RbA: Segmenting Unknown Regions Rejected by All using Mask Classifiers

by

Nazir Nayal

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Computer Science and Engineering

KOÇ ÜNİVERSİTESİ

September 4, 2023

RbA: Segmenting Unknown Regions Rejected by All using Mask Classifiers

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Nazir Nayal

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assist. Prof. Fatma Güney (Advisor)

_____

Prof. David Picard

_____

Assist. Prof. Carlo Masone

Date: _____

To infinity and beyond...

ABSTRACT

RbA: Segmenting Unknown Regions Rejected by All using Mask Classifiers
Nazir Nayal
Master of Science in Computer Science and Engineering
September 4, 2023

Fine-grained visual semantic understanding is essential for autonomous driving and many other computer vision tasks. To this end, segmentation tasks have been witnessing rapid advancements as a result of the growing number of benchmarks proposed. However, existing segmentation benchmarks generally assume a fixed set of semantic categories. Consequently, the development of segmentation methods has been centered around this assumption, while little attention has been poured into handling novel or out-of-distribution (OoD) samples that can potentially be encountered in real-life scenarios. This poses an issue for an autonomous vehicle, as it is crucial to identify unknown objects so that a safety warning can be issued to avoid disastrous consequences in case of failure. As a result, the task of OoD segmentation has been addressed separately, leading to the emergence of methods that adapt to the existing segmentation methods and disregard the performance of the main segmentation tasks. Additionally, these methods also generally suffer from a lack of smoothness and objectness in their predicted anomaly maps due to the reliance on models that follow the per-pixel classification paradigm. In this thesis, we explore the potential of region-level classification models for unknown segmentation as a unified architecture with an inherent ability to express uncertainty. We show that the object queries in mask classification models tend to behave like one vs. all classifiers. Based on this finding, we propose a novel outlier scoring function called Rejected by All (RbA) by defining the event of being an outlier as being rejected by all known classes. We also propose an objective that optimizes this proposed score for boosting the unknown segmentation performance using pseudo-outlier data without hurting the closed-set performance. RbA performs well under high domain shifts and is capable of separating sources of uncertainty, such as at the boundaries, due to known class ambiguity. We evaluate RbA on several unknown segmentation benchmarks and show that it achieves state-of-the-art performance with significant margins compared to previous pixel-level unknown segmentation methods. We report extensive ablation experiments that validate the effectiveness of RbA.

ÖZETÇE

Yüksek Lisans Tez Başlığı
Nazir Nayal
Bilgisayar Bilimleri ve Mühendisliği, Yüksek Lisans
4 Eylul 2023

İnce taneli görsel anlamsal anlayış, otonom sürüş ve diğer birçok bilgisayarla görme görevi için gereklidir. Bu amaçla, segmentasyon görevlerinde, önerilen kıyaslamaların sayısının artmasının bir sonucu olarak hızlı gelişmelere tanık olunmaktadır. Bununla birlikte, mevcut segmentasyon kriterleri genellikle sabit bir anlamsal kategori kümesini varsaymaktadır. Sonuç olarak, segmentasyon yöntemlerinin geliştirilmesi bu varsayım etrafında yoğunlaşırken, gerçek hayat senaryolarında karşılaşılabilecek yeni veya dağıtım dışı (OoD) örneklerin ele alınmasına çok az önem verilmiştir. Arıza durumunda feci sonuçlardan kaçınmak için bir güvenlik uyarısının verilebilmesi amacıyla bilinmeyen nesnelerin tanımlanması çok önemli olduğundan, bu durum otonom bir araç için sorun teşkil etmektedir. Sonuç olarak OoD segmentasyon görevinin ayrı ayrı ele alınması, mevcut segmentasyon yöntemlerine uyum sağlayan ve ana segmentasyon görevlerinin performansını göz ardı eden yöntemlerin ortaya çıkmasına yol açmıştır. Ek olarak, bu yöntemler piksel başına sınıflandırma paradigmasını takip eden modellere dayanmalarından dolayı tahmin edilen anormallik haritalarında genel olarak düzgünlük ve nesnellik eksikliğinden de muzdariptir. Bu tezde, doğası gereği belirsizliği ifade etme yeteneğine sahip birleşik bir mimari olarak bilinmeyen bölümleme için bölge düzeyinde sınıflandırma modellerinin potansiyelini araştırıyoruz. Maske sınıflandırma modellerindeki nesne sorgularının, tüm sınıflandırıcılara göre tek bir sınıflandırıcı gibi davranma eğiliminde olduğunu gösterdik. Bu bulguya dayanarak, aykırı değer olma olayını bilinen tüm sınıflar tarafından reddedilmek olarak tanımlayarak hepsi Tarafından Reddedilen (RbA) adı verilen yeni bir aykırı değer puanlama fonksiyonu öneriyoruz. Ayrıca, kapalı küme performansına zarar vermeden sözde aykırı veriler kullanarak bilinmeyen bölümleme performansını artırmak için önerilen bu puanı optimize eden bir hedef de öneriyoruz. RbA, yüksek etki alanı değişimleri altında iyi performans gösterir ve bilinen sınıf belirsizliği nedeniyle sınırlar gibi belirsizlik kaynaklarını ayırma yeteneğine sahiptir. RbA'yı çeşitli bilinmeyen segmentasyon kriterlerine göre değerlendiriyoruz ve önceki piksel düzeyinde bilinmeyen

segmentasyon yöntemleriyle karşılaştırıldığında önemli marjlarla en son teknolojiye sahip performansa ulaştığını gösteriyoruz. RbA'nın etkinliğini doğrulayan kapsamlı ablasyon deneylerini rapor ediyoruz.

ACKNOWLEDGMENTS

Well well well... A long chain of events caused by my interactions with various people resulted in this thesis in its final form and subtle details. First of all, I extend my limitless gratitude to my parents: Baba Amthal and Mama Reemo, and my siblings: Nasim, Lais, and Lolo (known only by me as Dabdobeh). Since the moment I was born (or the moment they were born), and especially through the difficult times we lived together during war and in the long periods I stayed away from them, their unbounded love, patience, and strength equipped me with the necessary motivation and mental strength leading to this work. I also thank my aunts Soso and Jojo, and Uncle Bassam, for their vital impact and support at all times. Then, my finest appreciation goes to my advisors, Fatma Güney, for her unwavering and continuous support and amazing mentorship, and João F. Henriques, for his incredible guidance and feedback. I would also like to thank Prof. David Picard and Assist. Prof. Carlo Masone for agreeing to join my thesis defense jury and reviewing my work.

A special hat tip to my friend and project accomplice, Mısra ($\sim \mid \sim$). Our discussions and her support (academic and non-academic) are the cornerstone of this work and its stability. Collaborating with her has been an absolute delight.

Furthermore, infinite gratitude for all my friends who made my MSc experience a remarkably pleasant one:

Volkan (hojjam): for the encouragement and support that helped me run my first 5k, and then 10k, for being a dedicated fan of my meme-craft in which you have been featured multiple times, and for the delightful company over the last year, especially at the time of writing this thesis.

Rabia (Snape-ia): for the entertaining bets that I won, the meme and laugh inspawrations, the congenial company, and a lot of things to think about.

Shadi: for the long-term comradeship, and the wise and uplifting sayings (everything will fade, let's just go eat Shawarma).

Hamza and Esra: for being my second family away from home and for your immense support and warmth in my times of distress.
Merve: for the motivating advice, continuous support despite the distance, and free-of-charge therapy sessions :P
Moayad: for tolerating my bullying and superiority :3, and for the long-term pleasant companionship.
Mert: for the very pleasant company in the lab and shared coffee enthusiasm.
Ege: for barely ever refusing to join us for coffee and invariably positive attitude.
Taher: for ordering water and keeping us hydrated.
Further thanks to Ali, Sadra, Ainaz, and my friends at the KUIS AI center for the beautiful memories and interactions.
Again, I would like to additionally acknowledge an anonymous person who had quite the impact on me (and symbolically this work) without knowing the extent of the effect.
Also, warm appreciation for the creator of coffee, Linkin Park for their healing music, Hiroyuki Sawano for Attack on Titan's exquisite soundtracks, Chainsaw Man opening KICK BACK by Kenshi Yonezu, which was quite motivating near deadlines, and Abdullah Al-Ghurair Foundation for Education for changing my life through their STEM undergrad scholarship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

OoD	Out-of-Distribution
ID	In-Distribution
RbA	Rejected by All

Chapter 1

INTRODUCTION

1.1 Overview

Robust and reliable perception is the first step to realizing autonomous intelligence. Perception in this context encompasses the ability to determine the existence of objects or entities within a specific range around an autonomous agent, as well as to semantically classify them to contemplate their effect on the decision-making process. For autonomous vehicles, even though it seems that the environment surrounding a vehicle is restricted compared to a drone navigating a wild forest, for instance, it remains significantly challenging due to the unbounded patterns that can manifest on the road and the limitations of the available solutions.

Numerous methods have been developed by the deep learning research community to solve perception under different settings, such as object detection, semantic segmentation, trajectory prediction, etc. These methods are typically evaluated and tested on datasets curated with the closed-set assumption, i.e., the fixed number of semantic concepts that the dataset provides. The field has witnessed novel methods, bringing the performance closer to perfect on these datasets. However, there exists a gap that limits the applicability of these methods to real-life applications, namely the rare and unknown concepts outside the training datasets. Tailoring the design of architectures and training paradigms towards closed-set datasets rendered the model unequipped to handle novel or outlier inputs sufficiently. These models fail in the open-set case by over-confidently assigning the labels of known classes to unknowns [Hein et al., 2019, Nguyen et al., 2015]. Therefore, the research areas of out-of-distribution detection, open-world recognition, and anomaly detection have been simultaneously witnessing growth in the number of contributions to minimize the gap between performance on benchmarks and real life.

The unexpected and rare cases encountered by the perception system of an autonomous vehicle can be categorized into several levels as proposed in [Breitenstein et al., 2020]. This work focuses on the object-level corner case of perception in driving scenes. Specifically, we address the problem of segmentation of unknown or novel categories, i.e. the semantic concepts that have not been encountered and supervised during training. Detecting novel objects in the surroundings of a vehicle is crucial for the safety and planning of the vehicle. The distribution of potential novel semantic concepts that can be encountered has a long tail of unknowns, such as wild animals, vehicle debris, litter, etc., manifesting in small quantities on the existing driving datasets [Yu et al., 2020, Caesar et al., 2020, Cordts et al., 2016]. Apart from the rareness in terms of quantity, the diversity of unknowns in terms of attributes such as appearance, size, and location adds to the difficulty of the problem.

1.2 Motivation

The existing approaches to segmenting unknown categories can be divided into two depending on whether they use supervision for unknown objects: *with outlier supervision* and *without outlier supervision*. In both settings, the model has access to a fixed set of known classes during training, i.e. inlier or in-distribution. The goal of the task is to identify the pixels belonging to an unknown class, i.e. anomalous, outlier, or out-of-distribution (OoD), while maintaining the performance on the known classes. These two conditions are necessary for the scalability and safety of an autonomous system.

For the methods without outlier supervision, or unsupervised for short, efforts have been focused on constructing a scoring function that quantifies the uncertainty of a sample belonging to the training distribution. Earlier approaches resort to quantifying the disagreement of an ensemble of models [Lakshminarayanan et al., 2017] as a measurement of uncertainty, which is computationally costly for training and inference. Others proposed Monte Carlo dropout [Gal and Ghahramani, 2016], which is less costly but still requires multiple forward passes, therefore costly for practical inference. More recent approaches attempt to extract uncertainty from the predictive class probability distribution, as in the case of the maximum class probability [Hendrycks and Gimpel, 2017] or maximum logit [Hendrycks et al., 2022]. However, these approaches require the probability predic-

tions to be calibrated, which is not guaranteed in practice [Szegedy et al., 2013, Nguyen et al., 2015, Guo et al., 2017, Minderer et al., 2021, Jiang et al., 2018]. Calibration in this context means that the predicted class probabilities represent the true probabilities or likelihoods of the predicted outcomes. In other words, a calibrated classifier is capable of predicting class probabilities that reliably reflect a meaningful confidence score. Lack of calibrations is common among deep learning classifiers and manifests in the form of consistently assigning too high or too low class probabilities compared to the actual outcome. This lack of calibration leads to unreliable confidence and uncertainty estimates.

In the case of outlier supervision, additional data that is representative of potential outliers is used for additional supervision. The model can utilize outlier data to learn a more discriminative representation against outliers. Typically, data samples from another dataset with a sizeable semantic taxonomy from a disjoint domain are used for this purpose [Chan et al., 2021b], or in other cases; outlier objects are artificially pasted to driving images [Grcić et al., 2022, Tian et al., 2022] in order to maintain the contextual information of driving scenes. Despite the better performance compared to unsupervised methods [Tian et al., 2022], depending on the choice and quantity of outlier data, outlier supervision can bias the model to a particular distribution of outliers and negatively affect the in-distribution performance.

Most existing methods for unknown segmentation suffer from several drawbacks:

Lack of Objectness: The first issue stems from defining the semantic segmentation problem as a per-pixel classification problem. Per-pixel classification has been the main paradigm used to develop semantic segmentation models. In this paradigm, each pixel is classified into a semantic class independently. The loss calculation aggregates the error from all pixels. The convenience of this paradigm arises as a generalization from standard image classification task, which has been studied thoroughly over many years. However, semantic supervision and decision-making at the pixel level limit an image's smoothness and access to global information. Most existing unknown segmentation methods build their solutions on top of per-pixel classification models, leading to a lack of smoothness and a failure to preserve the boundaries of objects in the predictions of unknowns, as shown in Fig. 1.1.

Uncertainty at Boundaries: The existing methods also suffer from high false positive rates due to failing to separate the sources of uncertainty, especially on datasets in the wild such as Road Anomaly [Lis et al., 2019]. False positives in the context of unknown object segmentation means assigning high anomaly or outlier scores to known regions in the input image. For example, on the boundary regions that separate semantic categories, segmentation models typically predict weak scores for the two inlier classes separated by the boundary, causing these regions to be confused as outliers by unsupervised score-based methods that estimate the uncertainty of a pixel using the class confidence probabilities [Hendrycks et al., 2022, Hendrycks and Gimpel, 2017, Tian et al., 2022].

Negative impact on closed-set performance: It has been shown that using outlier supervision in training outperforms unsupervised methods [Tian et al., 2022, Grcić et al., 2022] in terms of unknown segmentation metrics. However, this is achieved at the cost of lower closed-set performance. This is disadvantageous as the primary goal of outlier detection is to reinforce the closed-set task, not inhibit it.

1.3 Contributions

This thesis aims to tackle the aforementioned limitations in existing methods for unknown segmentation. It challenges the per-pixel segmentation paradigm by exploring the potential of region-level classification for unknown segmentation. Region-level or mask classifiers such as MaskFormer [Cheng et al., 2021], Mask2Former [Cheng et al., 2022], or OneFormer [Jain et al., 2023] have been proposed recently with the goal of unifying the three segmentation tasks: semantic, instance, and panoptic segmentation. The idea is to use learnable object queries that interact with the input image features using several layers of cross-attention and then allow each object query to predict a region in the image and assign it to a semantic category. Unlike per-pixel classification models, the semantic-cross entropy loss during training is computed per region instead of pixel, and an additional mask localization loss is used for mask supervision.

