Open-World Semantic Segmentation Including Class Similarity

Matteo Sodano[1] Federico Magistri[1] Lucas Nunes[1] Jens Behley[1] Cyrill Stachniss[1,2]
[1]Center for Robotics, University of Bonn [2]Lamarr Institute for Machine Learning and Artificial Intelligence
{firstname.lastname}@igg.uni-bonn.de

Abstract

Interpreting camera data is key for autonomously acting systems, such as autonomous vehicles. Vision systems that operate in real-world environments must be able to understand their surroundings and need the ability to deal with novel situations. This paper tackles open-world semantic segmentation, i.e., the variant of interpreting image data in which objects occur that have not been seen during training. We propose a novel approach that performs accurate closed-world semantic segmentation and, at the same time, can identify new categories without requiring any additional training data. Our approach[1] additionally provides a similarity measure for every newly discovered class in an image to a known category, which can be useful information in downstream tasks such as planning or mapping. Through extensive experiments, we show that our model achieves state-of-the-art results on classes known from training data as well as for anomaly segmentation and can distinguish between different unknown classes.

1. Introduction

Autonomous systems need to understand their surroundings to operate robustly. To this end, semantic scene understanding based on sensor data is key and numerous variants exist, such as object detection [17, 52], semantic and instance segmentation [40, 42, 68], and panoptic segmentation [29, 30]. Over the last decade, we witnessed tremendous progress in scene interpretation for autonomous vehicles using machine learning. A central challenge for most learning-based systems is scenes in which novel and previously unseen objects occur. Such *open-world* settings, i.e., the fact that not everything can be covered in the training data, have to be considered when building vision systems for human-centered environments and real-world settings. For example, autonomous cars in cities will eventually experience situations or objects they have not seen before. They should be able to identify them, for example, to change into a more conservative mode of operation.

Today, high-quality datasets such as Cityscapes [13] or MS COCO [35] allow deep learning methods to achieve outstanding performance in closed-world scene understanding tasks. A prominent task is semantic segmentation [18], which aims to assign a semantic category to each pixel in an image. Systems operating under the *closed-world* assumption [56] typically cannot correctly recognize an object that belongs to none of the known categories. Often, they tend to be overconfident and assign such an object to one of the known classes. We believe that for applications targeting reliability and robustness under varying conditions, the closed-world assumption has to be relaxed, and we need to move towards open-world setups. Additionally, a measure of class similarity can help downstream tasks. For example, predicting that an area of the image is unknown but similar to the class car or another type of moving vehicle can be

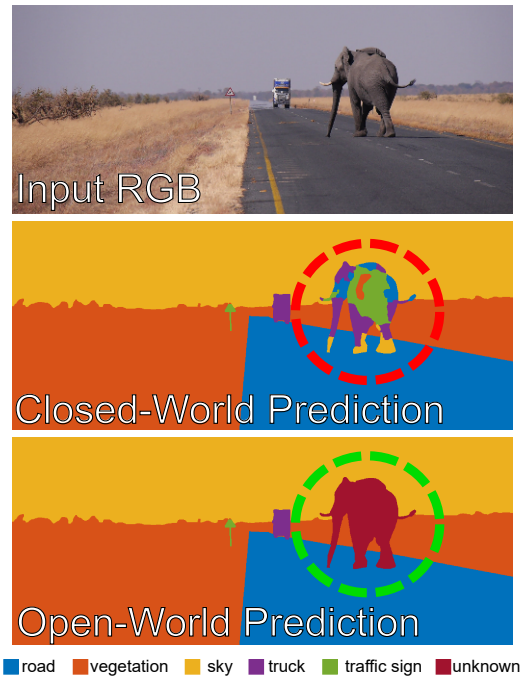


Figure 1. Given an image containing a previously-unseen object (top), closed-world methods for semantic segmentation classify the pixels belonging to that object as one of the known classes (center, red circle). Our goal is to segment the unknown object and identify it as a semantic class different to the previously-known ones (bottom, green circle).

[1]Code: https://github.com/PRBonn/ContMAV

used in planning or tracking to estimate the motion of the object, or in mapping to discard that class from the map.

This paper investigates the problem of *open-world* semantic segmentation. Given an image at test time, we aim to have a model that is able to detect any pixel that belongs to a category that was unseen at training time and is also able to distinguish between different new categories. The first problem is called *anomaly segmentation* [9] and aims to achieve a binary segmentation between known and unknown. The second problem, called *novel class discovery* [21], aims to obtain a pixel-wise classification of novel samples into different classes starting from the knowledge of previously seen, labeled samples. We aim to investigate how to solve both tasks jointly in a neural network setting. We extend best-practice approaches for anomaly detection for classification tasks [2, 15, 70] and provide compelling results for both, anomaly segmentation and novel class discovery. See Fig. 1 for an example of the targeted output.

The main contribution of this paper is a novel approach for open-world segmentation based on an encoder-decoder convolutional neural network (CNN). We propose a new method that simultaneously performs accurate closed-world semantic segmentation while constraining all known classes towards their learned feature descriptor, thanks to a loss function we introduce. We combine operations on the feature space with binary anomaly segmentation that allows us to distinguish between different novel classes and provide a measure of similarity for every newly discovered class to a known category. We implemented and thoroughly tested our approach. In sum, our contributions are the following:

- A fully-convolutional neural network that achieves state-of-the-art performance on anomaly segmentation while providing compelling closed-world performance.
- A loss function that allows us to distinguish among different novel classes, and to provide a similarity score for each novel class to the known categories.
- Extensive experiments on multiple datasets, including the public benchmark SegmentMeIfYouCan, where we rank first in three out of five metrics.

2. Related Work

Semantic segmentation under closed-world settings achieved outstanding performances in different domains, such as autonomous driving [7, 43, 45, 65], indoor navigation [25, 32, 59], or agricultural robotics [12, 44, 53, 67]. However, the closed-world assumption should be relaxed when developing systems for navigating in the wild. In such cases, we need to move towards open-world setups.

Anomaly Detection and Classification. The open-world setting was initially explored for classification, where anomalous samples had to be recognized and discarded. This problem was tackled in different ways in the literature. Simple strategies such as thresholding the softmax activations [8, 23], using a background class for tackling unknown samples [6, 46, 63], and using model ensembles [33, 64] represent a solid starting point in theory. In practice, however, closed-world predictions tend to be overconfident by showing a peak in the softmax even for unknown samples [48, 61]. Additionally, it is impossible to train with all possible examples of unknown objects. To deal with this, modifications to the softmax layer have been proposed [2]. Other approaches rely on maximizing the entropy [15] or on energy scores [38], which are supposedly less susceptible to the aforementioned overconfidence issue. Even though these approaches can be easily adapted to the segmentation problem, they are limited in the sense that they rely on the output of the CNN to be "uncertainty-aware" to some extent, in order to be able to modify the scores, consider the output entropies, or similar.

In contrast, we operate on the feature space of the semantic segmentation to not only classify pixels correctly but also match their feature to a unique class descriptor, in order to use the distance from it as a measure of "unknown-ness".

