
| Road Inpain. [46]       | ×         | - - - - - - - - -                      | 54.1 47.1 57.6 39.5 36.0               |
| Image Resyn. [47]       | ×         | 52.3 25.9 39.7 11.0 12.5               | 37.7 4.7 16.6 20.5 8.4                |
| ObsNet [1]              | ×         | 75.4 26.7 44.2 52.6 45.1               | - - - - - - - - -                      |
| NFlowJS [24]            | ×         | 56.9 34.7 36.9 18.0 14.9               | 85.6 0.4 45.5 49.5 50.4               |
| Ours (NLS)              | ×         | 86.1 15.9 56.3 41.4 42.0               | 87.8 3.3 47.4 56.2 50.4               |
| Max. Entropy [42]       | ×         | 85.5 15.0 49.2 39.5 28.7               | 85.1 0.8 47.9 62.6 48.5               |
| DenseHybrid [25]        | ×         | 78.0 9.8 54.2 24.1 31.1               | 87.1 0.2 45.7 50.1 50.7               |
| PEBAL [65]              | ×         | 49.1 40.8 38.9 27.2 14.5               | 5.0 12.7 29.9 7.6 5.5                |
| SynBoost [4]            | ×         | 56.4 61.9 34.7 17.8 10.0               | 71.3 3.2 44.3 41.8 37.6               |
| Ours (NLS)              | ×         | 91.3 10.7 57.8 49.0 46.6               | 91.7 0.4 54.3 59.1 57.4               |

Table 1. Results on the SMIYC Benchmark. We report results on both the anomaly and the obstacle track. Both tracks cover a wide variety of scenarios and unknown objects. We report both pixel-level (AP, FPR@95) and component-level metrics (sIoU, PPV, mean F1). We show the results with (lower part) and without (upper part) outlier supervision with the best in bold and the second best underlined for each.

objects that do not intersect with Cityscapes inliers. In our case, we sample only 300 images at the beginning and fix them, then an image is randomly chosen and pasted on the Cityscapes images with probability $p_{out}$. We experimentally set $p_{out}$ to 0.1.

Evaluation Metrics: For comparison to previous methods on the Road Anomaly and the Fishyscapes, we report Average Precision (AP), Area under ROC Curve (AuROC), and False Positive rate at the threshold of 95% True Positive Rate (FPR@95). On SMIYC, the public benchmark reports AP and FPR@95 for per-pixel metrics as well as component-level metrics that are designed to measure the statistics of detected objects [10]. Specifically, the proposed metrics aim at quantifying true positives (TP), false negatives (FN), and false positives (FP) of detected unknown objects. Please see the benchmark paper [10] for more details on these metrics.

4.3. Quantitative Results

4.3.1 Segment Me If You Can Benchmark

Table 1 shows the results on anomaly and obstacle tracks of the public SMIYC benchmark. Without outlier supervision, NLS outperforms all the models, including those trained with outlier supervision, in AP while maintaining a competitive FPR@95. In terms of component metrics, the gains with NLS are more pronounced, which is due to an improved ability to characterize objectness, compared to the previous work. With outlier supervision, the performance gap improves with respect to the previously best method consistently across both tracks: +5.8% and +4.6% in AP and +1.5% and +6.7% in mean F1 for anomaly and obstacle tracks respectively.

DenseHybrid [25] achieves a slightly better FPR@95 on the anomaly and obstacle tracks but NLS achieves +13.3% and +4.6% better AP respectively, and better performance in all component-level metrics. ObsNet [1] achieves impressive performance at the component-level, however, not at the pixel-level. NLS consistently performs well across both tracks in both pixel and component-level metrics.

SMIYC is characterized by high domain shift and diversity of objects in terms of size and appearance, making it particularly challenging. While some methods, like DenseHybrid [25], rely on more diverse data when fine-tuning, NLS with mask classification shows that no outlier supervision is necessary for performing well under domain shift, thereby surpassing the limitations of the existing methods.