We discover the properties of this family of models, which allow better calibration of confidence values. Then, we exploit these properties to boost the performance of the existing OoD methods that rely on predicted class scores such as max logit [Hendrycks

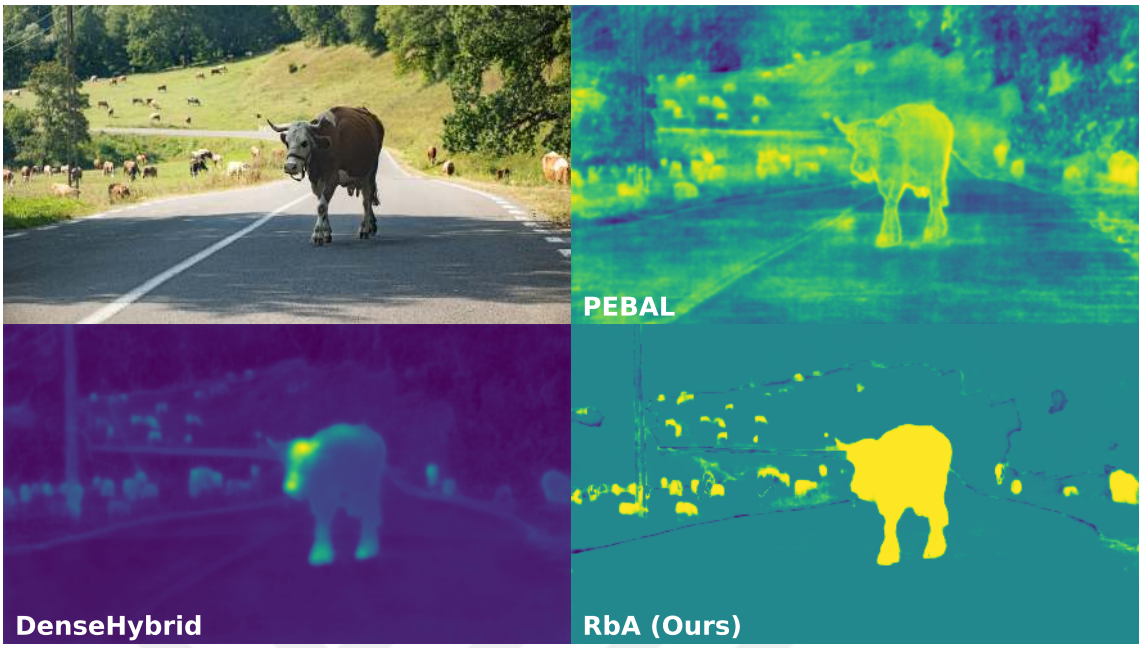


Figure 1.1: **Preserving objectness and eliminating noise.** While state-of-the-art methods PEBAL [Tian et al., 2022] and DenseHybrid [Grcić et al., 2022] suffer from a lack of smoothness and objectness with high false positive rates, our proposed method RbA clearly segments the unknown objects. It reduces false positives by eliminating uncertainty at semantic boundaries and in ambiguous background regions. compared to recent state-of-the-art methods PEBAL [Tian et al., 2022] and DenseHybrid [Grcić et al., 2022]. It also performs well for different object scales and under high distribution shifts.

et al., 2022] and energy-based ones [Grcić et al., 2022, Tian et al., 2022, Liu et al., 2020]. More specifically, Based on exploring the behavior of object queries in mask classification, we find that most object queries tend to behave like one vs. all classifiers. Consequently, we propose a novel outlier scoring function termed Rejected by All (RbA) based on this one vs. all behavior of object queries. We define the event of a pixel being an outlier as a result of being rejected by all known classes. In other words, we define being an outlier as a complementary event whose probability can be expressed in terms of the known class probabilities. We show that this scoring function can eliminate irrelevant sources of uncertainty, as in the case of boundaries, resulting in a considerably lower false positive rate on all datasets. We show the effectiveness of RbA on both semantic and panoptic

segmentation mask classifiers.

Additionally, we propose an objective to optimize the proposed outlier scoring function using a limited amount of outlier data. By fine-tuning a very small portion of the model with this objective, our method outperforms the state-of-the-art on challenging datasets with high distribution shifts such as Road Anomaly [Lis et al., 2019] and SMIYC [Chan et al., 2021a]. Notably, we achieve this without affecting the closed-set performance. The proposed RbA can be adapted for both semantic and panoptic segmentation settings.

Our contributions can be summarized as follows:

- We postulate and study the inherent ability of mask classification models to express uncertainty and use this strength to boost the performance of several existing unknown semantic and panoptic segmentation methods.
- Based on our finding that object queries behave approximately as one vs. all classifiers, we propose a novel outlier scoring function that represents the probability of being an outlier as not being any of the known classes. The proposed scoring function helps to eliminate uncertainty in ambiguous inlier regions such as semantic boundaries.
- We propose a loss function that optimizes our scoring function using minimal outlier data. The proposed objective exceeds the state-of-the-art by only fine-tuning a very small portion of the model without affecting the closed-set performance.

Chapter 2

RELATED WORK AND BACKGROUND

2.1 Semantic Segmentation Paradigms

Since the success of Fully Convolutional Networks (FCN) [Shelhamer et al., 2015], 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 [Chen et al., 2018a, Chen et al., 2018b, Dai et al., 2017, Zhao et al., 2017, Yang et al., 2018], and by aggregating multi-scale contextual information [Yu and Koltun, 2016, Yuan et al., 2020]. Recent work shifted towards transformer-based architectures [Xie et al., 2021, Strudel et al., 2021, Yuan et al., 2021, Zheng et al., 2021, Yang et al., 2021] and attention mechanisms [Harley et al., 2017, Li et al., 2018, Zhao et al., 2018, He et al., 2019, Li et al., 2019, Huang et al., 2019, Fu et al., 2019].

On the other hand, mask classification has been mainly adopted by instance segmentation and object detection models [He et al., 2017, Hariharan et al., 2014, Carion et al., 2020] 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 [Wang et al., 2021] 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 [Carreira et al., 2012, Hariharan et al., 2014], they were quickly outperformed by the per-pixel methods until recently. MaskFormer variants [Cheng et al., 2021, Cheng et al., 2022] apply query-based mask classification and attention to obtain a unified segmentation model that shows competitive performance with specialized semantic and instance segmentation architectures across benchmarks [Lin et al., 2014, Cordts et al., 2016, Zhou et al., 2017, Neuhold et al., 2017].

2.2 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 [Hendrycks and Gimpel, 2017, Liang et al., 2018], ensembles [Lakshminarayanan et al., 2017], Bayesian approximation [Mukhoti and Gal, 2018], Monte Carlo dropout [Gal and Ghahramani, 2016], or by learning to estimate its confidence [Devries and Taylor, 2018]. However, posterior probabilities of a closed-set model are not necessarily calibrated, leading to overconfident predictions on unseen categories [Szegedy et al., 2013, Nguyen et al., 2015, Guo et al., 2017, Minderer et al., 2021, Jiang et al., 2018]. Therefore, follow-up work focuses on making a clear distinction between inliers and outliers by using true class probabilities [Corbière et al., 2019], unnormalized logits instead of softmax probabilities [Hendrycks et al., 2022], standardized class-wise logits [Jung et al., 2021], and the distance to learned prototypes of known classes [Cen et al., 2021]. 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 [Lee et al., 2018, Ren et al., 2019] while subsequent works focus on reconstruction. Several works rely on the predicted segmentation maps to resynthesize [Lis et al., 2019, Xia et al., 2020, Haldimann et al., 2019, Vojir et al., 2021] or inpaint the inputs [Lis et al., 2020] and measure discrepancy with comparison networks. Others apply localized adversarial attacks [Besnier et al., 2021], synthesize negatives using normalizing flows [Grcic et al., 2021], or combine Gaussian mixture models with discriminative representation learning [Liang et al., 2022]. 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 [Grcic et al., 2021]. Several works [Nalisnick et al., 2018, Serrà et al., 2020, Zhang et al., 2021] show that generative models tend to

estimate high likelihoods on out-of-distribution samples.

2.3 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 [Russakovsky et al., 2015] and 80 Million Tiny Images [Torralba et al., 2008] for OoD. Given data, OoD detection can be simply treated as binary classification [Bevandic et al., 2018, Bevandic et al., 2019]. Outlier data can also be used to estimate the distributional uncertainty of OoD samples [Malinin and Gales, 2018] or to fine-tune parametrized OoD detectors [Hendrycks et al., 2019]. The energy score has been proposed as a better alternative to softmax in terms of separation [Liu et al., 2020]. SynBoost [Biase et al., 2021] is a supervised image resynthesis method that treats void regions as anomalies to obtain an uncertainty signal.

Recent work uses a subset of COCO [Lin et al., 2014] or ADE20K [Zhou et al., 2017], either as entire images [Chan et al., 2021b] or after cut-and-paste into the inlier scenes [Tian et al., 2022, Grcić et al., 2022]. Meta-OoD [Chan et al., 2021b] maximizes the entropy on outliers, whereas PEBAL [Tian et al., 2022] learns adaptive energy-based penalties by abstention learning. Combining likelihood and posterior evaluation, DenseHybrid [Grcić et al., 2022] achieves state-of-the-art results. However, for each benchmark, different models are fine-tuned using multiple datasets [Zhou et al., 2017, Neuhold et al., 2017, Zendel et al., 2018] with high distribution shifts, resulting in a higher degree of supervision and variety. Our model, on the other hand, can achieve better performance across benchmarks by using the same model and only a small subset of COCO [Lin et al., 2014] for fine-tuning.

2.4 One vs. all Classification

There exist some methods that attempt to utilize one vs. all classification for unknown segmentation [Franchi et al., 2022, Bevandić et al., 2021], where a separate classifier is trained for each known category in addition to a standard multi-class classifier. The class probabilities obtained by the one vs. all classifiers are merely used for calibrating the multi-class classifier. The outlier scoring function used is the negative maximum class

probability of the calibrated probabilities. On the contrary, our method utilizes the implicit one vs. all behavior of mask classifiers for explicitly defining the outlier probability.

2.5 Concurrent Work Exploring Mask-Classification

In parallel with our work, other methods emerged that explore and utilize the properties of mask classification for unknown segmentation. EAM [Grcic et al., 2023] applies maximum class probability to estimate per-mask uncertainty and then aggregates the score over all the predicted masks to obtain the anomaly score. They also utilize pseudo outliers by pasting objects onto driving scenes and assigning them the void class label in order to encourage masks to avoid predictions in anomalous regions. Mask2Anomaly [Rai et al., 2023], on the other hand, modifies the attention modules in the transformer modules to allow for focusing on background regions in addition to foreground. They also utilize additional outlier data to contrastively improve the separability between knowns and unknowns. These contributions improve the outlier score obtained by maximum softmax probability (MSP) [Hendrycks and Gimpel, 2017] score. Similar to concurrent methods, we also utilize unknown data to improve our proposed scoring function. Differently, our proposed outlier scoring function is defined based on the implicit one vs. all behavior of object queries, and our objective focuses solely on suppressing the confidence of unknown pixels.

Chapter 3

METHODOLOGY

3.1 Overview

In this chapter, we provide a detailed description of our formulation, as well as the observations and insights that led to it. We first show an analysis of the mask classification models and their behavior. Then, based on our analysis, we propose a novel scoring function to exploit the implicit one vs. all behavior in these models. We mathematically define the probability of being an outlier probability as the "none of the above" option for the model. Then, we outline our proposed training objective to optimize our proposed scoring function RbA with minimal outlier data. Moreover, we describe the application of RbA for semantic and panoptic segmentation. Finally, we present an analysis of the behavior of the predicted logits and how it aligns with the properties of RbA.

3.2 Mask Classification with Mask2Former Architecture

Mask2Former consists of three main parts: the backbone, the pixel decoder, and the transformer decoder. The backbone processes the input image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ associated with ground truth $\mathbf{y} \in \mathcal{K}^{H\times W}$ for a predefined set of known classes $\mathcal{K} = \{1, \dots, K\}$. The backbone extracts features at multiple scales and different hidden dimension sizes. Then, the pixel decoder further processes the multi-scale features to produce high-resolution per-pixel features $\mathbf{F}(\mathbf{x}) \in \mathbb{R}^{C_p\times H\times W}$. The transformer decoder takes the resulting multi-scale features $\{\mathbf{f}_i\}_{i=1}^{D}$, where D is number of scales, 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 using self-attention modules and with the multi-scale features $\mathbf{f}_i$ using masked cross-attention layers. In each layer of the transformer decoder, the queries interact with features of only a single scale. The features are exposed to queries in a round-robin order.

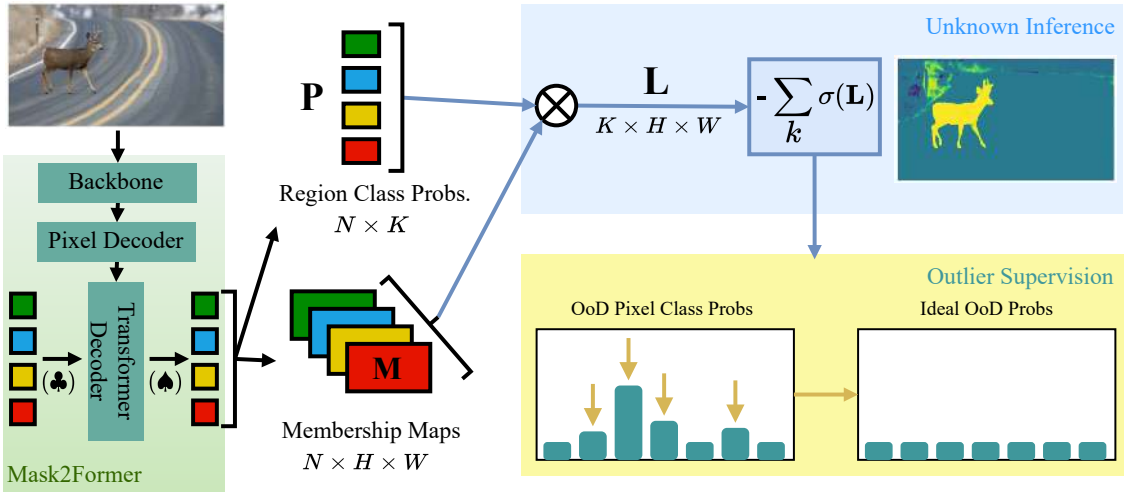


Figure 3.1: **Overview.** This figure provides an illustration of our proposed outlier scoring function RbA and the objective to optimize it as defined in (3.6). The class logit scores $\mathbf{L}$ are aggregated as the product of region class probabilities $\mathbf{P}$, and mask predictions $\mathbf{M}$ are pooled over all regions. We define the RbA as the probability of not being assigned to any of the known classes. With the proposed objective, we push the probabilities of known classes down in the outlier pixels.

After refinement, the object queries are used for region prediction and classification. Regions are predicted by also using pixel features $\mathbf{F}$. 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 σ to the result:

$$\mathbf{M}(\mathbf{x}) = \sigma(\mathbf{Q}_p\, \mathbf{F}(\mathbf{x})) \tag{3.1}$$

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

region prediction and classification losses as the cost. For region prediction, a weighted combination of dice loss [Milletari et al., 2016] and binary cross-entropy is applied to the binary mask predictions. For classification, cross-entropy loss is used.