Open-World Segmentation. Open-world or anomaly segmentation extends the anomaly detection task by trying to predict whether each individual pixel in an image is an anomaly or not. Some methods rely on the estimation of the uncertainty on the prediction with Bayesian deep learning [16, 33, 57], or on the gradient [37, 41]. Other works use an additional dataset for out-of-distribution samples, in order to help the CNN recognize categories that do not belong to the standard training set [5, 10]. Recently, generative models have also been used, since in the reconstruction phase they will accurately resynthesize only the known areas, while unknown objects will suffer from a lower reconstruction quality, and can be recognized by looking at the most dissimilar areas between the input and the output [31, 36, 76]. Due to the limitation of available training data, many unsupervised approaches use synthetic anomaly data and train an anomaly detector which is either distance-based [39, 55, 62] or reconstruction-based [3, 72, 74], with the latter sharing the same concept as the generative models mentioned above. Vision-language models based on CLIP [50, 51, 77] are also gaining interest in the context of anomaly segmentation [27]. Lately, a lot of research interest is also going in the direction of anomaly segmentation in video streams because of its application in intelligent surveillance systems [60, 69, 73].

Differently from these approaches, we do not require additional data for training and do not rely on uncertainty estimation or generative models. In contrast, by operating on the feature space of the semantic segmentation task, we can define a distinct region for known and unknown classes. Furthermore, we leverage the feature descriptors of the known classes to recognize different unknown classes and find the most similar known class.

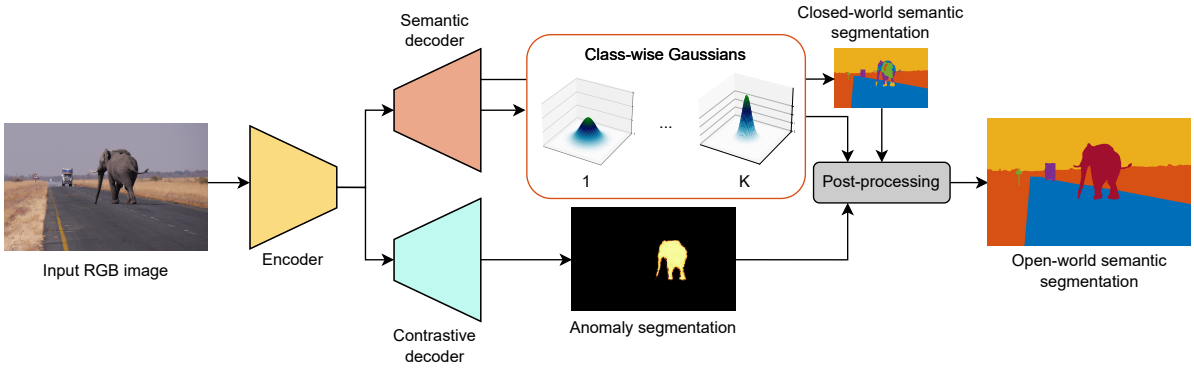


Figure 2. Given an RGB image as input, our network processes it by means of an encoder and two decoders. The semantic decoder produces a closed-world semantic segmentation and a Gaussian model for each known category. The class Gaussian models are built from a learned class descriptor (mean) and the variance of all predictions from it. A 3D example is shown in the image. The contrastive decoder provides an anomaly segmentation output. A post-processing phase finally achieves open-world semantic segmentation.

3. Our Approach

In this work, we tackle the problem of open-world semantic segmentation. In addition to handling known classes, we are particularly interested in segmenting all anomalous areas in an image, where previously unseen objects appear, and in differentiating between potentially multiple novel classes. We propose an approach (see Fig. 2) based on a convolutional neural network with one encoder and two decoders. The first decoder tackles semantic segmentation and operates on the feature space so that, for each class, features of pixels belonging to the same class are pushed together. The mean and variance of each individual class descriptor are stored representing Gaussian distributions that describe known classes. The second decoder performs binary anomaly segmentation. Results are finally merged to obtain open-world semantic segmentation, i.e. anomaly segmentation and novel class discovery.

3.1. General Network Architecture

Our network for open-world semantic segmentation is composed of one encoder and two decoders. We use a ResNet34 [22] encoder, where the basic ResNet block is replaced with the NonBottleneck-1D block [54], which allows a more lightweight architecture since all 3×3 convolutions are replaced by a sequence of 3×1 and 1×3 convolutions with a ReLU in between. For open-world segmentation, contextual information is valuable. Therefore, we expand the limited receptive field of ResNet by incorporating contextual information using a pyramid pooling module [75] after the encoding part. The features produced will be fed to two separate decoders, that share the same structural properties. In order to preserve the lightweight nature of the network, we use three SwiftNet modules [49] where we incorporate NonBottleneck-1D blocks, and two final upsampling modules based on nearest-neighbor and depth-wise convolutions, which reduce the computational load. We use encoder-decoder skip connections after each downsampling stage of the encoder to directly propagate more fine-grained features to the decoder.

3.2. Approach for Open-World Segmentation

Our approach for open-world segmentation builds upon the structure of the CNN we developed, and it exploits the double-decoder architecture for providing accurate segmentation of unknown regions. The first decoder, which we call "semantic decoder" in the following, targets semantic segmentation. We additionally manipulate the feature space to create a unique distinct descriptor for each known class. Our goal is to obtain a correct semantic segmentation for the known classes, but also produce pre-softmax features that are similar to the descriptor for each pixel of a certain class. In this way, we aim to detect as unknown classes all pixels whose feature vectors are substantially different from the descriptor of the class they have been assigned to. The second decoder, which we call "contrastive decoder" in the following, leverages the contrastive loss [11] and objectosphere loss [15] together, to place all features of known classes on the surface of a hypersphere while pushing the ones of unknown classes towards its center. In this way, the second decoder provides an anomaly segmentation, where the anomalous regions correspond to previously unseen classes. The two results are finally merged using an automated post-processing operation to obtain the final open-world segmentation.

In the following, we call $\Omega = \{(1, 1), \ldots, (H, W)\}$ the set of pixels in the image, $Y \in \{1, \ldots, K\}^{H \times W}$ the ground truth mask, and $\hat{Y} \in \{1, \ldots, K\}^{H \times W}$ the predicted mask, where H and W are the dimensions of the input image. Additionally, we denote with $\Omega_k = \{p \in \Omega \mid Y_p = k\}$ the set of pixels whose ground truth label is k, and with $\hat{\Omega}_k = \{p \in \Omega \mid \hat{Y}_p = Y_p\}$ the set of pixels that are true posi-

tives for class k, *i.e.*, the set of pixels whose ground truth label and predicted label are k. Finally, the square of a vector refers to the element-wise operation (Hadamard product):

$$\boldsymbol{v}^2 = \left[v_1^2, \ldots, v_n^2\right]^\top \tag{1}$$

Semantic Decoder. The aim of semantic segmentation is to predict a categorical distribution over K classes for all pixels in an image. We follow best practice and optimize it with the weighted cross-entropy loss

$$\mathcal{L}_{\text{sem}} = -\frac{1}{|\Omega|} \sum_{p \in \Omega} \omega_k \boldsymbol{t}_p^\top \log\left(\sigma(\boldsymbol{f}_p)\right), \tag{2}$$

where ω_k is a class-wise weight computed via the inverse frequency of each class in the dataset, $\boldsymbol{t} \in \mathbb{R}^{H \times W \times K}$ is a one-hot encoded pixel-wise ground truth label, $\boldsymbol{t}_p \in \mathbb{R}^K$ is a one-hot encoded pixel-wise ground truth label at pixel $p \in \Omega$, σ indicates the softmax operation, and $\boldsymbol{f}_p$ denotes the pre-softmax feature predicted for pixel p.