4.3.2 Road Anomaly & Fishyscapes LaF

Table 2 shows the results on the Road Anomaly [47] and the Fishyscapes Lost and Found (LaF) validation set [5]. In addition to NLS, we report the performance of max logit [33] based on logits as computed in Eq. 2. Max logit with mask classification outperforms all the existing methods owing
to the mostly correct behavior of logits (see Fig. 5). Using NLS results in a considerable improvement in FPR@95 on both datasets; by 4.22% on Road Anomaly and 8.69% on Fishyscapes LaF. This shows that NLS can better utilize the potential of mask classification. With outlier supervision, the results are further improved in all metrics. GMMSeg [43] and PEBAL [65] perform better in FPR@95 on Fishyscapes LaF but fall behind in AP with 10% and 5% gap, respectively. Furthermore, their success does not map to Road Anomaly.

Table 2 also shows the in-distribution performance on the validation set of Cityscapes. Fine-tuning NLS with outlier supervision does not hurt the closed-set performance, it rather improves it slightly. This can be attributed to the fact that detecting unknowns using mask classification is highly coupled with detecting known classes; when known classes are detected better, it becomes easier to find unknown regions ignored by known classes. Therefore, fine-tuning can improve unknown detection by enhancing closed-set performance.

### 4.4. Ablation Study

We validate the effectiveness of mask classification with Mask2Former [14] and the proposed scoring function NLS on Road Anomaly in comparison to a couple of per-pixel semantic segmentation methods with max logit [33] in Table 3. We choose PSA [48] due to its superior closed-set performance on Cityscapes validation set, and Segmenter [63] due to learnable object queries but without mask classification. Despite the impressive in-distribution performance of PSA, it performs poorly in outlier detection. Segmenter performs worse in terms of closed-set performance but better in terms of detecting outliers, proving the effectiveness of learnable object queries by specializing to each class. However, it cannot reach the performance of the Mask2Former with mask classification. NLS further improves the performance, especially by reducing the false positives compared to max logit and achieves the best performance in all metrics.

**Number of Decoder Layers:** Although more decoder layers improve the inlier performance, i.e. mIoU on Cityscapes, we found that using fewer decoder layers results in better performance in case of outliers. Fig. 7a highlights the decrease in performance in terms of the AP and FPR@95 on the Road Anomaly dataset as the number of decoder layers increases. We investigate this behavior by isolating the sources of error with respect to the number of decoder layers. Fig. 7b shows semantic and mask losses of the Mask2Former [14] averaged over the validation samples on Cityscapes. With more decoder layers, we observe that the semantic cross-entropy loss increases while the mask-related BCE and dice loss decrease. This shows that the increase in inlier mIoU with more decoder layers can be attributed to a better ability in detecting masks at the cost of a higher semantic error.

Fig. 7b also shows the mIoU evaluated by applying hard masking (see Fig. 4) on the specialized object queries. Specialized object queries perform worse with more decoder layers. The information loss in the object queries as well as the degradation in the semantic loss show the importance of semantics for outlier detection compared to precise masks.

**Number of Object Queries:** The original Mask2Former [14] uses 100 object queries. We train different models by varying the number of object queries to observe its effect on detecting outliers. We focus on the AP values on both Road Anomaly [47] with large objects and on Fishyscapes with an outlier supervision dataset.
Table 4a. The results improve with more iterations while the outlier performance drops in terms of AP and FPR@95 on Road Anomaly (a). Good performance is mainly due to better mask prediction at the cost of a higher semantics loss (b). The drop in mIoU with hard masking (HM mIoU) is another indicator of semantic information loss in the object queries with more decoder layers. The performance improves as the number of object queries increases (c).

Table 4b. Increasing the number of fine-tuning iterations while keeping low outlier exposure probability maintains a stable in-distribution performance. After 5K iterations, the average performance begins to drop (a). Increasing $p_{\text{out}}$ can lead to better AP at the cost of higher FPR@95 (b). A tighter threshold $\alpha_{\text{out}}$ does not show a particular improvement (c).

LaF [5] with small objects. Fig. 7c shows that more object queries result in better AP on both datasets. Even if some object queries specialize in predicting a particular class, the other object queries still play a role.