In inference, the class scores or logits $\mathbf{L}(\mathbf{x}) \in \mathbb{R}^{K \times H \times W}$ are calculated as the product of mask predictions with class predictions by broadcasting the class prediction to all the pixels within the region:

$$\mathbf{L}(\mathbf{x}) = \sum_{n=1}^{N} \mathbf{P}_n(\mathbf{x}) \mathbf{M}_n(\mathbf{x}) \tag{3.2}$$

We perform our analysis and build our method on top of the Mask2Former architecture [Cheng et al., 2022], which is an improved version of the initial MaskFormer [Cheng et al., 2021]. The improvements include using masked attention modules in place of standard cross-attention, where the attention computation is limited to the mask region predicted by the previous decoder layer. Additionally, multi-scale features are exposed for interaction with the learnable queries instead of only a single scale. There were also other minor improvements, such as switching the order between cross and self-attention and removing dropout for computational efficiency.

3.3 Independence of Object Queries

The logit term $\mathbf{L}$ as defined in Eq. 3.2 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. During training, the ground truth binary map of each class is matched to an object query using bipartite matching. Therefore, we find that object queries specialize in predicting a specific class after convergence. We empirically verify this behavior on another driving dataset on the validation set of BDD100K [Yu et al., 2020] using a model that is trained solely on Cityscapes. We 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%; Fig. 3.2 shows a visualization of this specialized behavior, in which we can see that for each of the closed-set classes, there is a single object query dominantly predicting it.

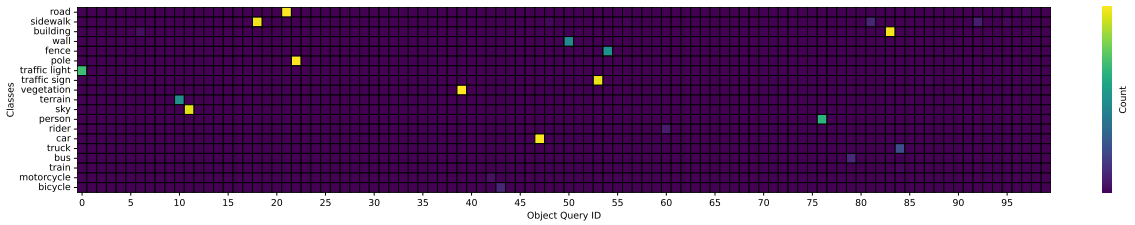


Figure 3.2: **Specialization of Object Queries.** Certain object queries specialize in predicting a specific class. For each query, we show how many times it predicts a region to belong to a class with high confidence. The sparsity in the plot clearly shows the specialization of queries.

After identifying which object query predicts which class, we test their independence, i.e. the ability of each object query to predict its class without relying on other object queries during their interactions in the transformer decoder. To evaluate the predictions of class k, we mask out all but its specialized query. We do this in one of two ways: 1) before the transformer decoder (♣ in Fig. 3.1), where each object query only interacts with the image features and not with each other (*hard masking*), or 2) after the transformer decoder (♠ in Fig. 3.1), allowing queries to interact with each other weakly (*soft masking*), but only considering the mask prediction of the specialized query is the class logits mask. In both cases, only the specialized query is used to predict the mask and class. Fig. 3.3 shows a clear illustration of how hard and soft masking are applied.

Fig. 3.4 shows per-class IoU scores on Cityscapes [Cordts et al., 2016] using both strategies compared to the original model without any masking. We observe that the performance in most of the classes is not affected compared to the original model. The drop in performance occurs only in rare classes, such as train, truck, or bus, indicating that their object queries rely on other queries in prediction, which explains the slightly better performance of soft masking than hard masking. This behavior of object queries resembles multiple independent binary classifiers implicitly embedded in a single model.

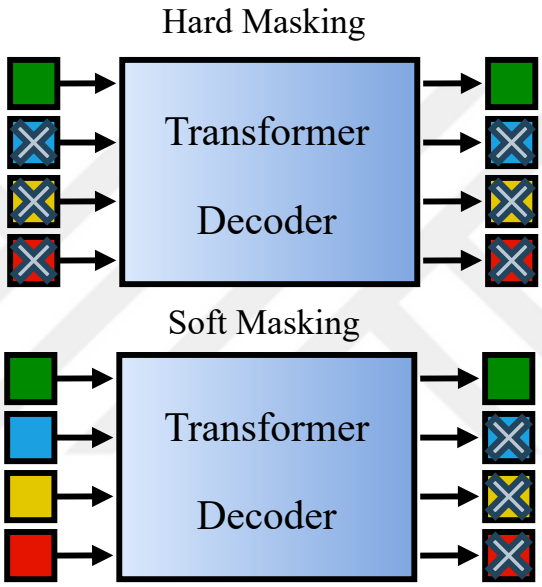


Figure 3.3: **Illustration of hard vs. soft masking of object queries.** In hard masking, when predicting class k, all but the object queries specialized to predict class k are masked before the transformer decoder so that the specialized query can only interact with the image features. In soft masking, the queries are allowed to interact within the transformer decoder but are dropped after, and only the specialized query is used to predict class k.

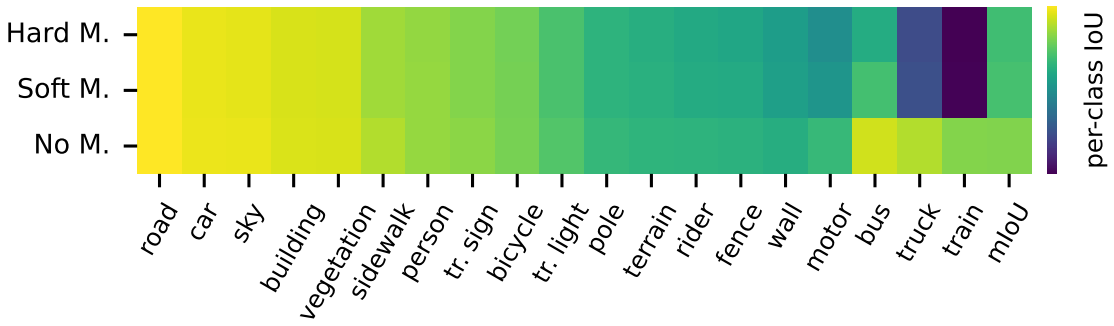


Figure 3.4: **Masking object queries.** We show the impact on per-class IoU on Cityscapes [Cordts et al., 2016] when using two types of masking: hard masking without any interactions between the queries (top) and soft masking by allowing interactions in the transformer decoder (middle) compared to the original model without masking (bottom). The constant color in most columns shows that most of the object queries can independently segment their particular classes. Without interacting with one another, specialized object queries can independently segment their particular classes.

3.4 Rejected by All (RbA) Scoring Function

Inspired by the independent behavior of object queries, we propose to model the prediction of each class as an independent binary classification problem. We consider it as K one vs. all classifiers where the predicted score for each class is independently modeled as follows:

$$p(\mathbf{y} = k|\mathbf{x}) = \sigma(\mathbf{L}_k(\mathbf{x})) \tag{3.3}$$

where $\mathbf{y} \in \mathcal{K}^{H \times W}$ is a random variable representing the predicted class label over a predefined set of known classes $\mathcal{K} = \{1, \ldots, K\}$ and σ is a normalization function applied to per class logits to map them to a probability, i.e. a value between 0 and 1. Based on this definition, we assume that the latent space is partitioned into $K + 1$ mutually exclusive and exhaustive regions, such that the label $K + 1$ represents the region where the outliers reside, rejected by all other known classes. By assuming the mutual exclusiveness of per-class probabilities for a given input, we can define the probability map of an input $\mathbf{x}$ being an outlier as follows:

$$
\begin{aligned}
p(\mathbf{y} = K+1|\mathbf{x}) &= 1 - p\left(\cup_{k=1}^{K} \mathbf{y} = k|\mathbf{x}\right) \\
&= 1 - \sum_{k=1}^{K} p(\mathbf{y} = k|\mathbf{x}) \\
&= 1 - \sum_{k=1}^{K} \sigma(\mathbf{L}_k(\mathbf{x}))
\end{aligned}
\tag{3.4}
$$

Dropping the constant 1 (which does not affect the optimization), we define our outlier scoring function RbA:

$$
\mathrm{RbA}(\mathbf{x}) = -\sum_{k=1}^{K} \sigma(\mathbf{L}_k(\mathbf{x}))
\tag{3.5}
$$

We choose σ to be the `tanh` function to map $\mathbf{L}_k > 0$ more uniformly to the range $[0, 1]$.

3.5 Optimizing RbA with Outlier Supervision

We propose to regularize our proposed scoring function RbA using supervision from pseudo outliers created by cutting and pasting objects that are semantically disjoint from the known categories onto driving scenes [Li et al., 2021, Xia et al., 2020]. Our goal is to improve the unknown segmentation while preserving the closed-set performance. Without retraining the entire model, we only fine-tune the mask prediction MLP and classification layer after the transformer decoder, which constitutes only 0.21% of the total model parameters. Interestingly, finetuning this significantly small portion yields considerable gains. We also study the effect of fine-tuning different components of the mask classification architecture, such as the pixel decoder, transformer decoder, and full model; we show the results of these experiments in Section 5.8. For outlier data, we use a modified version of Anomaly Mix proposed in [Tian et al., 2022], where objects from COCO dataset [Lin et al., 2014] are randomly cut and pasted on Cityscapes images [Cordts et al., 2016]. We regularize the scores by *maximizing RbA for outlier pixels*, with a squared hinge loss. This is also equivalent to suppressing high-confidence probabilities of known classes for outlier pixels as shown in Fig. 3.1. The loss is formally defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{RbA}} &= \sum_{\mathbf{x}\in\Omega_{out}} \left(\max(0, \alpha - \text{RbA}(\mathbf{x}))\right)^2 \\ &= \sum_{\mathbf{x}\in\Omega_{out}} \max\left(0, \alpha + \sum_{k=1}^{K} \sigma(\mathbf{L}_k(\mathbf{x}))\right)^2 \end{aligned} \tag{3.6}$$

where Ω_{out} is the set of outlier pixels. We experimentally set the hyper-parameter α to 5, but we found that any $\alpha > 0$ works well in practice. Check the ablations section for details (Section 5.7). There also exist other options to be used for computing the loss and regularizing RbA. These options include mean-squared error (L2), mean absolute error (L1), binary cross entropy (BCE), and KL divergence. We examine the effect of each loss in Section 5.2.

3.6 Applying RbA to Segmentation Tasks

We focus on applying RbA to semantic and panoptic segmentation. For semantic segmentation, the application is straightforward; all the unknown pixels, regardless of their specific semantic category or their assigned instances, are cast as unknowns. However, in the case of panoptic segmentation, the challenge of separating unknown objects into separate instances emerges. In cases where instances in the image are already separated, this does not cause an issue, as a simple clustering or connected components algorithm is sufficient for separation. Yet, in the cases where the pixels of different instances are adjacent, another mechanism is needed. In this work, we show that for the main existing panoptic benchmark, combining RbA with simple methods such as connected components can yield state-of-the-art, but investigating a more robust method that would work on complicated real-life examples remains for future work. We explain next how we apply RbA to panoptic segmentation.

Since MaskFormer [Cheng et al., 2021] and Mask2Former [Cheng et al., 2022] were developed with the goal of unifying the segmentation tasks architecture-wise, the flow of the computation remains the same as shown in Fig. 3.1. What distinguishes a semantic segmentation model from a panoptic one is the format of the label supervision. Therefore, the RbA score map can be calculated the same way as described in (3.5) for a model trained

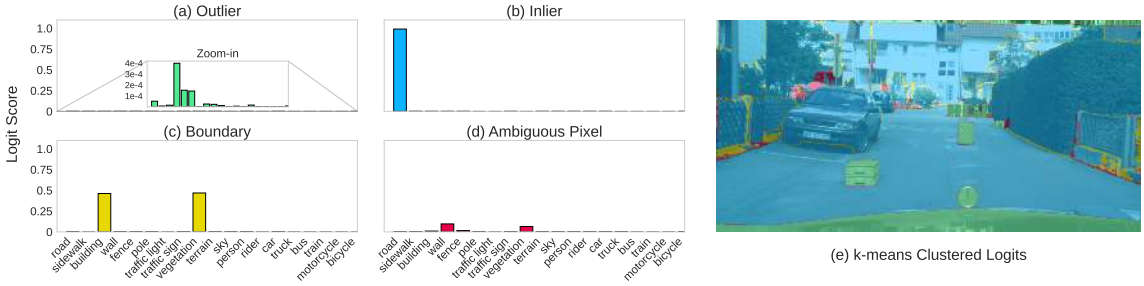


Figure 3.5: **Categorizing the behavior of logits.** Outlier pixels receive extremely low votes from object queries (a). Inlier pixels receive a high vote from a single object query (b). Boundary pixels separating two inlier classes receive moderate votes from both object queries (c). Ambiguous regions receive weak votes from multiple object queries (d). Clustering clearly outlines these four behaviors of logits (e). Pixels in (e) are color-coded with the same colors of the respective histograms (a-d).

on panoptic labels. After obtaining the RbA score map, we apply dilation and erosion operations of kernel size 3×3 in order to reduce the noise from the map. Finally, we extract the connected components from the RbA score map and consider each component as an unknown instance.

3.7 Analyzing RbA

The term $\mathbf{L}$ in Eq. 3.2 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. 3.5e. For an inlier pixel, only a single object query votes for it with high confidence (Fig. 3.5b), whereas true outlier pixels do not receive any votes from any object query (Fig. 3.5a). These two modes, especially the outliers in Fig. 3.5a, due to the one vs. all behavior, reduce the overconfidence issue in the existing outlier scoring functions used in max logit [Hendrycks et al., 2022] and energy-based methods [Grcić et al., 2022, Tian et al., 2022, Liu et al., 2020], therefore improve their results (Table 5.1).

However, there are pixels that disrupt the separability between the inliers and the

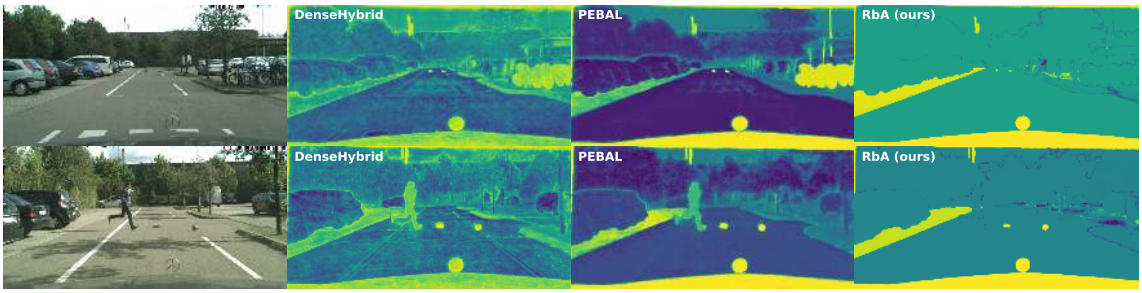


Figure 3.6: **Visual comparison to the state-of-the-art**. We show visualizations of outlier score maps predicted by our method, RbA, compared to the ones predicted by state-of-the-art methods PEBAL [Tian et al., 2022] and DenseHybrid [Grcić et al., 2022] trained using the same architecture as the RbA for a fair comparison. The other two methods falsely identify the inlier classes, such as person and bike, which are correctly ignored by the proposed RbA. It is noteworthy that RbA also eliminates the false positives in the background region, especially at the boundaries separating inliers and better preserves the smoothness of the outlier map compared to other methods despite being also trained with mask classification.

outliers, which max logit and energy-based methods fail to capture. For example, pixels on a boundary between two inlier classes (Fig. 3.5c) or ambiguous background pixels (Fig. 3.5d) end up with a higher anomaly score than the inliers, causing them to be mistaken as an outlier.