As mentioned above, we do not only want to perform standard semantic segmentation but also build a class descriptor to bring all pixels belonging to a certain class towards a certain region in the feature space. To achieve this, we accumulate the pre-softmax features, also called activation vectors, of all true positives for each class, where a true positive is a pixel that is correctly segmented. With this, we can store a running average class prototype, or mean activation vector, $\boldsymbol{\mu}_k \in \mathbb{R}^K$ for each class $k \in \{1, \ldots, K\}$:

$$\boldsymbol{\mu}_k = \frac{1}{|\hat{\Omega}_k|} \sum_{p \in \hat{\Omega}_k} \boldsymbol{f}_p. \tag{3}$$

We also iteratively compute the per-class variance $\boldsymbol{\sigma}_k^2 \in \mathbb{R}^K$ via sum of squares, as

$$\boldsymbol{\sigma}_k^2 = \frac{1}{|\hat{\Omega}_k|} \sum_{p \in \hat{\Omega}_k} \left(\boldsymbol{f}_p - \boldsymbol{\mu}_k\right)^2. \tag{4}$$

At the beginning of epoch e, we have the means $\boldsymbol{\mu}_k^{e-1}$ and variances $\boldsymbol{\sigma}_k^{e-1}$ accumulated in the previous epoch. At epoch e, we can steer the semantic segmentation to predict, for each pixel with ground truth class k, a feature vector equal to $\boldsymbol{\mu}_k^{e-1}$. For this, we introduce a feature loss function

$$\mathcal{L}_{\text{feat}} = \frac{1}{|\Omega|} \sum_{k=1}^{K} \sum_{p \in \Omega_k} \frac{\|\boldsymbol{f}_p - \boldsymbol{\mu}_k^{e-1}\|}{\boldsymbol{\sigma}_k^{e-1}}. \tag{5}$$

This loss function is not active during the first epoch since there is no accumulated mean yet. Thus, we perform standard semantic segmentation during the first epoch.

The semantic decoder is thus optimized with a weighted sum of the loss functions introduced above

$$\mathcal{L}_{\text{sdec}} = w_1 \, \mathcal{L}_{\text{sem}} + w_2 \, \mathcal{L}_{\text{feat}}. \tag{6}$$

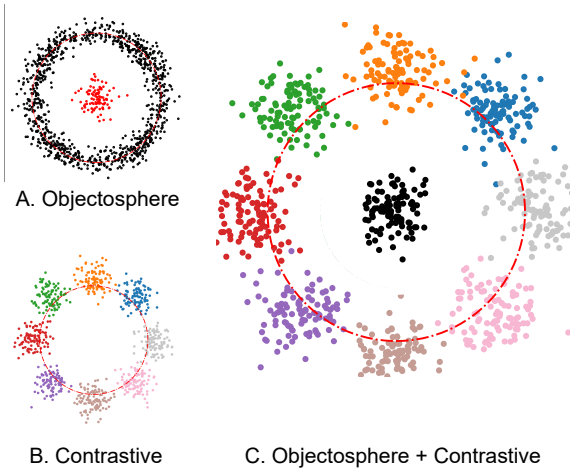


Figure 3. 2D visualization of the expected output of the contrastive decoder. The behavior of the objectosphere loss is shown in A, where all points coming from known classes (black) lie around the red (outer) circle of radius ξ, see Eq. (9), and the points from unknown classes lie around the origin. The contrastive loss is shown in B, where features lie on the unit circle. Together, they lead to a behavior similar to the one depicted in C.

Contrastive Decoder. The contrastive decoder explicitly aims for anomaly segmentation. Given an image of dimensions $H \times W$, where known and unknown classes are present, the goal of the contrastive decoder is to provide the basis for a binary prediction where 0 corresponds to known classes and 1 to unknown classes. We achieve this by means of a combination between the contrastive loss [11] and the objectosphere loss [15]. First, we compute the mean feature representation $\overline{\boldsymbol{f}}_k$ for class k in the current image as

$$\overline{\boldsymbol{f}}_k = \frac{1}{|\Omega_k|} \sum_{p \in \Omega_k} \boldsymbol{f}_p^d, \tag{7}$$

where $\boldsymbol{f}_p^d$ is the feature predicted at pixel p from the contrastive decoder (the equivalent of $\boldsymbol{f}_p$ for the semantic one). Then, we compute the contrastive loss $\mathcal{L}_{\text{cont}}$ such that $\overline{\boldsymbol{f}}_k$ approximates the normalized mean representation $\bar{\boldsymbol{\mu}}_k^{e-1}$ of the corresponding class in the previous epoch $\boldsymbol{\mu}_k^{e-1}$ and gets dissimilar from the other classes mean representation:

$$\mathcal{L}_{\text{cont}} = -\sum_{k=1}^{K} \log \frac{\exp\left(\overline{\boldsymbol{f}}_k{}^\top \bar{\boldsymbol{\mu}}_k^{e-1}/\tau\right)}{\sum_{i=1}^{K} \exp\left(\overline{\boldsymbol{f}}_k{}^\top \bar{\boldsymbol{\mu}}_i^{e-1}/\tau\right)}, \tag{8}$$

where τ is a temperature parameter. This way, the loss aims to make the features from the same class consistent with its running mean representation $\boldsymbol{\mu}_k^{e-1}$, while scattering all K classes around the unit hypersphere.

At the same time, we use the objectosphere loss $\mathcal{L}_{\text{obj}}$

over each pixel $p \in \Omega$ given by

$$\mathcal{L}_{\text{obj}} = \begin{cases} \max\left(\xi - \|\boldsymbol{f}_p\|^2, 0\right) & \text{, if } p \in \mathcal{D}_k \\ \|\boldsymbol{f}_p\|^2 & \text{, otherwise} \end{cases}, \tag{9}$$

where $\mathcal{D}_k$ is the set of pixels belonging to known classes. The remaining pixels, at training time, reduce to the unlabeled (void) areas of the image. This aims to make the norm of the feature vector $\|\boldsymbol{f}_p\|$ of pixels belonging to known classes $\mathcal{D}_k$ bigger than a threshold ξ, while the norm of the features of pixels belonging to unknown classes $\mathcal{D}_u$ gets reduced to 0. These two loss functions $\mathcal{L}_{\text{cont}}$ and $\mathcal{L}_{\text{obj}}$ together allows us to optimize towards a situation where the feature vectors of known classes are distributed along the surface of the K-dimensional hypersphere of radius ξ, while the feature vectors of unknown classes gets squashed to 0. A 2D example of the expected behavior is shown in Fig. 3.