Outlier Data Exposure: To show the effect of the number of fine-tuning iterations, we show results in different steps averaged over five different runs on Road Anomaly [47] in Table 4a. The results improve with more iterations while maintaining a low standard deviation. We also show the in-distribution performance, mIoU on the Cityscapes validation set. Impressively, more fine-tuning with outlier data slightly improves in-distribution performance. To check the effect of the frequency of outlier data exposed in fine-tuning, we experiment with different values of $p_{\text{out}}$ as shown in Table 4b. We choose $p_{\text{out}} = 0.1$ as it strikes a reasonable balance between performance and the number of outliers needed to improve the performance.

Outlier Bounds Optimization: As discussed in Section 3.4, the structure of the logits eliminates the need to tune $\alpha_{\text{in}}$ and $\alpha_{\text{out}}$ for outlier supervision. We support that by running fine-tuning experiments with different values of $\alpha_{\text{out}}$, and report the results averaged over five different runs on Road Anomaly dataset [47] to account for the variance caused by sampling different outlier data. As shown in Table 4c, the AP and FPR@95 values do not show a general trend in terms of the mean. We use $\alpha_{\text{out}} = -0.1$ based on its slightly better performance on average.

5. Conclusion and Future Work

In this work, we explore the potential of mask classification to segment unknown objects. We analyze its properties and demonstrate its inherent ability to express uncertainty by operating as an implicit one vs. all classifier. This allows the existing methods, such as max logit [33], to perform considerably well. Upon analyzing the behavior of logits, we propose a novel anomaly score function that eliminates irrelevant sources of uncertainty such as inlier boundaries and ambiguous backgrounds, leading to a considerable decrease in false positive rates. We further demonstrate that with minimal fine-tuning in terms of memory and computation, the proposed anomaly score can achieve additional gains without affecting the in-distribution performance.

As this work represents an initial attempt to utilize mask classification for unknown segmentation, its properties can be further explored with potential improvements. Given the increased ability to preserve objectness, open-world incremental learning is one step closer as unknown masks are more reliable as a source of supervision. While current efforts are limited to static image datasets, temporal or depth information can provide important cues to detect unknowns.

Acknowledgements: We thank A. Kaan Akan, Ali Safaya, Gökay Aydemir, and Shadi Hamdan for their valuable feedback. This project is funded by the Royal Society Newton Fund Advanced Fellowship (NAF\R\2202237). Nazir Nayal and Msra Yavuz are supported by the KUIS AI fellowship.
References

[1] Victor Besnier, Andrei Bursuc, David Picard, and Alexandre Briot. Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation. In ICCV, 2021. 3, 6

[2] Petra Bevandic, Ivan Kreso, Marin Orsic, and Sinisa Segvic. Discriminative out-of-distribution detection for semantic segmentation. arXiv.org, 1808.07703, 2018. 3

[3] Petra Bevandic, Ivan Kreso, Marin Orsic, and Sinisa Segvic. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In GCPR, 2019. 3

[4] Giancarlo Di Biase, Hermann Blum, Roland Y. Siegwart, and César Cadena. Pixel-wise anomaly detection in complex driving scenes. In CVPR, 2021. 3, 6, 7, 14

[5] Hermann Blum, Paul-Edouard Sarlin, Juan I. Nieto, Roland Y. Siegwart, and César Cadena. The fishscapes benchmark: Measuring blind spots in semantic segmentation. IJCV, 129:3119–3135, 2021. 5, 6, 8, 13

[6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11618–11628, 2020. 1

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, 2020. 2

[8] João Carreira, Rui Caseiro, Jorge P. Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In ECCV, 2012. 2

[9] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In ICCV, 2021. 2

[10] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Y. Siegwart, Mathieu Salzmann, P. Fua, and Matthias Rottmann. SegmentMELYouCan: A benchmark for anomaly segmentation. arXiv.org, 2104.14812, 2021. 2, 5, 6

[11] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In ICCV, 2021. 1, 3, 7, 14

[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. PAMI, 40:834–848, 2018. 2

[13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018. 2

[14] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In CVPR, 2022. 2, 4, 5, 7, 12, 13

[15] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In NeurIPS, 2021. 2, 4

[16] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In NeurIPS, 2019. 2

[17] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In CVPR, 2016. 1, 2, 3, 5, 12, 13

[18] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In ICCV, 2017. 2

[19] Terrance Devries and Graham W. Taylor. Learning confidence for out-of-distribution detection in neural networks. arXiv.org, 1802.04865, 2018. 2