Boundary and ambiguous regions are commonly characterized by having more than one weak vote from object queries. Since RbA aggregates votes from all classes, summing these weak votes results in a lower outlier score and hence reduces the false positive rate.

Fig. 3.6 highlights the differences between the anomaly maps predicted by RbA and the state-of-the-art methods, also trained using Mask2Former. Note that RbA assigns low outlier scores at boundaries separating known classes.

Chapter 4

EXPERIMENTS

We describe in this chapter our experimental setting and discuss our results. First, we provide the details of the datasets used for training and evaluation. Then, we describe our implementation details and the metrics used for evaluation. Finally, we discuss the quantitative results in comparison with previous state the art and show qualitative results to highlight the performance quality of our contributions.

4.1 Datasets

We detail below the datasets used for both semantic and panoptic segmentation separately.

4.1.1 Semantic Segmentation

We train our models mainly on the Cityscapes dataset [Cordts et al., 2016], which consists of 2975 training and 500 validation images. It contains 19 classes which are considered as the base known categories in most of the anomaly segmentation benchmarks. The class names in the dataset can be seen in Fig. 3.4. For outlier evaluation, we consider multiple datasets:

Segment Me If You Can (SMIYC) benchmark: [Chan et al., 2021a] contains two different Tracks: Anomaly Track and Obstacle Track. The anomaly track has 100 hidden test 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: [Lis et al., 2019] is an earlier and smaller version of SMIYC. It consists of 60 images with diverse objects in diverse environments. Another crucial difference compared to SMIYC is that segmentation labels of Road Anomaly do not account for

void regions that are irrelevant to the outlier detection tasks. In other words, predicting an irrelevant region outside the road in Road Anomaly would be punished as a false positive, whereas this issue is handled in SMIYC by including void class regions that do not contribute to the metric computation.

Fishyscapes Lost&Found: [Blum et al., 2021] contains 100 validation and hidden 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 the other datasets. Following most previous methods, we report our results mainly on the validation set due to the slowness of processing the submissions to the official benchmark to obtain the metrics on the hidden test set.

Mapillary Vistas: [Neuhold et al., 2017] contains 18k training images, which we use as auxiliary inlier data in the outlier supervision finetuning to boost the performance. The class labels of Mapillary are grouped so that they match the same taxonomy of Cityscapes. We used this dataset when finetuning with outlier supervision for a fair comparison with concurrent work [Grcic et al., 2023].

4.1.2 Panoptic Segmentation

Following [Hwang et al., 2021], the COCO [Lin et al., 2014] dataset with panoptic annotations is used for evaluating open-set panoptic segmentation. The dataset consists of 118K training and 5K validation samples, containing 80 thing classes and 54 stuff classes. To create the open-set setting, some classes from the known thing classes are removed from the training set and used for evaluating the unknown metrics. The authors in [Hwang et al., 2021] propose three different splits depending on the percentage of known classes removed from the thing classes. We evaluate our model on the most difficult split, which removes 20% of the known classes to the unknown set. The unknown classes of the used split are: `car, cow, pizza, toilet, boat, tie, zebra, stop sign, dining table, banana, bicycle, cake, sink, cat, keyboard, bear`.

4.2 Implementation Overview

4.2.1 Semantic Segmentation

We follow the setup of [Cheng et al., 2022] for closed-set training on Cityscapes. We use the Swin-B [Liu et al., 2021] architecture as the backbone. Differently, we use only one decoder layer in the transformer decoder instead of nine (see Section 5.4). For outlier supervision, we fine-tune the mask prediction MLP and the classification layer for 2K iterations with a batch size of 16 using the standard loss functions used in [Cheng et al., 2022] in addition to the RbA loss defined in Eq. 3.6. Previous work [Tian et al., 2022] samples 300 new images every epoch out of 40K COCO images with objects different from Cityscapes inliers. In our case, we sample 300 images only at the beginning and fix them; then, at each iteration, an image is randomly chosen and pasted on inlier images with probability p_{out}. We experimentally set p_{out} to 0.1 (see Section 5.6).

4.2.2 Panoptic Segmentation

Following [Hwang et al., 2021] for a fair comparison, we train a Mask2Former model from scratch with an Image Net 21K pretrained ResNet50 backbone for 280K iterations using a batch size of 18 using 3 RTX A6000 GPUs. We report RbA results without using outlier supervision for panoptic segmentation. Since the semantic taxonomy of COCO is very diverse, selecting pseudo-outlier samples for outlier supervision requires extra careful selection and filtering, therefore, we leave it for potential future work.

4.3 Evaluation Metrics

4.3.1 Semantic Segmentation

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 [Chan et al., 2021a]. Specifically, the proposed metrics aim at quantifying true positives (TP), false negatives (FN), and false positives

Method	OoD	Extra	Anomaly Track					Obstacle Track				
	Data	Net.	AP ↑	FPR ↓	sIoU gt ↑	PPV ↑	mean F1 ↑	AP ↑	FPR ↓	sIoU gt ↑	PPV ↑	mean F1 ↑
Emb. Density [Blum et al., 2021]	✗	✗	37.5	70.8	33.9	20.5	7.9	0.8	46.4	35.6	2.9	2.3
MC Dropout [Mukhoti and Gal, 2018]	✗	✗	28.87	69.47	20.49	17.26	4.26	4.88	50.31	5.49	5.77	1.05
ODIN [Liang et al., 2018]	✗	✗	33.06	71.68	19.53	17.88	5.15	22.12	15.28	21.62	18.50	9.37
Ensemble [Lakshminarayanan et al., 2017]	✗	✗	17.66	91.06	16.44	20.77	3.39	1.06	77.20	8.63	4.71	1.28
MSP [Hendrycks and Gimpel, 2017]	✗	✗	27.97	72.05	15.48	15.29	5.37	15.72	16.60	19.72	15.93	6.25
Mahalanobis [Lee et al., 2018]	✗	✗	20.04	86.99	14.82	10.22	2.68	20.90	13.08	13.52	21.79	4.70
JSRNet [Vojir et al., 2021]	✗	✓	33.6	43.9	20.2	29.3	13.7	28.1	28.9	18.6	24.5	11.0
Road Inpain.[Lis et al., 2020]	✗	✓	-	-	-	-	-	54.1	47.1	**57.6**	39.5	36.0
Image Resyn.[Lis et al., 2019]	✗	✓	52.3	25.9	39.7	11.0	12.5	37.7	4.7	16.6	20.5	8.4
ObsNet[Besnier et al., 2021]	✗	✓	75.4	26.7	44.2	**52.6**	**45.1**	-	-	-	-	-
NFlowJS[Grcic et al., 2021]	✗	✓	56.9	34.7	36.9	18.0	14.9	85.6	**0.4**	45.5	49.5	**50.4**
RbA (Ours)	✗	✗	**86.1**	**15.9**	**56.3**	41.4	42.0	**87.8**	3.3	47.4	**56.2**	**50.4**
SynBoost[Biase et al., 2021]	✓	✓	56.4	61.9	34.7	17.8	10.0	71.3	3.2	44.3	41.8	37.6
Max. Entropy[Li et al., 2019]	✓	✗	85.5	15.0	49.2	39.5	28.7	85.1	0.8	47.9	62.6	48.5
DenseHybrid[Grcić et al., 2022]	✓	✗	78.0	9.8	54.2	24.1	31.1	87.1	0.2	45.7	50.1	50.7
PEBAL[Tian et al., 2022]	✓	✗	49.1	40.8	38.9	27.2	14.5	5.0	12.7	29.9	7.6	5.5
EAM†*[Grcic et al., 2023]	✓	✗	93.8	**4.1**	**67.1**	**53.8**	**60.9**	92.9	0.5	**65.9**	**76.5**	**75.6**
Mask2Anomaly*[Rai et al., 2023]	✓	✗	88.7	14.6	55.3	51.7	47.2	93.22	0.2	55.7	75.4	68.2
RbA* (Ours)	✓	✗	90.9	11.6	55.7	52.1	46.8	91.8	0.5	58.4	58.8	60.9
RbA†* (Ours)	✓	✗	**94.5**	4.6	64.9	47.5	51.9	**95.1**	**0.1**	54.3	59.1	57.4

Table 4.1: **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. † denotes that the model was trained using additional inlier data. $*$ denotes that a mask classification model is used.

(FP) of detected unknown objects. Please see the benchmark paper [Chan et al., 2021a] for more details on these metrics.

4.3.2 Panoptic Segmentation

Following [Hwang et al., 2021], we report the Panoptic Quality (PQ), Segmentation Quality (SQ), and Recognition Quality (RQ) computed on known things and stuff classes and unknown classes separately.

Method	OoD Data	Extra Net.	Road Anomaly AUC ↑	Road Anomaly AP ↑	Road Anomaly FPR ↓	FS LaF AUC ↑	FS LaF AP ↑	FS LaF FPR ↓
MSP (R101) [Hendrycks and Gimpel, 2017]	✗	✗	73.76	20.59	68.44	86.99	6.02	45.63
Entropy (R101) [Hendrycks and Gimpel, 2017]	✗	✗	75.12	22.38	68.15	88.32	13.91	44.85
Mahalanobis [Lee et al., 2018]	✗	✗	76.73	22.85	59.20	92.51	27.83	30.17
SML [Jung et al., 2021]	✗	✗	81.96	25.82	49.74	<u>96.88</u>	36.55	14.53
GMMSeg (SF) [Liang et al., 2022]	✗	✗	<u>89.37</u>	<u>57.65</u>	<u>44.34</u>	**97.83**	<u>50.03</u>	<u>12.55</u>
SynthCP [Xia et al., 2020]	✗	✓	76.08	24.86	64.69	88.34	6.54	45.95
RbA (Ours)	✗	✗	**95.60**	**78.45**	**11.83**	96.43	**60.96**	**10.63**
SynBoost (WRN38) [Biase et al., 2021]	✓	✓	81.91	38.21	64.75	96.21	60.58	31.02
Maximized Entropy [Chan et al., 2021b]	✓	✗	-	-	-	93.06	41.31	37.69
PEBAL [Tian et al., 2022]	✓	✗	88.85	44.41	37.98	<u>98.52</u>	64.43	6.56
Mask2Anomaly* [Rai et al., 2023]	✓	✗	-	79.70	13.45	-	46.04	<u>4.36</u>
EAM†* [Grcic et al., 2023]	✓	✗	-	69.40	7.70	-	**81.50**	**4.20**
RbA* (Ours)	✓	✗	<u>97.99</u>	<u>85.42</u>	<u>6.92</u>	**98.62**	70.81	6.30
RbA†* (Ours)	✓	✗	**98.64**	**90.28**	**4.92**	96.62	<u>80.35</u>	4.58

Table 4.2: **Results on Road Anomaly and Fishyscapes LaF.** 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. We report the results of RbA both with and without outlier supervision. Our method RbA notably improves the results in all metrics on both datasets. † denotes that the model was trained using additional inlier data. $*$ denotes that a mask classification model is used.

4.4 Quantitative Results

4.4.1 Segment Me If You Can Benchmark

Table 4.1 shows the results on anomaly and obstacle tracks of the public SMIYC benchmark. Without outlier supervision, RbA outperforms all the models that do not use outlier supervision in training in terms of AP and FPR@95 and performs competitively compared with methods that employ outlier supervision. In terms of component-level metrics, the gains with RbA are more pronounced, which is due to an improved ability to characterize objectness compared to the previous works that are built on top of per-pixel classification

models. With outlier supervision, the performance gap improves with respect to the previous best methods (PEBAL [Tian et al., 2022], DenseHybrid [Grcić et al., 2022]) in all metrics.

Compared to concurrent work that utilizes mask-classification (EAM [Grcic et al., 2023] and Mask2Anomaly [Rai et al., 2023], RbA outperforms both of the methods in pixel-level metrics for the Obstacle Track and performs best in terms of AP on the anomaly track and second best in FPR95. In terms of component-level metrics, EAM achieves the best performance, and RbA performs competitively compared to Mask2Anomaly. The gap in performance with respect to component-level metrics between pixel-level and region-level models is apparent. However, determining the factors that correlate with the performance of component-level metrics among region-level models is an interesting direction to investigate in future work.

In Table 4.1, We report the results on two different models with outlier supervision; the difference is that one utilizes additional inlier data from the Mapillary Vistas dataset in the finetuning stage. Since our RbA scoring function represents the probability of not belonging to unknown classes, its performance heavily relies on the quality of known class predictions. Therefore, by adding additional diverse inlier data during training, we show that the outlier segmentation performance is considerably boosted. In Table 4.3, we show the pixel-level metrics on both anomaly and obstacle tracks using two different backbones with the additional Mapillary data.

Model	Anomaly Track		Obstacle Track	
	AP	FPR	AP	FPR
RbA (Swin-B)	94.46	4.60	93.68	0.15
RbA (Swin-L)	93.78	4.59	95.12	0.08

Table 4.3: **Results on SMIYC using additional data from Mapillary dataset.** In the outlier finetuning stage, the model is exposed additionally to samples from the Mapillary Vistas [Neuhold et al., 2017] dataset. We show the results using two different backbones, Swin-L and Swin-B.

Model	Known Metrics			Unknown Metrics		
	PQ	SQ	RQ	PQ	SQ	RQ
EOSPN [Hwang et al., 2021]	37.4	76.2	46.2	11.3	73.8	15.3
EAM [Grcic et al., 2023]	43.5	82.0	52.2	13.2	73.4	18.0
RbA (Ours)	**47.0**	**82.2**	**56.3**	**24.8**	**79.2**	**31.4**

Table 4.4: **Open-set panoptic segmentation results.** We show the results for known and unknown stuff classes on the Open-Set version of COCO dataset. We compare with a per-pixel classification method (EOSPN) and a mask classification method (EAM). RbA outperforms both methods on all metrics.

4.4.2 Road Anomaly & Fishyscapes LaF

Table 4.2 shows the results on the Road Anomaly [Lis et al., 2019] and the Fishyscapes Lost and Found (LaF) validation set [Blum et al., 2021]. Without outlier supervision, RbA improves the state-of-the-art significantly in almost all metrics on both datasets, even outperforming methods with outlier supervision in some metrics. With minimal supervision from a limited number of outlier objects, we obtain significant performance gains without hurting the closed-set performance (Table 5.1).

In comparison with mask-classification-based concurrent methods, RbA outperforms both EAM and Mask2Anomaly on Road Anomaly in all metrics and performs competitively on FS LaF. EAM performs best in FS LaF, while Mask2Anomaly falls behind.