The contrastive decoder is optimized with a weighted sum of the two losses given by

$$\mathcal{L}_{\text{cdec}} = w_3\,\mathcal{L}_{\text{cont}} + w_4\,\mathcal{L}_{\text{obj}}. \tag{10}$$

Post-Processing for Anomaly Segmentation. To obtain the open-world predictions at test time, we fuse the outputs of the two decoders. The semantic encoder provides a standard closed-world semantic segmentation but, thanks to the loss function that operates directly on the feature space that we introduced, we can obtain an open-world segmentation. In fact, we computed mean $\boldsymbol{\mu}_k \in \mathbb{R}^K$ and variance $\boldsymbol{\sigma}_k^2 \in \mathbb{R}^K$ of each class, meaning that, for each class, we can easily build a multi-variate normal distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k$ is the mean, and $\boldsymbol{\Sigma}_k = \text{diag}(\boldsymbol{\sigma}_k^2)$ is the covariance matrix, which reduces to the diagonalization of the variance $\boldsymbol{\sigma}_k^2$ under the assumption that all classes are independent. After building the Gaussian model of each class in the dataset, given a pixel p whose predicted feature $\boldsymbol{f}_p$ would correspond to class $k, \forall k$, we compute a fitting score by means of the squared exponential kernel

$$s_k(\boldsymbol{f}_p) = \exp\left(-\frac{1}{2}(\boldsymbol{f}_p - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{f}_p - \boldsymbol{\mu}_k)\right). \tag{11}$$

Then, for each pixel, we take the highest score

$$s(p) = \max_k s_k(\boldsymbol{f}_p), \tag{12}$$

and, if it is low, then the pixel of interest is in the tail of the Gaussian, and is considered as a novel class, leading to an open-world prediction $\mathcal{U}_{\text{sem}}$ of the semantic decoder. We can obtain a pixel-wise score $s^{\text{sem}}_{\text{unk},\,p}$ for being unknown $s^{\text{sem}}_{\text{unk},\,p} = 1 - s(p)$.

The contrastive decoder leads to a second open-world prediction $\mathcal{U}_{\text{cont}}$ by considering as unknown all pixels whose feature norm is below a certain threshold. In particular, we can obtain a pixel-wise score $s^{\text{cont}}_{\text{unk},\,p}$ for being unknown

$$s^{\text{cont}}_{\text{unk},\,p} = \max\left(0, \left(1 - \frac{\|\boldsymbol{f}_p\|^2}{\xi}\right)\right), \tag{13}$$

where $\boldsymbol{f}_p$ is the predicted feature at pixel p. This score is 1 when the norm of the feature vector is 0, and 0 when the norm is bigger than ξ, as described in Eq. (9).

Finally, we fuse the two predictions to obtain a cumulative pixel-wise score for being unknown as

$$s_{\text{unk},\,p} = \frac{1}{2}\left(s^{\text{sem}}_{\text{unk},\,p} + s^{\text{cont}}_{\text{unk},\,p}\right). \tag{14}$$

If $s_{\text{unk},\,p}$ is above a threshold δ, the pixel is considered belonging to an unknown class.

Post-Processing for Open-World Semantic Segmentation. When a pixel is considered unknown, we need to store its activation vector and decide whether it belongs to an already-discovered class or a new one. Given the set of mean activation vectors for G unknown classes discovered so far $\mathcal{F} = \{\boldsymbol{f}_u^1, \ldots, \boldsymbol{f}_u^G\}$, we take the vector $\boldsymbol{f}_u^g$ that minimizes the distance from the querying vector. If the distance between $\boldsymbol{f}_u^g$ and $\boldsymbol{f}_p$ is below a threshold η, then the pixel belongs to this class, and the mean activation vector gets updated, otherwise it creates a new unknown class $\boldsymbol{f}_u^{g+1}$. This allows us to have a virtually unlimited number of novel classes.

3.3. Class Similarity

As a byproduct of the open-world segmentation, our method can also predict the most similar known category for each unknown sample. As explained in Sec. 3.2, it does not suffice for a feature vector to have the highest activation in the k-th spot for being matched to class k. A sample can have the highest activation for a certain class k but its score computed with Eq. (11) is higher for another class $\tilde{k} \neq k$, meaning that the sample is more inside the area of influence of class $\tilde{k}$ despite having a higher activation on class k. As the most similar class, we propose to choose the one that provides the highest score given by $\tilde{k} = \text{argmax}_k\, s_k(\boldsymbol{f}_p)$.

4. Experimental Evaluation

The main focus of this work is an approach for open-world semantic segmentation that also provides a measure of class similarity. We present experiments to show the capabilities of our method. The results of our experiments support our claims, which are: (i) our model achieves state-of-the-art results for anomaly segmentation while performing competitively on the known classes, (ii) our approach can distinguish between different unknown classes, and (iii) our approach can provide a similarity score for each novel class to the known ones.

(a) Input RGB (b) Closed-world prediction (c) Open-world prediction

Figure 4. Results from the validation set of SegmentMeIfYouCan. We show the input RGB overlayed with the ground truth unknown mask (a), the prediction of our closed-world model (b), and the prediction of our approach for open-world segmentation (c). In the open-world prediction, the unknown class is shown in red.

Table 1. **Left**. Comparison between closed-world and open-world model on the known classes of the training datasets. Our OW approach does not harm closed-world semantic segmentation. **Right**. Results from the public leaderboard of the SegmentMeIfYouCan benchmark. We separate methods that use external data, i.e. out of distribution (OoD) data with semantic labels different from the ones in Cityscapes [9], during training. Our approach ranked overall top 1 for FPR95, PPV and mean F1, and top 6 for AUPR and sIoU (fourth and sixth, respectively) on January 31st, 2024.

Approach	mIoU [%] ↑	
	CityScapes	BDDAnomaly
CW	71.1	64.1
OW	70.8	62.8

Approach	OoD	Pixel-Level		Component-Level		
		AUPR [%] ↑	FPR95 [%] ↓	sIoU gt [%] ↑	PPV [%] ↑	mean F1 [%] ↑
DenseHybrid [19]	✓	78.0	9.8	54.2	24.1	31.1
RbA [47]	✓	**94.5**	4.6	**64.9**	47.5	51.9
Maskomaly [1]	✗	93.4	6.9	55.4	51.2	49.9
RbA [47]	✗	86.1	15.9	56.3	41.4	42.0
ContMAV (ours)	✗	90.2	**3.8**	54.5	**61.9**	**63.6**

4.1. Experimental Setup

We use two datasets for validating our method: SegmentMeIfYouCan [9] and BDDAnomaly [24]. Since ground truths are available for the test set of BDDAnomaly, we use it for ablation studies and experiments on class similarity.

We evaluate our methods with the metrics proposed in the SegmentMeIfYouCan public benchmark for pixel-level performance: area under the precision-recall curve (AUPR) and the false positive rate at a true positive rate of 95% (FPR95). For SegmentMeIfYouCan, we report also component-level metrics provided by the benchmark. As explained, our approach is not limited to anomaly segmentation, but performs open-world semantic segmentation. Thus, we also report the mean intersection-over-union (mIoU) on the known classes, to show that our open-world segmentation approach does not underperform on the known classes when compared to the closed-world equivalent (see Tab. 1, left). Finally, we report the mIoU between the newly-discovered classes and their respective highest-overlapping ground truth class to be discovered.