[20] Xianzhi Du, Barret Zoph, Wei-Chih Hung, and Tsung-Yi Lin. Simple training strategies and model scaling for object detection. arXiv.org, 2107.00057, 2021. 12

[21] J. Fu, J. Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In CVPR, 2019. 2

[22] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In ICML, 2016. 1, 2

[23] Golnaz Ghiasi, Yin Cui, A. Srinivas, Rui Qian, Tsung-Yi Lin, Ekin Dogus Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In CVPR, 2021. 12

[24] Matej Grcic, Petra Bevandi’c, and Sinisva vSegvi’c. Dense anomaly detection by robust learning on synthetic negative data. arXiv.org, 2112.12833, 2021. 3, 6, 14

[25] Matej Gric’c, Petra Bevand’ic, and Siniša Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In ECCV, 2022. 1, 3, 6, 13, 14, 15, 16

[26] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In ICML, 2017. 1, 2

[27] David Haldimann, Hermann Blum, Roland Y. Siegwart, and César Cadena. This is not what I imagined: Error detection for semantic segmentation through visual dissimilarity. arXiv.org, 2009.00676, 2019. 3

[28] Bharath Hariharan, Pablo Arbeláez, Ross B. Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In ECCV, 2014. 2

[29] Adam W. Harley, Konstantinos G. Derpanis, and Iasonas Kokkinos. Segmentation-aware convolutional networks using local attention masks. In ICCV, 2017. 2

[30] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multiscale filters for semantic segmentation. In ICCV, 2019. 2

[31] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In ICCV, 2017. 2

[32] Matthias Hein, Maxym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In CVPR, 2019. 1
[33] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohamadreza Mostajabi, Jacob Steinhardt, and Dawn Xiaohong Song. Scaling out-of-distribution detection for real-world settings. In ICML, 2022. 2, 6, 7, 8, 13, 17
[34] Dan Hendrycks and Kevin Gimpel. A baseline for detecting missclassified and out-of-distribution examples in neural networks. In ICLR, 2017. 1, 2, 7, 14
[35] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In ICLR, 2019. 3
[36] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, Humphrey Shi, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In ICCV, 2019. 2
[37] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In NeurIPS, 2018. 1, 2
[38] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene recognition. In ICCV, 2021. 2, 7, 14
[39] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In NeurIPS, 2017. 1, 2
[40] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In NeurIPS, 2018. 3, 7
[41] Hancho Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. In BMVC, 2018. 2
[42] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In ICCV, 2019. 2, 6
[43] Chen Liang, Wenguang Wang, Jiaxu Miao, and Yi Yang. GMMSeg: Gaussian mixture based generative semantic segmentation models. arXiv.org, 2210.02025, 2022. 3, 7, 14
[44] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In ICLR, 2018. 2
[45] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In ECCV, 2014. 2, 3, 5, 12
[46] Krzysztof Lis, Sina Honari, P. Fua, and Mathieu Salzmann. Detecting road obstacles by erasing them. arXiv.org, 2012.13633, 2020. 3, 6
[47] Krzysztof Lis, Krishna Kanth Nakka, Pascal V. Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In ICCV, 2019. 2, 3, 5, 6, 7, 8, 13, 14
[48] Huajun Liu, Fuqiang Liu, Xinyi Fan, and Dong Huang. Polarized self-attention: Towards high-quality pixel-wise regression. ARXIV, 2107.00782, 2021. 7
[49] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, NeurIPS, 2020. 3
[50] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In ICCV, 2021. 5, 12
[51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019. 12, 13
[52] Andrey Malinin and Mark John Francis Gales. Predictive uncertainty estimation via prior networks. In NeurIPS, 2018. 3
[53] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 3DV, 2016. 4
[54] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Ann Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In NeurIPS, 2021. 1, 2
[55] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. arXiv.org, 1811.12709, 2018. 2
[56] Eric Nalisnick, Akihiko Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don’t know? In ICLR, 2018. 3
[57] Gerhard Neufeld, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In ICCV, 2017. 2
[58] Anh M Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In CVPR, 2015. 1, 2
[59] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In NeurIPS, 2019. 3
[60] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. IJCV, 115:211–252, 2015. 3
[61] Joan Serrà, David Alvarez, V. Gómez, Olga Slizovskaia, Jose F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. arXiv.org, 1909.11480, 2020. 3
[62] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015. 2
[63] Robin Strudel, Ricardo Garcia Pinel, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In ICCV, 2021. 2, 7
[64] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv.org, 1312.6199, 2013. 1, 2
[65] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and G. Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In ECCV, 2022. 1, 3, 5, 6, 7, 12, 13, 14, 15, 16
[66] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. PAMI, 30:1958–1970, 2008. 3
[67] Tomas Vojir, Tomáš Šipka, Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino, and Jiri Matas. Road anomaly detection by partial image reconstruction with segmentation coupling. In ICCV, 2021. 3, 6