4.4.3 Open-Set Panoptic Segmentation on COCO

In Table 4.4, we report the performance of RbA on the open-set split of COCO following [Hwang et al., 2021]. We compare with EOSPN [Hwang et al., 2021] and EAM [Grcic et al., 2023]. RbA outperforms both methods with a considerable margin in both the known and unknown metrics.

4.5 Qualitative Results

In Fig. 4.1 and Fig. 4.2, we show additional qualitative results of RbA compared to the previous per-pixel state-of-the-art methods PEBAL [Tian et al., 2022] and DenseHybrid [Grcić et al., 2022]. For other methods, we show both the outputs of the models reported in their respective repositories and our implementations of the methods using Mask2Former. The proposed scoring function RbA 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. 4.2) under adverse weather and lighting conditions. Moreover, compared to the baselines, RbA results in fewer false positives as a result of reducing confusion with inlier classes.

In Fig. 4.3, we show qualitative examples on the validation set of open set COCO. We show the known class predictions, as well as the RbA scoring maps for a randomly selected set of images.

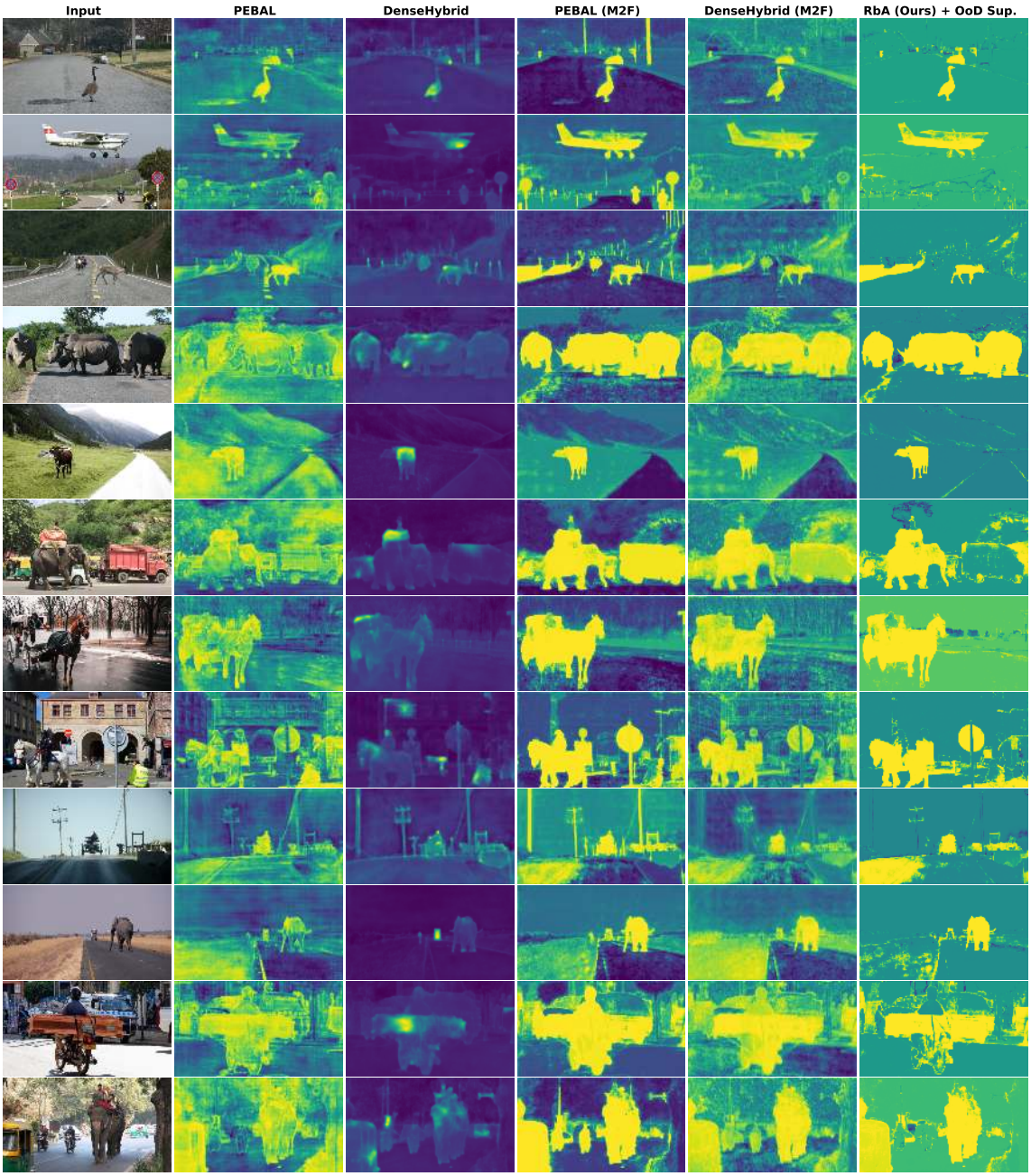


Figure 4.1: **Qualitative Results on SMIYC Anomaly Track.** On the anomaly track of the SMIYC benchmark, we compare RbA with outlier (OoD) supervision to the state-of-the-art methods PEBAL [Tian et al., 2022] and DenseHybrid [Grcić et al., 2022] using the models that were shared in their respective repositories, as well as the versions we trained using Mask2Former (M2F). RbA better distinguishes outliers from inliers and produces smoother outlier maps with fewer false positives.

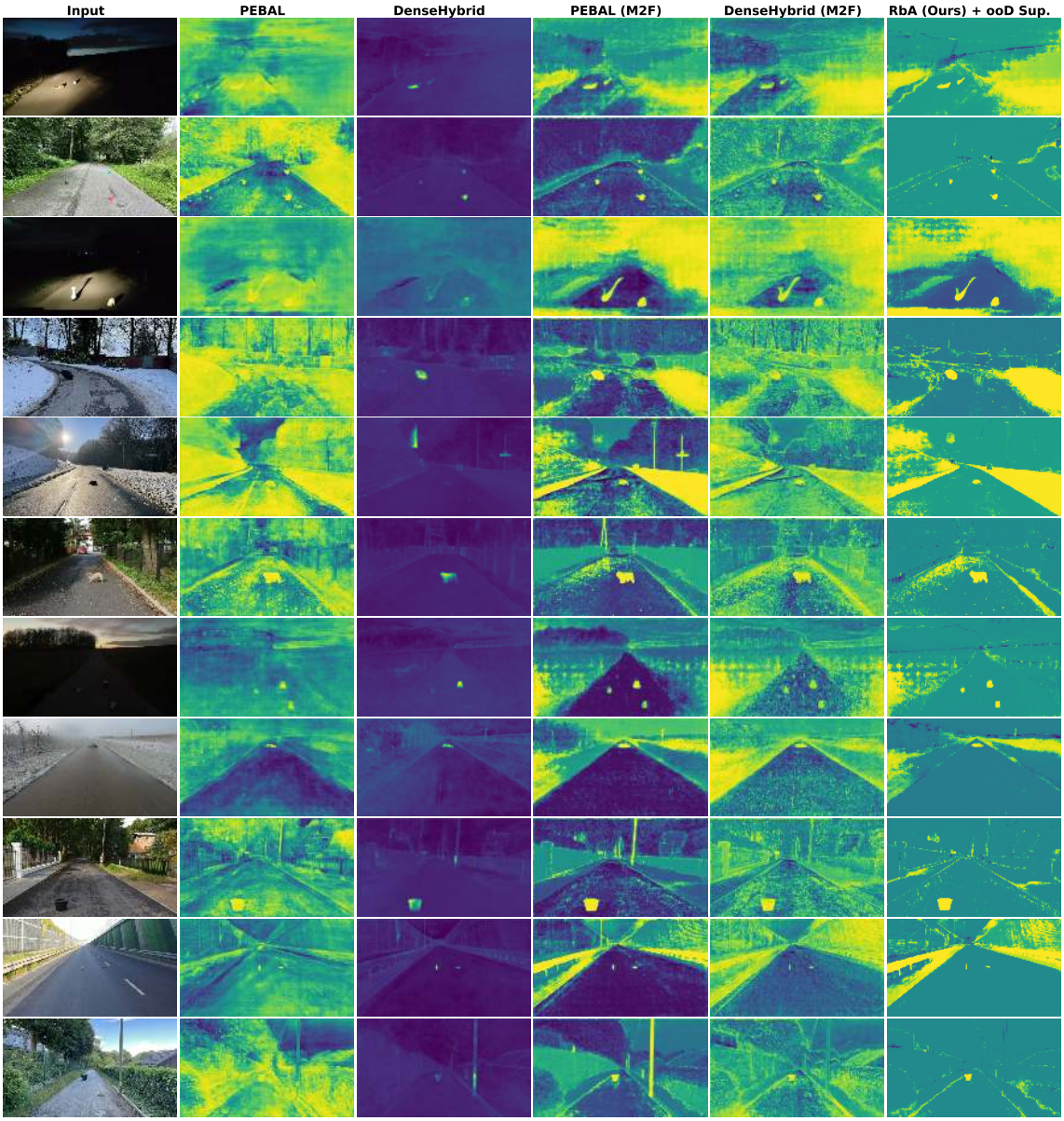


Figure 4.2: **Qualitative Results on SMIYC Obstacle Track.** We compare RbA with outlier supervision to the state-of-the-art methods PEBAL [Tian et al., 2022] and DenseHybrid [Grcić et al., 2022]. Under adverse weather and difficult lighting conditions, RbA can detect anomalies consistently better compared to DenseHybrid and reduce false positives more compared to PEBAL.

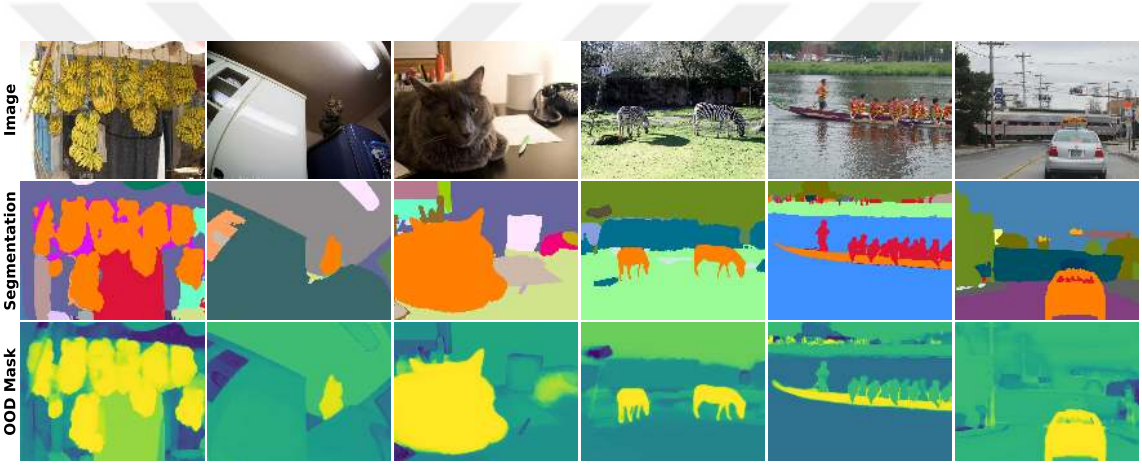


Figure 4.3: **Qualitative Results on COCO Open-Set Split** We show the known class predictions as well as the RbA score maps on images from the validation set of open-set COCO.

Chapter 5

ANALYSIS

In this chapter, we provide ablations to understand and analyze the effectiveness of RbA and justify our decision choices and hyperparameters. First, we show that mask classification improves the performance of the existing methods in OoD, but RbA better utilizes its potential. We then report the performance of the squared hinge loss compared to alternative loss functions. Afterward, we experiment with different backbones and show that optimizing for RbA improves the results with different backbones. Furthermore, we ablate architectural choices such as the number of transformer decoder layers, number of object queries, outlier data exposure, RbA loss hyperparameters, and fine-tuned components. Finally, we highlight some failure cases and limitations of our approach.

5.1 Other Methods with Mask Classification

To clearly demonstrate the effectiveness of our method and decouple it from the gains obtained by the properties of Mask2Former, we report the results of previous SOTA methods using Mask2Former, including PEBAL [Tian et al., 2022], DenseHybrid [Grcić et al., 2022], and Max Logit [Hendrycks et al., 2022] in Table 5.1.

The existing unknown segmentation methods perform remarkably better with Mask2Former; for example, the performance of PEBAL significantly improves compared to the official results reported in Table 4.2, which has been adapted from their paper [Tian et al., 2022].

As discussed in Section 3.3, the improvement comes from reducing the overconfidence issue owing to the independent behavior of object queries. Our method, RbA, performs better than the other methods in all metrics. More importantly, we achieve this performance in outlier segmentation without affecting the closed-set performance, unlike the other methods, such as PEBAL, causing a significant drop in mIoU. This experiment shows that

we can better utilize the properties of mask classification with RbA.

5.2 Alternative Loss Functions

We verify our choice of the loss function, which is a squared hinge loss, by optimizing our method with other commonly used loss functions. As can be seen in Table 5.2, squared hinge loss outperforms other loss functions. Mean Squared Error (MSE) and L1 result in a higher false positive rate. We define outlier segmentation as a binary classification problem and optimize it with BCE by using the outlier score given by the RbA as the positive class logit. While it improves the FPR compared to MSE and L1, it performs worse than the squared hinge loss in all metrics. Using KL Divergence, we minimize the distance between class probabilities of outlier pixels from a fixed distribution with maximum entropy. It performs comparably in FPR but poorly in AP, especially on the Fishyscapes LaF. Detailed formulations can be found in the appendix Section B.1.

Method	mIoU ↑	Road Anomaly		FS LaF	
		AP ↑	FPR ↓	AP ↑	FPR ↓
Max Logit [Hendrycks et al., 2022]	**82.25**	77.31	16.90	58.52	22.14
PEBAL [Tian et al., 2022]	75.32	<u>79.01</u>	<u>7.21</u>	<u>62.67</u>	25.60
DH [Grcić et al., 2022]	80.27	78.57	12.28	36.94	<u>21.12</u>
RbA (Ours)	<u>82.20</u>	**85.42**	**6.92**	**70.81**	**6.30**

Table 5.1: **Other methods with Mask2Former.** We show the performance of the state-of-the-art methods with Mask2Former. Our method RbA achieves the best results in all metrics with a clear margin, without affecting the closed-set performance, unlike previous methods. The mIoU before fine-tuning is shown in the first row of the table. The rest of the models are fine-tuned from the same checkpoint.

Method	Road Anomaly		FS LaF	
	AP ↑	FPR ↓	AP ↑	FPR ↓
KL Div.	79.91	11.33	63.58	8.78
MSE	80.71	15.79	<u>69.14</u>	22.06
L1	<u>80.94</u>	15.75	67.19	20.44
BCE	80.66	<u>10.29</u>	64.90	<u>6.89</u>
RbA (Ours)	**85.42**	**6.92**	**70.81**	**6.30**

Table 5.2: **Ablation study on alternative loss functions.** We compare our loss function based on the squared hinge loss to other commonly used loss functions. The results show that our method with squared hinge loss (RbA) performs the best in terms of OoD segmentation.

5.3 Different Backbones

We use the same Mask2Former model with Swin-B backbone [Liu et al., 2021] in almost all our experiments. In Table 5.3, we report the results with different backbones, including transformer-based Multiscale ViT (MViT) [Xiao et al., 2021] and Mix Transformer (MixT) [Xie et al., 2021] as well as convolutional WideResnet38 (WR38) [Zagoruyko and Komodakis, 2016] and ResNet101 (R101) [He et al., 2016] backbones. We keep all the other parameters the same as the default version for a fair comparison.