In all tables, we call our method "ContMAV", where "Cont" indicates the contrastive decoder and "MAV" the mean activation vector of the semantic decoder.

Training details and parameters. In all experiments, we use the one-cycle learning rate policy [58] with an initial learning rate of 0.004. We perform random scale, crop, and flip data augmentations, and optimize with Adam [28] for 500 epochs with batch size 8. We set $\xi = 1$, $\delta = 0.6$, $\tau = 0.1$, $\eta = 0.5$, and loss weights $w_1 = 0.9$, $w_2 = 0.1$, $w_3 = 0.5$, and $w_4 = 0.5$. For SegmentMeIfYouCan, we train only on Cityscapes. For BDDAnomaly, we train only on the training set of BDDAnomaly itself.

4.2. Anomaly Segmentation

The first set of experiments shows that our model achieves state-of-the-art results in anomaly segmentation, and thus also supports our first claim. Here, we aim for a binary segmentation between known classes and previously unseen classes. We report results on SegmentMeIfYouCan in Tab. 1, right and BDDAnomaly in Tab. 2. On SegmentMeIfYouCan, our method outperforms all baselines on

Table 2. Anomaly segmentation results on BDDAnomaly.

Approach	AUPR [%] ↑	FPR95 [%] ↓
MaxSoftmax [23]	3.7	24.5
Background [6]	1.1	40.1
MC Dropout [16]	4.3	16.6
Confidence [14]	3.9	24.5
MaxLogit [24]	5.4	14.0
ContMAV (ours)	**96.1**	**6.9**

FPR95 and ranks top 6 on the public leaderboard for AUPR. On the BDD datasets, our method outperforms all baselines on both metrics, providing compelling results for the task of anomaly segmentation. For the BDD datasets, in this experiment, we treat all the unknown categories as the same unknown class, without focusing on the fact they are, originally, separate classes. Our approach shows compelling results for anomaly segmentation, successfully dealing with challenging situations such as the case in which a known and an unknown object are overlapping, see Fig. 5. While SegmentMeIfYouCan is designed specifically for anomaly segmentation, having images where the anomalous objects are prominent, the BDD dataset is more challenging since objects belonging to bicycle or motorcycle can appear in very small areas of the image (see related figures in the supplementary material), making the task of anomaly segmentation more challenging and harder to solve.

4.3. Open-World Semantic Segmentation

The second experiment illustrates that our approach is capable of distinguishing between different unknown classes, rather than only stating whether something is known or unknown. We achieve this thanks to the feature loss function we introduced in Eq. (7). We conduct this experiment on BDDAnomaly since the test set is manually generated excluding images from the training and the validation set and thus the ground truth labels are available. Our approach is able to create multiple unknown classes, as explained in Sec. 3.2. To evaluate it,for each novel class we create we report the mIoU with the ground truth category that overlaps the most to it. We report results for our method together with results we would achieve without the feature loss function. Since this task is uncommon in the literature, we report one baseline approach as a performance lower bound, that uses the background class for the unknowns and performs K-means clustering in the feature space for this class. As a performance upper bound, we report the mIoU of the three classes in closed-world settings on the original BDD100K, where there is no unknown but every class is present at training time. Results are shown in Tab. 3. Our approach outperforms the baseline and provides satisfying results in distinguishing among different classes. Additionally, removing the feature loss function also provides good results for open-world segmentation, outperforming the baseline by a large margin. Thus, this experiment provides support for our second claim.

Table 3. Open-world semantic segmentation results on BDDAnomaly.

Approach	**mIoU** [%] ↑		
	Train	Motorcycle	Bicycle
Background + cluster	0	32.3	32.8
ContMAV (no feat loss)	48.1	53.8	39.9
ContMAV (with feat loss)	**62.4**	**62.2**	**56.8**
Closed-world	72.3	69.3	60.9

Table 4. Class similarity results on BDDAnomaly*.

Approach	**Accuracy** [%] ↑	
	Motorcycle	Train
Baseline	12.5	9.8
ContMAV with MA	39.9	27.6
ContMAV	**58.9**	**49.9**

4.4. Experiments on Class Similarity

The third experiment shows that our approach successfully assigns to each novel class its most similar known category, supporting our third claim. For this experiment, we manually created a lookup table (see supplementary material for further details) in which each class is assigned a ground truth label indicating its most similar category. For this experiment, we used the BDDAnomaly* dataset proposed by Besnier *et al*. [4], that is a modification of BDDAnomaly where only train and motorcycle are unknown (we report anomaly segmentation results on this dataset in the supplement). In the lookup table, the unknown class "motorcycle" is reported as similar to "car", while the unknown class "train" is reported as similar to "truck". We report one baseline that performs semantic segmentation on the known classes and has a stack of linear layers on the pre-softmax features that learns the lookup table. We compare with our same approach but taking the class that has the highest activation as most similar. We report pixel-wise accuracy results in Tab. 4. The results show that the classifier does not generalize well to the unknown classes. Considering only the highest activation is better than the "specialized" classifier, but still it is not a reliable measure of class similarity.

4.5. Ablation Studies

Finally, we provide ablation studies to investigate the contribution of the modules we introduced. We refer to each ablation study in the tables by the letter in the first column.

Anomaly Segmentation. First, we perform an ablation study on the anomaly segmentation pipeline (Tab. 5). We

Table 5. Ablation study on our anomaly segmentation pipeline on BDDAnomaly. $\mathcal{L}_{\text{feat}}$ refers to the feature loss in Eq. (5), and D_{cont} to the contrastive decoder. "PP" indicates the post-processing operation used for obtaining the open-world prediction: "Th" for softmax thresholding, "MA" for maximum activation, D_μ for the minimum distance from the mean activation vector, "Gs" for the Gaussian inference described in Sec. 3.2.

				BDDAnomaly	
	$\mathcal{L}_{\text{feat}}$	D_{cont}	PP	AUPR [%] ↑	FPR95 [%] ↓
A			Th	46.9	93.9
B	✓		Th	76.4	88.6
C		✓	Th	91.8	70.7
D	✓	✓	Th	94.1	54.4
E	✓		MA	75.9	89.9
F	✓	✓	MA	93.9	57.6
G		✓	–	91.8	69.7
H	✓		D_μ	94.2	57.0
I	✓	✓	D_μ	94.8	29.8
J	✓		Gs	94.2	55.8
K	✓	✓	Gs	**96.1**	**6.9**

investigate the contribution of the feature loss $\mathcal{L}_{\text{feat}}$, of the Gaussian post-processing described in Sec. 3.2, and of the contrastive decoder. We ablate different post-processing strategies. The first strategy is a softmax thresholding strategy where we consider a pixel as unknown if it has two or more activations above a threshold. The second strategy is based on the maximum softmax activation only and categorizes a pixel as unknown if its maximum activation is below a certain threshold. These two strategies yield similar performance, which is an expected outcome since they both rely on the standard final output vector. In the table, we can see that the thresholding strategy alone (A) has poor results, and its performance with the feature loss (B) is close to the performance of the maximum activation strategy with feature loss (E). Additionally, we notice how the thresholding without the feature loss but with the contrastive decoder (C) leads to better performance, that is however extremely similar to the one of the contrastive decoder only (G), suggesting that the contrastive decoder alone is better than a softmax thresholding strategy for this task. A further improvement comes from putting together the feature loss and the contrastive decoder, which leads to better results with both thresholding (D) and maximum activation (F). The other two post-processing strategies we employ are based on the output of the feature loss. One takes the minimum distance D_μ of the activation vector from the mean activation vectors we built during training, while the last one is the Gaussian querying. They lead to similar performance when the contrastive decoder is not used (H and J), and yield the top 2 performance when the contrastive decoder is used (I and K). The Gaussian querying provides a further improvement and achieves the best performance for this task.