[68] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Loddon Yuille, and Liang-Chieh Chen. MaX-DeepLab: End-to-end panoptic segmentation with mask transformers. In CVPR, 2021. 2

[69] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In ECCV, 2020. 3, 7

[70] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José Manuel Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In NeurIPS, 2021. 2

[71] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In CVPR, 2018. 2

[72] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In NeurIPS, 2021. 2

[73] Fisher Yu, Haofeng Chen, Xin Wang, Wenyi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In CVPR, 2020. 1, 4

[74] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In ICLR, 2016. 2

[75] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In ECCV, 2020. 2

[76] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. HRFormer: High-resolution transformer for dense prediction. arXiv.org, 2110.09408, 2021. 2

[77] Lily H. Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In ICML, 2021. 3

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

[79] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. PSANet: Point-wise spatial attention network for scene parsing. In ECCV, 2018. 2

[80] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In CVPR, 2021. 2

[81] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In CVPR, 2017. 2, 3

[82] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In ICLR, 2021. 12
Appendix

Abstract

This supplementary document contains the implementation details that are necessary to reproduce our approach (Section A), additional quantitative results on the Fishyscapes Benchmark, and an ablation of which components to fine-tune with outlier supervision on the Road Anomaly and the Fishyscapes Lost&Found datasets (Section B), additional qualitative results to showcase our method compared to state-of-the-art (Section C) and highlight some challenging cases that cause failure (Section D).

A. Implementation Details

A.1. Architecture

Fig. 8 illustrates the full Mask2Former architecture along with our unknown inference procedure. The main components of the model are the backbone, the pixel decoder, and the transformer decoder. We explain the details of each next.

Backbone: We use the Swin-B variant as the backbone [50]. It can take an RGB image with any resolution higher than 32 × 32 as input and outputs feature maps at several resolutions to the pixel decoder. Specifically, the output feature maps are downsampled with strides 4 (x4), 8 (x8), 16 (x16), and 32 (x32) with respect to the input image.

Pixel Decoder: Following [14], the pixel decoder mainly consists of 6 layers of deformable attention (MSDeformAttn) [82]. The multi-scale feature maps with strides x4, x8, x16, and x32 are processed with MSDeformAttn layers to produce f1, f2, and f3, respectively. In [14], the three processed feature maps are passed to 9 transformer decoder layers in a round-robin fashion. However, we found that using a single layer in the transformer decoder works better for unknown detection. Therefore, we only pass the last layer f3 to the transformer decoder. The feature map x4 is processed with a 1 × 1 filter-size convolutional layer and then added to the processed feature map f1 after bilinear upsampling. Finally, the output is passed to a 3 × 3 convolutional layer to produce per-pixel features F ∈ R C′p×H×W, where C′p = 256 is the embedding dimension. The computation can be summarized as follows:

F = Conv3x3(Conv1x1(x4) + Upsample(f1)) \quad (5)

Transformer Decoder: Learnable object queries Q ∈ R N×Cq are fed to the transformer decoder layer to be processed with the feature maps from the pixel decoder, where N = 100 is the number of object queries and Cq = 256 is the embedding dimension. A single transformer decoder layer mainly consists of a cross-attention layer followed by self-attention and feed-forward network (FFN), each of which is followed by a LayerNorm. The cross-attention operation is performed with mask-attention where each object query only attends to regions it predicted in the previous layer. Since we use only a single layer, each object query attends to the region it predicts directly from the input feature map before being processed in the transformer decoder. Learnable positional embeddings are added to the object queries. The transformer decoder outputs a refined set of object queries Qr that predict the regions and classify them.