Fine-tuning with RbA brings consistent improvements in all metrics for all backbones. While Swin-B performs the best, R101, MViT, and MixT can still outperform previous methods on Road Anomaly and achieve competitive results on Fishyscapes LaF. This experiment shows that the proposed scoring function improves the performance regardless of the backbone.

5.4 Number of Transformer Decoder Layers

As shown in [Cheng et al., 2022], more transformer decoder layers improve the inlier performance, i.e. mIoU on Cityscapes. However, we found that using fewer decoder layers results in better performance in terms of outliers. Fig. 5.1a highlights the decrease in

Backbone	Road Anomaly		FS LaF	
	AP ↑	FPR ↓	AP ↑	FPR ↓
R101 [He et al., 2016]	38.1 → 61.9	82.7 → 37.2	30.1 → 47.1	26.3 → 12.7
WR38 [Zagoruyko and Komodakis, 2016]	21.6 → 52.0	90.0 → 43.8	24.8 → 44.6	76.3 → 13.4
MViT [Xiao et al., 2021]	57.2 → 73.1	85.8 → 24.9	47.8 → 63.7	59.8 → **6.2**
MixT [Xie et al., 2021]	65.7 → 78.1	24.6 → 12.4	40.3 → 51.3	23.0 → 17.1
Swin-B [Liu et al., 2021]	**78.5** → **85.4**	**11.8** → **6.9**	**61.0** → **70.8**	**10.6** → 6.3

Table 5.3: **Ablation study on the backbone.** We show the effect of varying the backbone used for feature extraction on the OoD performance. Comparing the results before and after fine-tuning the proposed method, we observe clear improvements in the performance of all backbones. The results are shown in the format $x \rightarrow y$, which denotes that the result x is improved to y after applying outlier finetuning with RbA.

Num Iter	AP ↑	FPR@95 ↓
1000	83.20 ± 0.11	8.69 ± 0.04
2000	85.49 ± 0.08	**7.25** ± 0.04
3000	**85.72** ± 0.12	7.82 ± 0.11
4000	84.89 ± 0.07	7.45 ± 0.05
5000	84.66 ± 0.06	8.14 ± 0.03

(a) Number of Iterations

p_{out}	AP ↑	FPR@95 ↓	mIoU ↑
0.05	84.34 ± 0.15	9.12 ± 0.10	**82.28** ± 0.02
0.1	85.48 ± 0.12	**7.24** ± 0.06	82.15 ± 0.09
0.2	**85.74** ± 0.11	7.90 ± 0.11	81.54 ± 0.02
0.4	84.97 ± 0.11	9.19 ± 0.05	81.27 ± 0.06

(b) Outlier Selection Probability

α	AP ↑	FPR@95 ↓
-0.1	84.01 ± 0.18	9.25 ± 0.06
-0.01	84.41 ± 0.14	9.06 ± 0.11
0.0	84.30 ± 0.06	9.10 ± 0.07
2	85.30 ± 0.07	7.80 ± 0.06
5	85.24 ± 0.08	**6.95** ± 0.08
10	**85.61** ± 0.10	7.26 ± 0.07

(c) Outlier Threshold

Table 5.4: **Ablation study on outlier supervision.** Finetuning with RbA loss for 2000-3000 iterations achieves the best performance and the performance deteriorates after 3000 iterations as shown in a. A higher probability of exposure to outlier data results in a consistent decline in the closed-set performance. A probability of 0.1 achieves the best balance between outlier and inlier performance b. Finally, we ablate the RbA loss parameter α in c and find that the best results are achieved with $\alpha > 0$.

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

Module	Params (%)	mIoU	Road Anomaly		FS LaF	
			AP ↑	FPR ↓	AP ↑	FPR ↓
Full Model	100	80.81	76.00	9.50	**73.88**	6.02
Transformer Dec.	1.93	80.24	85.08	10.18	72.6	**5.51**
Pixel Dec.	4.66	81.59	75.83	10.51	69.8	6.77
MLP+Linear	**0.21**	**82.20**	**85.42**	**6.92**	70.81	6.30

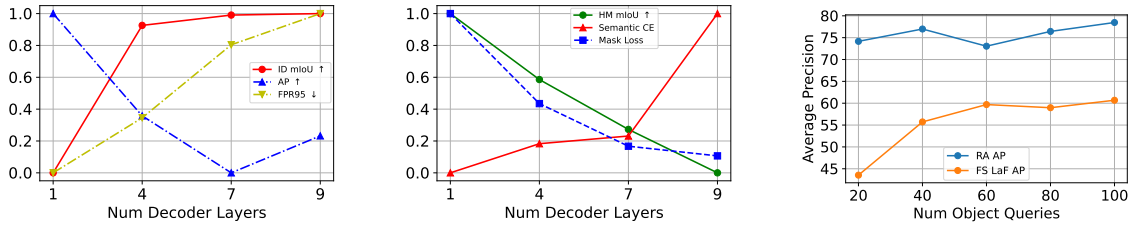
Table 5.5: **Ablation study on fine-tuning different modules.** We show the effect of fine-tuning different components of the model on the Road Anomaly and Fishyscapes LaF validation sets. Fine-tuning MLP+Linear maintains the best performance in unknown detection without sacrificing the closed-set performance.

with respect to the number of decoder layers. Fig. 5.1b shows semantic and mask losses of the Mask2Former [Cheng et al., 2022] 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 losses decrease. This shows that the increase in inlier mIoU with more decoder layers can be attributed to increased performance in detecting masks at the cost of a higher semantic error. By using fewer decoder layers, we regulate the semantic confusion, which helps to better align the logit scores, resulting in better outlier performance.

Fig. 5.1b also shows the mIoU evaluated by applying hard masking 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 increase in the semantic loss, show the importance of semantics for outlier segmentation compared to precise masks.

5.5 Number of Object Queries

The original Mask2Former [Cheng et al., 2022] uses 100 object queries. We train different models by varying the number of object queries to observe their effect on detecting outliers. We focus on the AP values on both Road Anomaly [Lis et al., 2019] with large objects



(a) Number of Decoder Layers (b) Behavior of Different Losses (c) Number of Object Queries

Figure 5.1: **Ablation study on architectural choices.** With more decoder layers, in-distribution performance on Cityscapes improves but 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.

and on Fishyscapes LaF [Blum et al., 2021] with small objects. Fig. 5.1c 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, especially for rare classes, as shown in Fig.3 in the main paper.

5.6 Outlier Data Exposure

We perform an experiment to show the effect of the number of iterations required for fine-tuning with our RbA loss in Table 5.4a. We report the AP and FPR metrics on the Road Anomaly dataset, averaged over 5 different runs to eliminate the effect of randomness. We can see that the best results can be achieved after around 2000 and 3000 iterations and then begin to degrade. We also evaluate the effect of the amount of outlier data exposed during training, which is controlled by the parameter p_{out}. We experiment with different values for p_{out} and report the outlier performance on Road Anomaly and closed-set performance on Cityscapes averaged over 5 different runs in Table 5.4b. We can see that more exposure to outlier data negatively affects the closed-set performance. Consequently, even the outlier

segmentation performance starts to degrade for $p_{out} > 0.2$. We choose the $p_{out} = 0.1$ because it strikes a reasonable balance between outlier and closed-set performance.

5.7 RbA Loss Parameter

In Table 5.4c, we report the performance of our loss function $\mathcal{L}_{\text{RbA}}$ using different values of α, averaged over 5 different runs. We find that positive values of α work similarly well and set α to 5 in our experiments. Positive values of α push the class probabilities of known classes for unknown pixels to zero. It also keeps a continuous gradient signal even if the probabilities of known classes become exactly zero. The higher α is, the higher the continuous gradient signal remains.

5.8 Which Component to Finetune ?

In Table 5.5, we analyze the effect of fine-tuning different parts of the model on validation sets of Road Anomaly [Lis et al., 2019] and Fishyscapes Lost and Found [Blum et al., 2021] using RbA loss. The alternative components we experimented with are the full model, only the transformer decoder (blue + pink in Fig. A.1), only the pixel decoder (red in Fig. A.1), and MLP+Linear layers (pink in Fig. A.1). On the Road Anomaly dataset, fine-tuning MLP+Linear achieves the best performance in terms of AP and FPR.

On Fishyscapes LaF, the best AP is achieved by fine-tuning the entire model, while the best FPR is obtained by fine-tuning only the transformer decoder. Both options cause a decrease in performance on Road Anomaly and negatively affect the closed-set performance. Fine-tuning MLP+Linear achieves the best balance between outlier detection and closed-set performance and is the least costly option in terms of the number of parameters finetuned.

5.9 Failure Cases

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

5.9.1 Tractors and Boats

As shown in Fig. 5.2, RbA fails to detect tractors and boats as outliers due to their similarity to inlier vehicle instances. Although the objects are partially identified, the model cannot decisively predict the whole object regions as outliers. The existing methods either segment the boats and tractors at the cost of more false positives, as in the case of PEBAL or also suffer from a lack of smoothness, as in the case of DenseHybrid.

5.9.2 Far away Animals

As shown in Fig. 5.3, 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.

5.9.3 Toy Cars

Fig. 5.4 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.

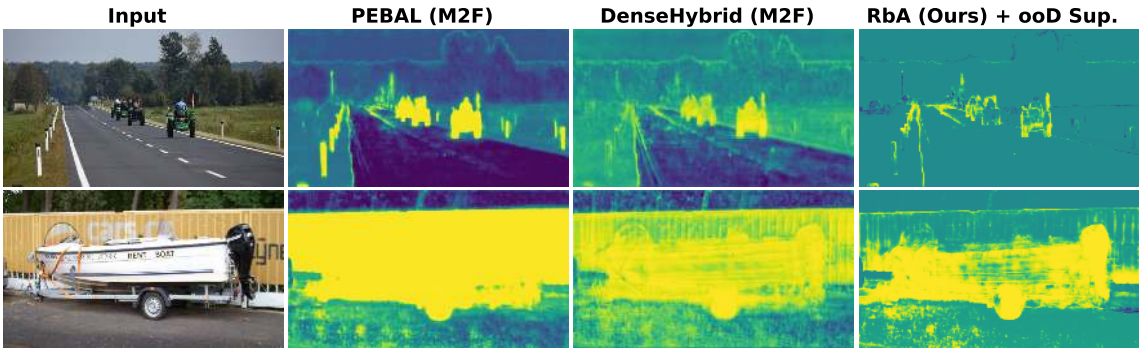


Figure 5.2: **Failure Cases: Tractors and Boats.** Due to their high similarity to inlier car and truck classes, unknown objects like tractors or boats can sometimes predicted as inliers.

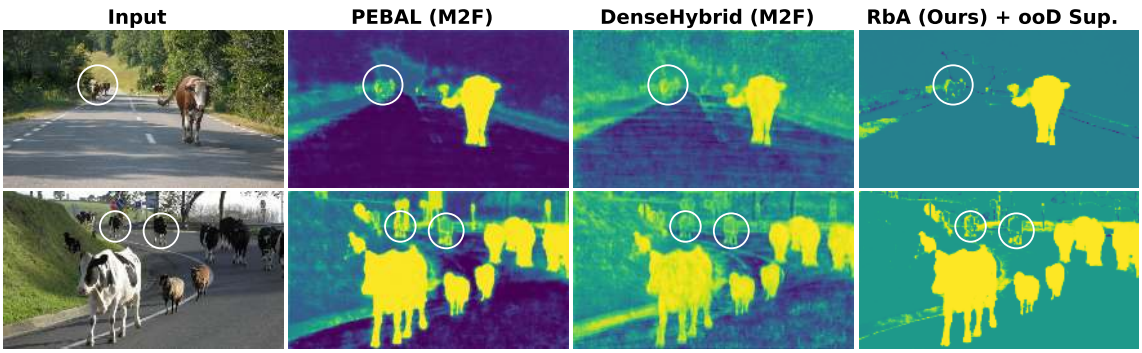


Figure 5.3: **Failure Cases: Animals Confused As Pedestrians.** As the pedestrian is one of the most frequent classes on Cityscapes, the model sometimes predicts animals that appear at a distance as pedestrians (highlighted in circles) on images from the SMIYC Anomaly Track.

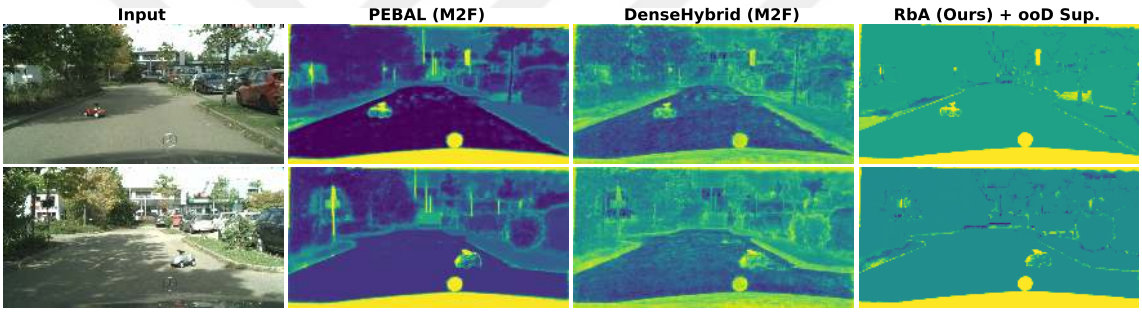


Figure 5.4: **Failure Cases: Toy Cars Predicted As Real Cars.** One confusing anomaly case for our model is small toy cars placed in front of the vehicle. Even though they can be semantically considered as cars, they are considered obstacles in a real driving scenario.

Chapter 6

CONCLUSION

6.1 Summary

In this thesis, we explore the properties and potential of the mask classification paradigm concerning uncertainty estimation and unknown segmentation. The motivation of this work is to overcome several limitations of existing methods, which include the lack of objectness in unknown map predictions, high false positive rates, and the sacrificing of closed-set performance. We demonstrate that object queries in mask classifiers behave implicitly like one vs. all classifiers, and their independent behavior reduces the overconfidence issue in the predicted class scores, resulting in improvements in the performance of the existing scoring-based unknown segmentation methods such as max logit and energy-based methods. By treating the processed object queries of mask classification as multiple one vs. all classifiers, we propose a novel outlier scoring function called Rejected by All (RbA) score defined in terms of known class probabilities. RbA represents the probability of a pixel not belonging to any of the known categories. We also propose an objective to optimize the proposed RbA score using supervision from a limited amount of outlier data, obtaining significant performance gains without affecting the closed-set performance, thereby overcoming a main limitation of the previous unknown segmentation approaches. We show that the RbA eliminates irrelevant sources of uncertainty, such as inlier boundaries and ambiguous background regions, leading to a considerable decrease in false positive rates. Moreover, RbA can preserve objectness and smoothness due to the region-level inductive biases learned by the mask classification model. We also utilize the applicability of mask classification models to panoptic segmentation to show that the RbA score can also be used to tackle open-set panoptic segmentation. We evaluate RbA on several benchmarks and show that it outperforms methods that use pixel-level segmentation models with large margins for both semantic and panoptic segmentation. We presented an analysis of the

effectiveness of RbA in terms of architecture and hyperparameters and their impact on unknown segmentation performance.