Table 6. Ablation study on our class similarity approach on BDDAnomaly*. $\mathcal{L}_{\text{feat}}$ refers to the feature loss in Eq. (5), and D_{cont} to the contrastive decoder. "PP" indicates the post-processing operation used for obtaining the open-world prediction: "MA" for maximum activation, D_μ for the minimum distance from the mean activation vector, "Gs" for the Gaussian inference described in Sec. 3.2.

				Accuracy [%] ↑	
	$\mathcal{L}_{\text{feat}}$	D_{cont}	PP	Motorcycle	Train
L	✓		MA	38.4	25.9
M	✓	✓	MA	39.9	27.6
N	✓		D_μ	53.5	41.7
O	✓	✓	D_μ	54.3	42.1
P	✓		Gs	57.8	48.6
Q	✓	✓	Gs	**58.9**	**49.9**

Class Similarity. The second ablation study targets the class similarity (Tab. 6). The presence of the contrastive decoder does not substantially improve the performance, since the class similarity originates from the semantic decoder. Still, numbers when the contrastive decoder is active (M, O, Q) or inactive (L, N, P) are slightly different since the contrastive decoder affects the shared encoder via backpropagation. The performance of class similarity is poor when we rely on the standard maximum activation (L and M), while it improves when it is based on the minimum distance D_μ of the activation vector from the mean activation vectors built during training (N and O). The Gaussian post-processing achieves the best performance for both classes (P and Q), proving the effectiveness of our approach.

5. Conclusions

In this paper, we presented a novel approach for open-world semantic segmentation on RGB images based on a double decoder architecture. Our method manipulates the feature space of the semantic segmentation for identifying novel classes and additionally indicates the known categories that are most similar to the newly discovered ones. We implemented and evaluated our approach on different datasets and provided comparisons with other existing models and supported all claims made in this paper. The experiments suggest that our double-decoder strategy achieves compelling open-world segmentation results. In fact, with our approach, we are able to detect all anomalous regions in an image and distinguish between different novel classes.

Acknowledgments. This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017008 (Harmony).

References

[1] Jan Ackermann, Christos Sakaridis, and Fisher Yu. Maskomaly: Zero-shot mask anomaly segmentation. In *Proc. of British Machine Vision Conference (BMVC)*, 2023. 6

[2] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed Students: Student-Teacher Anomaly Detection With Discriminative Latent Embeddings. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[4] Victor Besnier, Andrei Bursuc, David Picard, and Alexandre Briot. Triggering Failures: Out-Of-Distribution detection by learning from local adversarial attacks in Semantic Segmentation. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2021. 7, 1, 2

[5] Petra Bevandić, Ivan Kreso, Marin Orsić, and Sinisa Segvić. Simultaneous semantic segmentation and outlier detection in presence of domain shift. *Pattern Recognition*, 11824:33–47, 2019. 2

[6] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The Fishyscapes Benchmark: Measuring blind spots in semantic segmentation. *Intl. Journal of Computer Vision (IJCV)*, 129:3119–3135, 2021. 2, 7, 3

[7] Shubhankar Borse, Ying Wang, Yizhe Zhang, and Fatih Porikli. Inverseform: A loss function for structured boundary-aware segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[8] Douglas O. Cardoso, João Gama, and Felipe M.G. França. Weightless neural networks for open set recognition. *Machine Learning*, 106(9-10):1547–1567, 2017. 2

[9] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. In *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2021. 2, 6, 1

[10] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proc. of the Intl. Conf. on Machine Learning (ICML)*, 2020. 3, 4, 1

[12] Thomas A. Ciarfuglia, Ionut M. Motoi, Leonardo Saraceni, Mulham Fawakherji, Alberto Sanfeliu, and Daniele Nardi. Weakly and semi-supervised detection, segmentation and tracking of table grapes with limited and noisy data. *Computers and Electronics in Agriculture*, 205:107624, 2023. 2

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

[14] Terrance DeVries and Graham W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint*, arXiv:1802.04865, 2018. 7

[15] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2018. 2, 3, 4, 1

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing model uncertainty in deep learning. In *Proc. of the Intl. Conf. on Machine Learning (ICML)*, 2016. 2, 7

[17] Ross Girshick. Fast R-CNN. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2015. 1

[18] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[19] Matej Grcić, Petra Bevandić, and Sinisa Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2022. 6

[20] Steve Halligan, Douglas G Altman, and Susan Mallett. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European Radiology*, 25:932–939, 2015.

[21] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[23] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2017. 2, 7, 3

[24] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *Proc. of the Intl. Conf. on Machine Learning (ICML)*, 2022. 6, 7, 1, 2

[25] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[26] Galadrielle Humblot-Renaux, Sergio Escalera, and Thomas B Moeslund. Beyond auroc & co. for evaluating out-of-distribution detection performance. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[27] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[28] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2015. 6

[29] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic Feature Pyramid Networks. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[30] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic Segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[31] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[32] Abhijit Kundu, Xiaoqi Yin, Alireza Fathi, David Ross, Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru. Virtual multi-view fusion for 3d semantic segmentation. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2020. 2

[33] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2017. 2

[34] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2018.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2014. 1

[36] Krzysztof Lis, Sina Honari, Pascal Fua, and Mathieu Salzmann. Detecting road obstacles by erasing them. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024. 2

[37] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future Frame Prediction for Anomaly Detection – A New Baseline. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[38] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2020. 2

[39] Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-Measurable Anomaly Detection. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[40] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[41] Kira Maag and Tobias Riedlinger. Pixel-wise gradient uncertainty for convolutional neural networks applied to out-of-distribution segmentation. *arXiv preprint*, arXiv:2303.06920, 2023. 2

[42] Elias Marks, Matteo Sodano, Federico Magistri, Louis Wiesmann, Dhagash Desai, Rodrigo Marcuzzi, Jens Behley, and Cyrill Stachniss. High precision leaf instance segmentation for phenotyping in point clouds obtained under real field conditions. *IEEE Robotics and Automation Letters (RA-L)*, 2023. 1

[43] Andres Milioto and Cyrill Stachniss. Bonnet: An Open-Source Training and Deployment Framework for Semantic Segmentation in Robotics using CNNs. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2019. 2

[44] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2018. 2

[45] Andres Milioto, Leonard Mandtler, and Cyrill Stachniss. Fast Instance and Semantic Segmentation Exploiting Local Connectivity, Metric Learning, and One-Shot Detection for Robotics. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2019. 2

[46] Noam Mor and Lior Wolf. Confidence prediction for lexicon-free ocr. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2018. 2