Region Class Prediction: The refined object queries Qr are fed into a single linear layer followed by a softmax to produce the class probability of each region P ∈ RN×K, where K is the number of classes.

Membership Maps Prediction: The refined object queries Qr are also fed into a 3-layer MLP, so that Qr’s dimensionality matches that of F. Then, Qr and F are multiplied before being fed into a sigmoid activation to produce the per-pixel membership maps M ∈ RN×H×W.

A.2. Closed-Set Training

Loss Functions: Before applying any loss function, bipartite matching is used to match object queries to ground truth binary masks, where each mask contains all the pixels of a certain class. The matching cost is computed as a weighted sum of the individual losses. The classification is performed with the cross-entropy loss. A weighted combination of dice loss and binary cross-entropy is used to predict regions.

Hyper-Parameters: Following [14], the model is trained for 90K iterations using a batch size of 16. AdamW [51] optimizer is used with 0.05 for weight decay and an initial learning rate of 10⁻⁴, which is reduced using a polynomial scheduler. The learning rate for the backbone is multiplied by 0.1.

Data Augmentation: We use the same augmentations as in [14]. First, the short side of the input image is resized by a scale uniformly chosen between [0.5 − 2]. Then a random crop of size 512 × 1024 is applied. After that, large-scale jittering augmentation [20, 23] is applied with a random horizontal flip.

A.3. Outlier Supervision

Data Sampling: We use AnomalyMix proposed in [65] for outlier supervision. After eliminating the samples that contain Cityscapes classes [17], around 40K images remain for outlier supervision on the COCO [45] dataset. For a single fine-tuning experiment, we randomly sample 300 images and fix them throughout the fine-tuning phase.

Fine-tuned Components: For all the fine-tuning experiments, we only fine-tune the 3-layer MLP and linear layers
Figure 8. Detailed Architecture. This figure provides a more detailed view of the Mask2Former [14] architecture, including our modifications and unknown inference computation. We use a single transformer decoder layer as opposed to the original implementation that uses 9 layers. Therefore, only a single scale feature $f_3$ from the last layer is passed from the pixel decoder to the transformer decoder. For outlier supervision, all modules are frozen except for the MLP and the linear layers shown in pink.

shown in pink in Fig. 8. Their weights together constitute approximately 0.21% of the entire model parameters.

**Hyper-Parameters:** After the model is trained on the closed-set setting, we fine-tune it for 5K iterations on Cityscapes [17] using the setting of the closed-set training; AdamW [51] optimizer with 0.05 weight decay and $10^{-4}$ initial learning with polynomial scheduling. For every Cityscapes image used in fine-tuning, an object from the 300 COCO samples is uniformly chosen and pasted on the Cityscapes image with probability $p_{out} = 0.1$, which is independent for each image. The NLS score for outlier pixels is optimized to be higher than $\alpha_{out} = -0.1$ and for inlier pixels to be lower than $\alpha_{out} = -1.0$ (see the Ablations in the main paper).

**B. Additional Quantitative Results**

Table 5 shows additional quantitative results on the Fishyscapes Benchmark [5]. Due to technical difficulties on the official benchmark management side, we could only evaluate the results of fine-tuning the pixel decoder for the outlier supervision result. As shown in Table 5, without outlier supervision, NLS achieves second-best AP in Fishyscapes LaF and best in all other metrics. With outlier supervision, NLS consistently improves AP on both sets with a wide margin and obtains the best FPR@95 on FS Static.

To avoid confusion about fine-tuning specific modules, we also report the results of an ablation study in Table 6. We analyze the effect of fine-tuning different parts of the model on validation sets of Road Anomaly [47] and Fishyscapes Lost and Found [5]. While on the Road Anomaly, the best AP is obtained by fine-tuning the MLP and the linear layers, on Fishyscapes LaF, the best results are obtained by fine-tuning the pixel decoder. We chose to report the results in the main paper by fine-tuning the MLP and the linear layers since they amount to a smaller number of parameters to fine-tune and do not compromise the closed-set performance.