6.2 Future Work

As this work represents an initial attempt to utilize mask classification for unknown segmentation, its properties can be further explored with potential improvements. There exist several concurrent works that utilize different aspects of mask classification already. However, its properties also have a broader range of promising applications. We describe here some potential future directions that can be built on this work.

In this work, we presented a modest adaptation of RbA to panoptic segmentation. Even though it proved effective on the existing benchmark, it is still limited in cases where many instances exist adjacently. Hence, RbA can be improved further to sufficiently account for different instances for both panoptic and instance segmentation tasks.

Given the increased ability to preserve the objectness and smoothness of the predicted maps, quantified by the component level metrics proposed in the SMIYC benchmark, open-world incremental learning is one step closer. The predicted unknown masks become more reliable, thereby minimizing or eliminating the need for human annotations when encountering novel classes so that new labels can be used to expand the knowledge of incremental learning models. An idea is to utilize the tendency of object queries in mask classifiers to specialize in predicting a specific class each and dynamically add new object queries for novel classes to be learned incrementally.

While most current efforts focus on segmenting unknowns in static image datasets, video sequences are abundantly available in most driving datasets. Even if some frames of existing videos are not annotated, they still contain valuable cues that can be used to improve the segmentation of unknowns. Motion and behavior of objects such as animals or a wheelchair can be detected from videos in an unsupervised fashion and provide a meaningful signal to the segmentation model. Additionally, enforcing temporal consistency on the predicted RbA maps is likely to improve the quality of the predictions. One simple way to achieve consistency is by adding a loss that pushes RbA predictions across frames to be similar.

Apart from temporal data, more modalities with relevant cues to semantics can be utilized, such as depth, LiDAR, or even text. There exist many off-the-shelf depth prediction models that can be used to inject a 3-dimensional understanding of the vehicle surroundings in order to improve perception and the unknown prediction. Furthermore, the recent and rapid advancements of large language models (LLMs) can be utilized, especially cross-modal representations such as CLIP [Radford et al., 2021]. There are several recent works [Zhang et al., 2023, Lin et al., 2023, Ge et al., 2023, Kaul et al., 2023] that explore CLIP for segmentation generally and out-of-distribution detection specifically. Due to the rich semantics embedded in textual data, language-visual features have been found to generalize visual semantics quite robustly. Utilizing this property to enhance the performance of detecting known classes would reflect positively on the performance of RbA for segmenting unknowns. In addition, an interesting direction would be predicting RbA scores with respect to each input modality separately. For instance, for a model that expects RGB images and LiDAR as input, knowing that some LiDAR points have high anomaly scores but not their corresponding RGB regions would improve the interpretability of the system.

BIBLIOGRAPHY

[Besnier et al., 2021] Besnier, V., Bursuc, A., Picard, D., and Briot, A. (2021). Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation. In *ICCV*.

[Bevandic et al., 2018] Bevandic, P., Kreso, I., Orsic, M., and Segvic, S. (2018). Discriminative out-of-distribution detection for semantic segmentation. *arXiv.org*, 1808.07703.

[Bevandic et al., 2019] Bevandic, P., Kreso, I., Orsic, M., and Segvic, S. (2019). Simultaneous semantic segmentation and outlier detection in presence of domain shift. In *GCPR*.

[Bevandić et al., 2021] Bevandić, P., Krešo, I., Oršić, M., and Šegvić, S. (2021). Dense outlier detection and open-set recognition based on training with noisy negative images. *arXiv preprint arXiv:2101.09193*.

[Biase et al., 2021] Biase, G. D., Blum, H., Siegwart, R. Y., and Cadena, C. (2021). Pixel-wise anomaly detection in complex driving scenes. In *CVPR*.

[Blum et al., 2021] Blum, H., Sarlin, P.-E., Nieto, J. I., Siegwart, R. Y., and Cadena, C. (2021). The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *IJCV*, 129:3119–3135.

[Breitenstein et al., 2020] Breitenstein, J., Termöhlen, J.-A., Lipinski, D., and Fingscheidt, T. (2020). Systematization of corner cases for visual perception in automated driving. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1257–1264.

[Caesar et al., 2020] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2020). nuscenes: A multimodal

dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628.

[Carion et al., 2020] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *ECCV*.

[Carreira et al., 2012] Carreira, J., Caseiro, R., Batista, J. P., and Sminchisescu, C. (2012). Semantic segmentation with second-order pooling. In *ECCV*.

[Cen et al., 2021] Cen, J., Yun, P., Cai, J., Wang, M. Y., and Liu, M. (2021). Deep metric learning for open world semantic segmentation. In *ICCV*.

[Chan et al., 2021a] Chan, R., Lis, K., Uhlemeyer, S., Blum, H., Honari, S., Siegwart, R. Y., Salzmann, M., Fua, P., and Rottmann, M. (2021a). SegmentMeIfYouCan: A benchmark for anomaly segmentation. *arXiv.org*, 2104.14812.

[Chan et al., 2021b] Chan, R., Rottmann, M., and Gottschalk, H. (2021b). Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *ICCV*.

[Chen et al., 2018a] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. P., and Yuille, A. L. (2018a). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 40:834–848.

[Chen et al., 2018b] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*.

[Cheng et al., 2022] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In *CVPR*.

[Cheng et al., 2021] Cheng, B., Schwing, A. G., and Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*.

[Corbière et al., 2019] Corbière, C., Thome, N., Bar-Hen, A., Cord, M., and Pérez, P. (2019). Addressing failure prediction by learning model confidence. In *NeurIPS*.

[Cordts et al., 2016] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *CVPR*.

[Dai et al., 2017] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. (2017). Deformable convolutional networks. In *ICCV*.

[Devries and Taylor, 2018] Devries, T. and Taylor, G. W. (2018). Learning confidence for out-of-distribution detection in neural networks. *arXiv.org*, 1802.04865.

[Du et al., 2021] Du, X., Zoph, B., Hung, W.-C., and Lin, T.-Y. (2021). Simple training strategies and model scaling for object detection. *arXiv.org*, 2107.00057.

[Franchi et al., 2022] Franchi, G., Bursuc, A., Aldea, E., Dubuisson, S., and Bloch, I. (2022). One versus all for deep neural network incertitude (ovnni) quantification. *IEEE Access*.

[Fu et al., 2019] Fu, J., Liu, J., Tian, H., Fang, Z., and Lu, H. (2019). Dual attention network for scene segmentation. In *CVPR*.

[Gal and Ghahramani, 2016] Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*.

[Ge et al., 2023] Ge, Y., Ren, J., Zhao, J., Chen, K., Gallagher, A., Itti, L., and Lakshminarayanan, B. (2023). Building one-class detector for anything: Open-vocabulary zero-shot ood detection using text-image models.

[Ghiasi et al., 2021] Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., and Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*.

[Grcic et al., 2021] Grcic, M., Bevandi'c, P., and vSegvi'c, S. (2021). Dense anomaly detection by robust learning on synthetic negative data. *arXiv.org*, 2112.12833.

[Grcic et al., 2023] Grcic, M., Šarić, J., and Šegvić, S. (2023). On advantages of mask-level recognition for outlier-aware segmentation.

[Grcić et al., 2022] Grcić, M., Bevandić, P., and Šegvić, S. (2022). Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *ECCV*.

[Guo et al., 2017] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *ICML*.

[Haldimann et al., 2019] Haldimann, D., Blum, H., Siegwart, R. Y., and Cadena, C. (2019). This is not what I imagined: Error detection for semantic segmentation through visual dissimilarity. *arXiv.org*, 1909.00676.

[Hariharan et al., 2014] Hariharan, B., Arbeláez, P., Girshick, R. B., and Malik, J. (2014). Simultaneous detection and segmentation. In *ECCV*.

[Harley et al., 2017] Harley, A. W., Derpanis, K. G., and Kokkinos, I. (2017). Segmentation-aware convolutional networks using local attention masks. In *ICCV*.

[He et al., 2019] He, J., Deng, Z., and Qiao, Y. (2019). Dynamic multi-scale filters for semantic segmentation. In *ICCV*.

[He et al., 2017] He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2017). Mask R-CNN. In *ICCV*.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Hein et al., 2019] Hein, M., Andriushchenko, M., and Bitterwolf, J. (2019). Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*.

[Hendrycks et al., 2022] Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., and Song, D. X. (2022). Scaling out-of-distribution detection for real-world settings. In *ICML*.

[Hendrycks and Gimpel, 2017] Hendrycks, D. and Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*.

[Hendrycks et al., 2019] Hendrycks, D., Mazeika, M., and Dietterich, T. (2019). Deep anomaly detection with outlier exposure. In *ICLR*.

[Huang et al., 2019] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Shi, H., and Liu, W. (2019). Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*.

[Hwang et al., 2021] Hwang, J., Oh, S. W., Lee, J.-Y., and Han, B. (2021). Exemplar-based open-set panoptic segmentation network. In *CVPR*.

[Jain et al., 2023] Jain, J., Li, J., Chiu, M., Hassani, A., Orlov, N., and Shi, H. (2023). OneFormer: One Transformer to Rule Universal Image Segmentation.

[Jiang et al., 2018] Jiang, H., Kim, B., Guan, M., and Gupta, M. (2018). To trust or not to trust a classifier. In *NeurIPS*.

[Jung et al., 2021] Jung, S., Lee, J., Gwak, D., Choi, S., and Choo, J. (2021). Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *ICCV*.

[Kaul et al., 2023] Kaul, P., Xie, W., and Zisserman, A. (2023). Multi-modal classifiers for open-vocabulary object detection.

[Lakshminarayanan et al., 2017] Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*.

[Lee et al., 2018] Lee, K., Lee, K., Lee, H., and Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*.

[Li et al., 2021] Li, C.-L., Sohn, K., Yoon, J., and Pfister, T. (2021). Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9664–9674.

[Li et al., 2018] Li, H., Xiong, P., An, J., and Wang, L. (2018). Pyramid attention network for semantic segmentation. In *BMVC*.

[Li et al., 2019] Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., and Liu, H. (2019). Expectation-maximization attention networks for semantic segmentation. In *ICCV*.

[Liang et al., 2022] Liang, C., Wang, W., Miao, J., and Yang, Y. (2022). GMMSeg: Gaussian mixture based generative semantic segmentation models. *arXiv.org*, 2210.02025.

[Liang et al., 2018] Liang, S., Li, Y., and Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *ECCV*.

[Lin et al., 2023] Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., and He, X. (2023). Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15305–15314.

[Lis et al., 2020] Lis, K., Honari, S., Fua, P., and Salzmann, M. (2020). Detecting road obstacles by erasing them. *arXiv.org*, 2012.13633.

[Lis et al., 2019] Lis, K., Nakka, K. K., Fua, P. V., and Salzmann, M. (2019). Detecting the unexpected via image resynthesis. In *ICCV*.

[Liu et al., 2020] Liu, W., Wang, X., Owens, J., and Li, Y. (2020). Energy-based out-of-distribution detection. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *NeurIPS*.

[Liu et al., 2021] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.

[Loshchilov and Hutter, 2019] Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

[Malinin and Gales, 2018] Malinin, A. and Gales, M. J. F. (2018). Predictive uncertainty estimation via prior networks. In *NeurIPS*.

[Milletari et al., 2016] Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*.

[Minderer et al., 2021] Minderer, M., Djolonga, J., Romijnders, R., Hubis, F. A., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. (2021). Revisiting the calibration of modern neural networks. In *NeurIPS*.

[Mukhoti and Gal, 2018] Mukhoti, J. and Gal, Y. (2018). Evaluating bayesian deep learning methods for semantic segmentation. *arXiv.org*, 1811.12709.

[Nalisnick et al., 2018] Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. (2018). Do deep generative models know what they don't know? In *ICLR*.

[Neuhold et al., 2017] Neuhold, G., Ollmann, T., Rota Bulo, S., and Kontschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*.

[Nguyen et al., 2015] Nguyen, A. M., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable

visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

[Rai et al., 2023] Rai, S. N., Cermelli, F., Fontanel, D., Masone, C., and Caputo, B. (2023). Unmasking anomalies in road-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[Ren et al., 2019] Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., and Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. In *NeurIPS*.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *IJCV*, 115:211–252.

[Serrà et al., 2020] Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J. F., and Luque, J. (2020). Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv.org*, 1909.11480.

[Shelhamer et al., 2015] Shelhamer, E., Long, J., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *CVPR*.

[Strudel et al., 2021] Strudel, R., Pinel, R. G., Laptev, I., and Schmid, C. (2021). Segmenter: Transformer for semantic segmentation. In *ICCV*.

[Szegedy et al., 2013] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv.org*, 1312.6199.

[Tian et al., 2022] Tian, Y., Liu, Y., Pang, G., Liu, F., Chen, Y., and Carneiro, G. (2022). Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *ECCV*.

[Torralba et al., 2008] Torralba, A., Fergus, R., and Freeman, W. T. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. *PAMI*, 30:1958–1970.

[Vojir et al., 2021] Vojir, T., Šipka, T., Aljundi, R., Chumerin, N., Reino, D. O., and Matas, J. (2021). Road anomaly detection by partial image reconstruction with segmentation coupling. In *ICCV*.

[Wang et al., 2021] Wang, H., Zhu, Y., Adam, H., Yuille, A. L., and Chen, L.-C. (2021). MaX-DeepLab: End-to-end panoptic segmentation with mask transformers. In *CVPR*.

[Xia et al., 2020] Xia, Y., Zhang, Y., Liu, F., Shen, W., and Yuille, A. (2020). Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *ECCV*.

[Xiao et al., 2021] Xiao, H., Zhang, Y., Dana, K., and Shi, J. (2021). Multiscale vision transformers. In *CVPR*.

[Xie et al., 2021] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Álvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*.

[Yang et al., 2018] Yang, M., Yu, K., Zhang, C., Li, Z., and Yang, K. (2018). Denseaspp for semantic segmentation in street scenes. In *CVPR*.

[Yang et al., 2021] Yang, Z., Wei, Y., and Yang, Y. (2021). Associating objects with transformers for video object segmentation. In *NeurIPS*.

[Yu et al., 2020] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. (2020). BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*.

[Yu and Koltun, 2016] Yu, F. and Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *ICLR*.

[Yuan et al., 2020] Yuan, Y., Chen, X., and Wang, J. (2020). Object-contextual representations for semantic segmentation. In *ECCV*.

[Yuan et al., 2021] Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., and Wang, J. (2021). HRFormer: High-resolution transformer for dense prediction. *arXiv.org*, 2110.09408.

[Zagoruyko and Komodakis, 2016] Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. In *BMVC*.

[Zendel et al., 2018] Zendel, O., Honauer, K., Murschitz, M., Steininger, D., and Dominguez, G. F. (2018). Wilddash - creating hazard-aware benchmarks. In *ECCV*.

[Zhang et al., 2023] Zhang, J., Liu, R., Shi, H., Yang, K., Reiß, S., Peng, K., Fu, H., Wang, K., and Stiefelhagen, R. (2023). Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1136–1147.

[Zhang et al., 2021] Zhang, L. H., Goldstein, M., and Ranganath, R. (2021). Understanding failures in out-of-distribution detection with deep generative models. In *ICML*.

[Zhao et al., 2017] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *CVPR*.