[47] Nazir Nayal, Mısra Yavuz, João F. Henriques, and Fatma Güney. Segmenting unknown regions rejected by all. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2023. 6

[48] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[49] Marin Orsić, Ivan Kreso, Petra Bevandić, and Sinisa Segvić. In Defense of Pre-Trained ImageNet Architectures for Real-Time Semantic Segmentation of Road-Driving Images. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. of the Intl. Conf. on Machine Learning (ICML)*, 2021. 2

[51] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-Guided Dense Prediction With Context-Aware Prompting. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. of the Advances in Neural Information Processing Systems (NIPS)*, 2015. 1

[53] Gianmarco Roggiolani, Matteo Sodano, Tiziano Guadagnino, Federico Magistri, Jens Behley, and Cyrill

Stachniss. Hierarchical Approach for Joint Semantic, Plant Instance, and Leaf Instance Segmentation in the Agricultural Domain. *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2023. 2

[54] Eduardo Romera, José M. Alvarez, Luis M. Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Trans. on Intelligent Transportation Systems (T-ITS)*, 19(1):263–272, 2018. 3, 2

[55] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards Total Recall in Industrial Anomaly Detection. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[56] Stuart J. Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010. 1

[57] Hitesh Sapkota and Qi Yu. Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[58] Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 11006:369–386, 2019. 6

[59] Matteo Sodano, Federico Magistri, Tiziano Guadagnino, Jens Behley, and Cyrill Stachniss. Robust Double-Encoder Network for RGB-D Panoptic Segmentation. *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2023. 2

[60] Shengyang Sun and Xiaojin Gong. Hierarchical Semantic Contrast for Scene-Aware Video Anomaly Detection. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[61] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. *Domain Adaptation in Computer Vision Applications*, pages 37–55, 2017. 2

[62] Chin-Chia Tsai, Tsung-Hsuan Wu, and Shang-Hong Lai. Multi-scale patch-based representation learning for image anomaly detection and segmentation. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2022. 2

[63] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2021. 2

[64] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2018. 2

[65] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[66] Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. Openauc: Towards auc-oriented open-set recognition. *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2022.

[67] Jan Weyler, Federico Magistri, Elias Marks, Yue Linn Chong, Matteo Sodano, Gianmarco Roggiolani, Nived Chebrolu, Cyrill Stachniss, and Jens Behley. PhenoBench–A Large Dataset and Benchmarks for Semantic Image Interpretation in the Agricultural Domain. *arXiv preprint*, arXiv:2306.04557, 2023. 2

[68] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2021. 1

[69] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video Event Restoration Based on Keyframes for Video Anomaly Detection. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[70] Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit Boundary Guided Semi-Push-Pull Contrastive Learning for Supervised Anomaly Detection. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[71] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[72] Vitjan Zavrtanik, Matej Kristan, and Danijel Skoćaj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2021. 2

[73] Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting Completeness and Uncertainty of Pseudo Labels for Weakly Supervised Video Anomaly Detection. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[74] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. DeSTSeg: Segmentation Guided Denoising Student-Teacher for Anomaly Detection. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[75] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 2

[76] Ying Zhao. OmniAL: A Unified CNN Framework for Unsupervised Anomaly Localization. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[77] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-Based Language-Image Pretraining. In *Proc. of the*

IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2022. 2

Open-World Semantic Segmentation Including Class Similarity

Supplementary Material

A. Further Details on Experiments

A.1. Datasets and Metrics

We use two datasets for validating our method: SegmentMeIfYouCan [9] and BDDAnomaly [24]. SegmentMeIfYouCan relies on the semantic annotations of Cityscapes [13], and offers a public benchmark with a hidden test set for anomaly segmentation, where the goal is to segment objects that are not present on Cityscapes. Annotations are binary, since each object is either known or unknown. BDDAnomaly is a reorganization of BDD100K [71], where all images containing the classes train, motorcycle and bicycle have been discarded from the training and validation set to create an open-world test set. Since ground truth data is available for this dataset, we use it for ablation studies and experiments on class similarity. Additionally, we report results on a further modification of BDDAnomaly proposed by Besnier *et al.* [4], which we call BDDAnomaly*, where only train and motorcycle are considered as unknown classes. For metrics computation, we used the official evaluation pipeline of SegmentMeIfYouCan to enforce fairness and reproducibility[2]. We decided to not use the area under the ROC curve (AUROC), because recently several papers showed its limitations [20, 26, 66], as two models with the same performance may differ widely in terms of how clearly they separate in-distribution and out-of-distribution samples. In general, these works argue that AUROC is not a fair metric for comparing different approaches. This might be the reason why the official evaluation tool of SegmentMeIfYouCan, which we use in this work, does not report it.

A.2. Experiments on Hyperparameters

Hyperparameters search is usually a challenging problem when it comes to training neural networks. Usually, they are chosen empirically and only the configuration that works best is reported on the paper. In the following, we try to give some insight on our choice of hyperparameters and the reasoning behind them. We provide an analysis on the four hyperparameters (ξ, δ, τ, and η) in the following.

As discussed in Sec. 3.2 of the main paper, in the paragraph dedicated to the contrastive decoder, ξ is the radius of the hypersphere created by the objectosphere loss [15] in Eq. (9). In principle, this hyperparameter could take any value. However, we pair the objectosphere loss to the contrastive loss [11] in Eq. (8), which aims to distribute all feature vectors on the unit sphere. Thus, we expect that any choice of ξ that is different from 1 would harm performance, since it would reduce the synergy between the two loss functions operating on the same decoder. We report an experiment about this in Tab. 7. When $\xi < 1$, the performance is not dramatically harmed because the objectosphere loss aims to make the norm of the features belonging to the known pixels greater than ξ. Thus, the two losses do not work against each other. In contrast, when $\xi > 1$, the two loss functions try to achieve two tasks which are incompatible (features on the unit circle and, at the same time, with norm greater than 1), and performance suffers.

[2] https://github.com/SegmentMeIfYouCan/road-anomaly-benchmark

Table 7. Anomaly segmentation results on BDDAnomaly with different choices of the parameter ξ.

Approach	AUPR [%] ↑	FPR95 [%] ↓
ContMAV ($\xi = 0.75$)	92.2	18.7
ContMAV ($\xi = 1.25$)	83.4	55.2
ContMAV ($\xi = 1$)	**96.1**	**6.9**

Table 8. Anomaly segmentation results on BDDAnomaly with different choices of the parameter δ.