**C. Additional Qualitative Results**

In Fig. 9 and Fig. 10, we show additional qualitative results of NLS both with and without outlier supervision, compared to the state-of-the-art methods PEBAL [65] and DenseHybrid [25]. Our method with mask classification better preserves the objectness. The proposed scoring function NLS reduces the false positives on the boundaries of inliers and ambiguous background regions compared to the baselines. These improvements can be observed more prominently on the obstacle track (Fig. 10) under adverse weather and lighting conditions. The anomaly maps of the baselines are extracted using the pre-trained models provided in the respective public repositories.

In Fig. 11 we show more examples from Fishyscapes LaF [5] that demonstrate the effect of NLS in eliminating false positive compared to Max Logit [33].

**D. Failure Cases**

We analyze some failure cases of our method in this section. The main reason for the failure cases is the high similarity to the inlier classes.

**Tractors and Boats:** As shown in Fig. 12, NLS fails to detect tractors and boats due to their similarity to truck and vehicle instances. While outlier supervision slightly improves
### Table 5. Results on Fishyscapes Benchmark.
We show the results with (lower part) and without (upper part) outlier supervision by highlighting the best in bold and the second best underlined for each metric. NLS with outlier supervision significantly improves the AP on both Lost&Found and Static. We achieve the best performance in terms of FPR@95 on the Static while maintaining a comparable performance on the Lost&Found. Without outlier supervision, NLS achieves second best AP in Lost&Found and best in all other metrics.

### Table 6. Ablation Study on Fine-tuning Different Modules.
We show the effect of fine-tuning different components of the model on the Road Anomaly and Fishyscapes LaF validation sets. Fine-tuning MLP+Linear maintains the best performance in unknown detection without sacrificing the closed-set performance.

---

the segmentation, the model’s confusion is clear from a lack of smoothness over the object region. Our method produces competitive results with outlier supervision compared to state-of-the-art PEBAL [65] and DenseHybrid [25] which also use outlier supervision.

**Far away Animals:** As shown in Fig. 13, animals that are situated relatively far from the camera are confused as the inlier pedestrian class. This can be attributed to the dominance of pedestrian class in the training data as well as the similarity of legged animals to a pedestrian in appearance.

**Toy Cars:** Fig. 14 shows that the model fails to detect a toy car on the road and predicts it confidently as the inlier car class. While the class assignment can be considered semantically correct, it is still a hazard in a real driving scenario. Note that a small car can either be a toy car or a real car that is far away. Therefore, distinguishing real cars from toy cars might require additional information such as depth or scale.
Figure 9. **Qualitative Results on SMIYC Anomaly Track.** We compare NLS with and without outlier (OoD) supervision to the state-of-the-art methods PEBAL [65] and DenseHybrid [25] on the anomaly track of the SMIYC benchmark. Even without outlier supervision, NLS with mask classification produces more coherent anomaly maps compared to other methods. PEBAL achieves better coverage of anomalies compared to DenseHybrid but at the cost of a higher false positive rate.
Figure 10. **Qualitative Results on SMIYC Obstacle Track.** We compare NLS with and without outlier supervision to the state-of-the-art methods PEBAL [65] and DenseHybrid [25]. Under adverse weather and difficult lighting conditions, NLS can consistently detect anomalies compared to DenseHybrid and reduce false positives compared to PEBAL.
Figure 11. **Qualitative Comparison of NLS and Max Logit.** We provide more qualitative samples from Fishyscapes LaF [5] to highlight the advantages of NLS over Max Logit [33] while using a mask classifier for both. Note that NLS can better suppress the anomaly scores on the boundaries separating inlier classes as well as in ambiguous background regions at a higher distance from the camera.
Figure 12. **Failure Cases: Tractors and Boats.** Due to their high similarity to inlier car and truck classes, unknown objects like tractors or boats can be predicted as inliers. Outlier supervision can moderately help mitigate this limitation.

Figure 13. **Failure Cases: Animals Confused As Pedestrians.** As the pedestrian is one of the most frequent classes on Cityscapes, the model sometimes predicts animals that appear at a distance as pedestrians (highlighted in circles) on images from SMIYC Anomaly Track.

Figure 14. **Failure Cases: Toy Cars Predicted As Real Cars.** One confusing anomaly case for our model is small toy cars placed in front of the vehicle. Even though they can be semantically considered as cars, they are considered obstacles in a real driving scenario.