[Zhao et al., 2018] Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C. C., Lin, D., and Jia, J. (2018). PSANet: Point-wise spatial attention network for scene parsing. In *ECCV*.

[Zheng et al., 2021] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. S., and Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*.

[Zhou et al., 2017] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. (2017). Scene parsing through ADE20K dataset. In *CVPR*.

[Zhu et al., 2021] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021). Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*.

Appendix A

IMPLEMENTATION DETAILS

A.1 Architecture

Fig. A.1 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.

A.1.1 Backbone

We mainly use the Swin-B variant as the backbone [Liu et al., 2021]. 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 ($\mathbf{x}_4$), 8 ($\mathbf{x}_8$), 16 ($\mathbf{x}_{16}$), and 32 ($\mathbf{x}_{32}$) with respect to the input image.

A.1.2 Pixel Decoder

Following [Cheng et al., 2022], the pixel decoder mainly consists of 6 layers of deformable attention (MSDeformAttn) [Zhu et al., 2021]. The multi-scale feature maps with strides $\mathbf{x}_8$ $\mathbf{x}_{16}$, and $\mathbf{x}_{32}$ are processed with MSDeformAttn layers to produce $\mathbf{f}_1$, $\mathbf{f}_2$, and $\mathbf{f}_3$, respectively. In [Cheng et al., 2022], 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 $\mathbf{f}_3$ to the transformer decoder. The feature map $\mathbf{x}_4$ is processed with a 1×1 filter-size convolutional layer and then added to the processed feature map $\mathbf{f}_1$ after bilinear upsampling. Finally, the output is passed to a 3×3 convolutional layer to produce per-pixel features $\mathbf{F} \in \mathbb{R}^{C_p \times H \times W}$, where $C_p = 256$ is the embedding dimension. The computation can be summarized as follows:

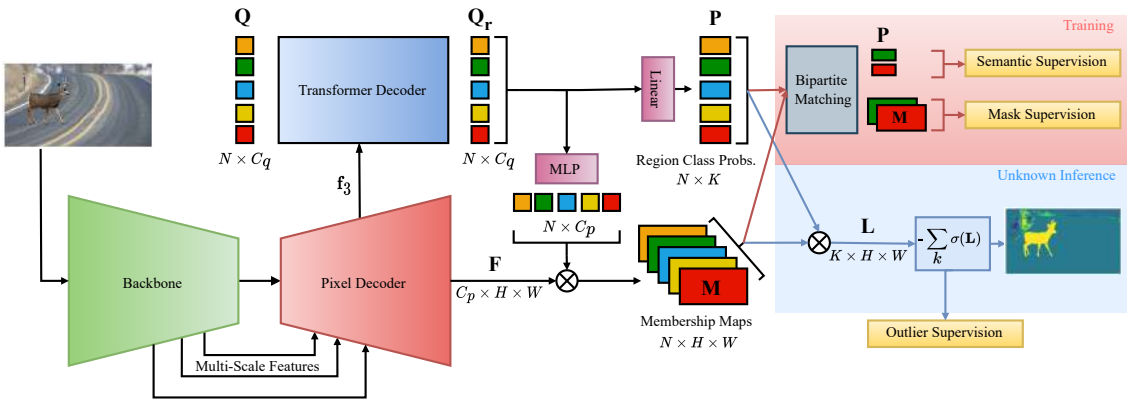


Figure A.1: **Detailed architecture.** This figure provides a more detailed view of the Mask2Former [Cheng et al., 2022] architecture, including our modifications and unknown inference computation. We use a single transformer decoder layer as opposed to the original implementation that uses 9 layers. Therefore, only a single scale feature $\mathbf{f}_3$ 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.

$$\mathbf{F} = \text{Conv}_{3\times3}\left(\text{Conv}_{1\times1}(\mathbf{x}_4) + \text{Upsample}(\mathbf{f}_1)\right) \tag{A.1}$$

A.1.3 Transformer Decoder

Learnable object queries $\mathbf{Q} \in \mathbb{R}^{N\times C_q}$ 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 $C_q = 256$ is the embedding dimension. A single transformer decoder layer 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, and it predicts directly from the input feature map before being processed in the transformer decoder. Learnable positional embeddings are added to the object queries. Pixel features are supported by sinusoidal positional embeddings

and learnable level embeddings. The transformer decoder outputs a refined set of object queries $\mathbf{Q}_r$ that predict the regions and classify them.

A.1.4 Region Class Prediction

The refined object queries $\mathbf{Q}_r$ are fed into a single linear layer followed by a softmax to produce the class probability of each region $\mathbf{P} \in \mathbb{R}^{N \times K}$ where K is the number of classes.

A.1.5 Membership Maps Prediction

The refined object queries $\mathbf{Q}_r$ are also fed into a 3-layer MLP so that $\mathbf{Q}_r$'s dimensionality matches that of $\mathbf{F}$. Then, $\mathbf{Q}_r$ and $\mathbf{F}$ are multiplied before being fed into a sigmoid activation to produce the per-pixel membership maps $\mathbf{M} \in \mathbb{R}^{N \times H \times W}$.

A.2 Closed-Set Training

A.2.1 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.

A.2.2 Hyper-Parameters

Following [Cheng et al., 2022], the model is trained for 90K iterations using a batch size of 16. AdamW [Loshchilov and Hutter, 2019] 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.

A.2.3 Data Augmentation

We use the same augmentations as in [Cheng et al., 2022]. 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 [Du et al., 2021, Ghiasi et al., 2021] is applied with a random horizontal flip.

A.3 Outlier Supervision

A.3.1 Data Sampling

We use a slightly modified version of AnomalyMix proposed in [Tian et al., 2022] for outlier supervision. After eliminating the samples that contain Cityscapes classes [Cordts et al., 2016], around 40K images remain for outlier supervision on the COCO [Lin et al., 2014] dataset. For a single fine-tuning experiment, we randomly sample 300 images and fix them throughout the entire fine-tuning phase.

A.3.2 Fine-tuned Components

For all the fine-tuning experiments, we only fine-tune the 3-layer MLP and linear layers shown in pink in Fig. A.1. Their weights together constitute approximately 0.21% of the entire model parameters.

A.3.3 Hyper-Parameters

After the model is trained on the closed-set setting, we fine-tune it for 2000 iterations on Cityscapes [Cordts et al., 2016] using the setting of the closed-set training; AdamW [Loshchilov and Hutter, 2019] 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 RbA score for outlier pixels is optimized with the squared hinge loss $\mathcal{L}_{RbA}$ using $\alpha = 5$.

Appendix B

EXPERIMENTS DETAILS

In this section, we provide further illustrations and detailed settings of our analysis and ablation experiments. We first perform an experiment to verify the specialization of object queries. We then provide a detailed formulation of our ablations on the loss functions and other methods using Mask2Former, including the hyper-parameters that we use to obtain the results presented in the main paper.

B.1 Other Loss Functions

In the analysis section of the thesis, we perform an ablation study with different loss functions in comparison to squared hinge loss. In this section, we provide the formulation and the parameter setting of each loss function. In the following, Ω_{out} denotes the set of outlier pixels, and Ω_{in}, the set of inlier pixels on an image.

B.1.1 Mean-Squared Error (MSE)

With MSE loss, we optimize the RbA to be closer to $\alpha = 5$ for outlier pixels:

$$\mathcal{L}_{\text{MSE}} = \sum_{\mathbf{x} \in \Omega_{out}} \left(\text{RbA}(\mathbf{x}) - \alpha\right)^2 \tag{B.1}$$

B.1.2 Mean Absolute Error (L1)

Similar to MSE, we optimize the RbA with L1 loss by setting α to 5:

$$\mathcal{L}_{L1} = \sum_{\mathbf{x} \in \Omega_{out}} \left|\text{RbA}(\mathbf{x}) - \alpha\right| \tag{B.2}$$

B.1.3 Binary Cross Entropy (BCE)

We formulate the scoring of outliers as a per-pixel binary classification problem, where outliers correspond to the positive class. We use the RbA score as the logit score for the positive class. Assuming that y corresponds to the binary label (outlier vs. inlier) of a pixel $\mathbf{x}$, we optimize the BCE loss as follows:

$$\mathcal{L}_{\text{BCE}} = \sum_{\mathbf{x} \in \Omega_{out}} y \cdot \log\left(\text{RbA}(\mathbf{x})\right) + (1 - y) \cdot \log\left(1 - \text{RbA}(\mathbf{x})\right) \tag{B.3}$$

B.1.4 KL Divergence

We use the KL Divergence to minimize the distance between the predicted class distribution to a fixed distribution. For inlier pixels, we minimize the distance to the Dirac delta function where the correct class has a probability of 1.0. For outlier pixels, we minimize the distance to a uniform distribution where the entropy is maximum. Even though this loss function does not optimize our proposed score function RbA, it helps us estimate the contribution of RbA by pushing the predicted class probabilities toward the ideal distributions for outliers and inliers. Let $\mathbf{L}_p(\mathbf{x})$ be the class probability distribution of a pixel $\mathbf{x}$ with ground truth label y. Let $\mathbf{P}_{in}(y)$ be a fixed probability distribution where class y has probability 1.0. Let $\mathcal{U}$ be a fixed uniform probability distribution. Formally, the loss function is defined as follows:

$$\begin{aligned}
\mathcal{L}_{in} &= \sum_{\mathbf{x} \in \Omega_{in}} D_{KL}(\mathbf{L}_p(\mathbf{x}) \parallel \mathbf{P}_{in}(y)) \\
\mathcal{L}_{out} &= \sum_{\mathbf{x} \in \Omega_{out}} D_{KL}(\mathbf{L}_p(\mathbf{x}) \parallel \mathcal{U}) \\
\mathcal{L}_{KL} &= \frac{1}{2}\left(\mathcal{L}_{in} + \mathcal{L}_{out}\right)
\end{aligned} \tag{B.4}$$

B.2 Other Methods with Mask2Former

To disentangle the contribution of our scoring function RbA from the architecture, we presented the results of the state-of-the-art methods PEBAL [Tian et al., 2022] and DenseHybrid [Grcić et al., 2022] using the Mask2Former architecture. Here, we provide the

details of this ablation experiment for each method. For a fair comparison, we train both methods by fine-tuning the same components as the RbA scoring function, that is, the MLP+Linear layers as shown in Fig. A.1, and also using the same outlier data supervision method as described in the thesis.

B.2.1 PEBAL

The optimized objective as per the official implementation [Tian et al., 2022] consists of three components. First, there is the pixel-wise anomaly abstention loss (PAL) with the abstention term and penalty defined as follows:

$$\mathcal{L}_{PAL} = -\sum_{\mathbf{x}\in\Omega} \log\left(\mathbf{L}_y(\mathbf{x}) + \frac{\mathbf{L}_{K+1}(\mathbf{x})}{a(\mathbf{x})}\right)$$

where Ω is the set of all pixels on a given image, $y \in 1, \ldots, K+1$ is the ground truth class of pixel $\mathbf{x} \in \Omega$, and $K+1$ is the class for the outliers. In Mask2Former, class $K+1$ is assumed to be the no object class, therefore we avoid dropping it from the region class probabilities term $\mathbf{P}$ (see Fig. A.1) while fine-tuning with the PEBAL objective. The abstention penalty term $a(\mathbf{x})$ is defined as follows:

$$\begin{aligned} a(\mathbf{x}) &= \left(-E(\mathbf{x})\right)^2 \\ E(\mathbf{x}) &= -\log\sum_{k=1}^{K}\exp(\mathbf{L}_k(\mathbf{x})) \end{aligned} \tag{B.5}$$

where $E(\mathbf{x})$ is the free energy function. When the penalty term is high, the prediction is discouraged from abstaining and vice versa.

The second component of the loss optimizes the energy terms such that it is maximized for outlier pixels and minimized for inlier pixels as follows:

$$\begin{aligned} \mathcal{L}_{energy} = &\sum_{\mathbf{x}\in\Omega_{in}} \max(0, E(\mathbf{x}) - m_{in})^2 + \\ &\sum_{\mathbf{x}\in\Omega_{out}} \max(0, m_{out} - E(\mathbf{x}))^2 \end{aligned} \tag{B.6}$$

where m_{in} and m_{out} are hyper-parameters to be set. In our experiments, we experimentally use $m_{out} = -2.5$ and $m_{in} = -3.5$.

The last component is a regularization term for the smoothness and sparsity of the predicted energy map:

$$\mathcal{L}_{reg} = \sum_{\mathbf{x}\in\Omega} \beta_1 |E(\mathbf{x}) - E(\mathcal{N}(\mathbf{x}))| + \beta_2 |E(\mathbf{x})| \tag{B.7}$$

where $\mathcal{N}(\mathbf{x})$ is the set of vertical and horizontal neighboring pixels of $\mathbf{x}$, and β_1 and β_2 are hyper-parameters. We use $\beta_1 = 3 \times 10^{-7}$ and $\beta_2 = 5 \times 10^{-5}$.

The full objective is defined as the weighted sum of the three loss functions:

$$\mathcal{L}_{PEBAL} = \mathcal{L}_{PAL} + \beta \mathcal{L}_{energy} + \mathcal{L}_{reg} \tag{B.8}$$

where we set $\beta = 0.1$. Starting from the same checkpoint that we use fine-tuning RbA, we optimize the model with $\mathcal{L}_{PEBAL}$ for 5K iterations. We set other hyper-parameters to be the same as the ones that we use for RbA.

B.2.2 DenseHybrid

Following the official implementation of DenseHybrid [Grcić et al., 2022], we added an additional outlier prediction head $\mathbf{D}(\mathbf{x}) \in \mathbb{R}^{2\times H\times W}$ to the Mask2Former model which is defined as follows:

$$\mathbf{D}(\mathbf{x}) = \mathrm{Conv}_{3\times 3}\left(\mathrm{ReLU}(\mathrm{BatchNorm}(\mathbf{x}))\right) \tag{B.9}$$

The outlier prediction head takes the output feature map of the highest resolution from the pixel decoder and predicts a binary output for every pixel denoting the probability of being an outlier. The objective for DenseHybrid is defined as follows:

$$\mathcal{L}_{DH} = \mathrm{CE}(\mathbf{L}(\mathbf{x}_{in}), \mathbf{Y}_{in}) + \beta_1 \mathrm{CE}(D(\mathbf{x}), \mathbf{Y}_{out}) + \beta_2 \mathcal{L}_o \tag{B.10}$$

where CE is short for the cross-entropy loss, $\mathbf{x}_{in} \in \Omega_{in}$ denotes the set of inlier pixels, $\mathbf{Y}_{in}$ denotes the ground truth for the closed-set and $\mathbf{Y}_{out}$ denotes the binary map where the outlier pixels are set to one. We experimentally set the hyper-parameters $\beta_1 = 0.3$ and $\beta_2 = 0.03$. $\mathcal{L}_o$ is defined as follows:

$$\mathcal{L}_o = \frac{1}{|\Omega_{out}|} \sum_{\mathbf{x} \in \Omega_{out}} \log \sum_{k=1}^{K} \exp\left(\mathbf{L}_k(\mathbf{x})\right) + \mathrm{sg}[\mathrm{mean}(\mathbf{L}(\mathbf{x}))] \tag{B.11}$$

where sg is short for the stop gradient operation, and mean denotes the mean of all the elements of the input tensor.