Approach	AUPR [%] ↑	FPR95 [%] ↓
ContMAV ($\delta = 0.4$)	86.6	41.2
ContMAV ($\delta = 0.8$)	89.1	30.1
ContMAV ($\delta = 0.6$)	**96.1**	**6.9**

The threshold δ, which we also introduced in Sec. 3.2, in the paragraph dedicated to the post-processing, is our "unknown-ness threshold". In fact, we obtain a score $s_{\text{unk}} \in [0, 1]$ and have to decide whether a pixel belongs to an unknown category based on this score. The score is given by

$$s_{\text{unk}} = \frac{1}{2}\left(s_{\text{unk}}^{\text{sem}} + s_{\text{unk}}^{\text{cont}}\right), \tag{15}$$

where $s_{\text{unk}}^{\text{seg}}$ and $s_{\text{unk}}^{\text{cont}}$ are the scores coming from the semantic and the contrastive decoders, respectively. Notice that, since the final score is a standard mean of the two, setting the threshold to a low value would make us label a pixel as unknown also in the case in which only one score is high but the other is not. This would create a lot of false positives, and we expect performance aligned with models G and J in Tab. 5 of the main paper. Those two models, in fact, only have one active decoder, and setting a low δ causes a similar behavior. Setting the threshold too high is, in contrast, achievable only when both decoder heads are very confident in their prediction of unknown, and it could cause a high number of false negatives. Thus, we choose $\delta = 0.6$, that is a good compromise and provides good results (see Tab. 8).

We do not optimize the temperature parameter of the contrastive loss τ and perform all experiments with $\tau = 0.1$, as suggested by Chen *et al.* [11].

The hyperparameter η, also introduced in Sec. 3.2, in the paragraph dedicated to the post-processing, does not affect the prediction of a pixel as unknown, but it plays a role in the class discovery. In fact, it represents the minimum distance needed to decide whether a feature categorized as unknown is a class of its own and does not belong to any of the already-discovered new classes. Setting this threshold heavily depends on the data distribution. A

Table 9. Class discovery results on BDDAnomaly with different choices of the parameter η. For each class of interest, the discovered one with greater IoU is chosen and reported.

Approach	mIoU [%] ↑ Train	Motorcycle	Bicycle	N_U
ContMAV ($\eta = 0.3$)	0.0	23.4	0.0	1
ContMAV ($\eta = 0.9$)	30.5	31.1	18.9	12
ContMAV ($\eta = 0.6$)	**62.4**	**62.2**	**56.8**	4

very high threshold would create a lot of classes, and its usefulness would be limited. On the other hand, a low threshold would put all classes together, providing nothing more than an anomaly segmentation. We report results in Tab. 9, where we also report the number N_U of new classes created, for which the ground truth value is 3 (i.e., the number of unknown classes in BDDAnomaly).

A.3. Further Details on Anomaly Segmentation

In Sec. 4 of the main paper, we reported extensive experiments on SegmentMeIfYouCan [9] and BDDAnomaly [24]. SegmentMeIfYouCan is a public benchmark for anomaly segmentation, with a hidden test set and a public leaderboard. Our method, called ContMAV, ranks first overall on three out of five metrics, namely FPR95, PPV and mean F1, and it ranks fourth on AUPR and sixth on sIoU. Further details and baselines results can be found on SegmentMeIfYouCan's official website: `https://segmentmeifyoucan.com/leaderboard`.

Besnier *et al.* [4] proposed a modification of BDDAnomaly [24], where only two classes (train and motorcycle) are considered as unknown. We call this dataset BDDAnomaly*. Differently from BDDAnomaly, the images containing bicycle are not discarded from the training and validation set, but are kept and bicycle is considered a known class. We test our method also on this dataset, using the same training details and parameters discussed above. We report our results in Tab. 10.

A.4. Further Details on Class Similarity

In Sec. 4.3 of the main paper, we reported our experiment on class similarity, and mentioned the creation of a lookup table in which each class is assigned a ground truth label indicating its most similar category. We chose the most similar class based on the relevance in an autonomous driving scenario. For example, truck is paired to bus since one could expect a similar behavior between these two traffic participants. Some classes, such as sky, are not assigned any label for the most similar category. The lookup table is reported in Tab. 11. For this experiment, we decided to use BDDAnomaly* because we did not find a valid correspondence for the class bicycle. The only vehicle that belongs to known classes is car, and in fact our method on BDDAnomaly achieves, for bicycle, a 43.2% similarity score with car. A more modern dataset, with more vehicle classes such as electric scooters, would provide better candidates for class similarity.

Table 10. Anomaly segmentation results on BDDAnomaly*.

Approach	AUPR [%] ↑	FPR95 [%] ↓
MaxSoftmax [23]	80.1	63.5
Background [6]	75.3	68.1
MC Dropout [16]	82.6	61.1
ODIN [34]	81.7	60.6
ObsNet + LAA [4]	82.8	60.3
ContMAV (ours)	**92.9**	**43.9**

Table 11. Look-up table for class similarity. The unknowns are specified in the context of BDDAnomaly*.

Category	Most Similar	Type
Road	Sidewalk	stuff
Sidewalk	Road	stuff
Building	Wall	stuff
Wall	Fence	stuff
Fence	Wall	stuff
Pole	Sign	stuff
Light	Sign	stuff
Vegetation	Terrain	stuff
Terrain	Vegetation	stuff
Sky	–	stuff
Person	Rider	thing
Rider	Person	thing
Car	Truck	thing
Truck	Bus	thing
Bus	Truck	thing
Bicycle	–	thing
Train	Truck	thing, unknown
Motorcycle	Car	thing, unknown

A.5. Architectural Choices

As reported in Sec. 3.1, we used a modified version of ResNet34. Still, our contribution does not include any of the modules presented there, such as the NonBottleneck-1D block or the average pyramid pooling module, whose contributions are reported in the related papers [54, 75]. Therefore, we do not provide ablation studies on these components, but rather all of our models and ablations use them.

B. Further Details on the Contrastive Decoder

The contrastive decoder, which we explain in details in Sec. 3.2, is optimized with a combination of two loss functions, namely the objectosphere and the contrastive loss. Fig. 3 intuitively shows the idea behind it, and what the ideal output in the 2D case would be. However, the

Table 12. Architectural Efficiency

Approach	GFLOPs ↓	Training Parameters ↓
Maskomaly [23]	937	215M
Mask2Anomaly [6]	258	**23M**
ContMAV (ours)	**84**	48M

feature vectors that the contrastive decoder predicts are K-dimensional, where K is the number of known classes (*i.e.*, 19 in our case). In order to verify whether the output of the decoder is aligned with our expectation, we define two thresholds ζ and ρ. Then, given $\boldsymbol{f}_p^d$, *i.e.*, the feature predicted at pixel p from the contrastive decoder, we want $1-\zeta < ||\boldsymbol{f}_p^d||_2 < 1+\zeta$ for all $\boldsymbol{f}_p^d$ whose ground truth label is a known class, and $||\boldsymbol{f}_p^d||_2 < \rho$ for all $\boldsymbol{f}_p^d$ whose ground truth label is an unknown class. The former means that the norms of the vectors belonging to known classes should be in a "tube" of radius ζ around 1, which is our ξ parameter as explained in Tab. 7. The latter means that the norms of the vectors belonging to unknown classes (which, at training time, are the unlabeled portions of the image), should be smaller than ρ. We choose $\zeta = 0.2$ and $\rho = 0.4$, and we find that 86.5% of the vectors belonging to known classes fall into the tube, and that 79.9% of the vectors belonging to unknown classes are smaller than ρ. This verifies that the output is aligned to our expectation. To visually show the result, we would need to apply a dimensionality reduction approach such as principal component analysis. However, linear dimensionality reduction techniques always lead to loss of information, and the new dimensions may offer no concrete interpretability.

C. Architectural Efficiency

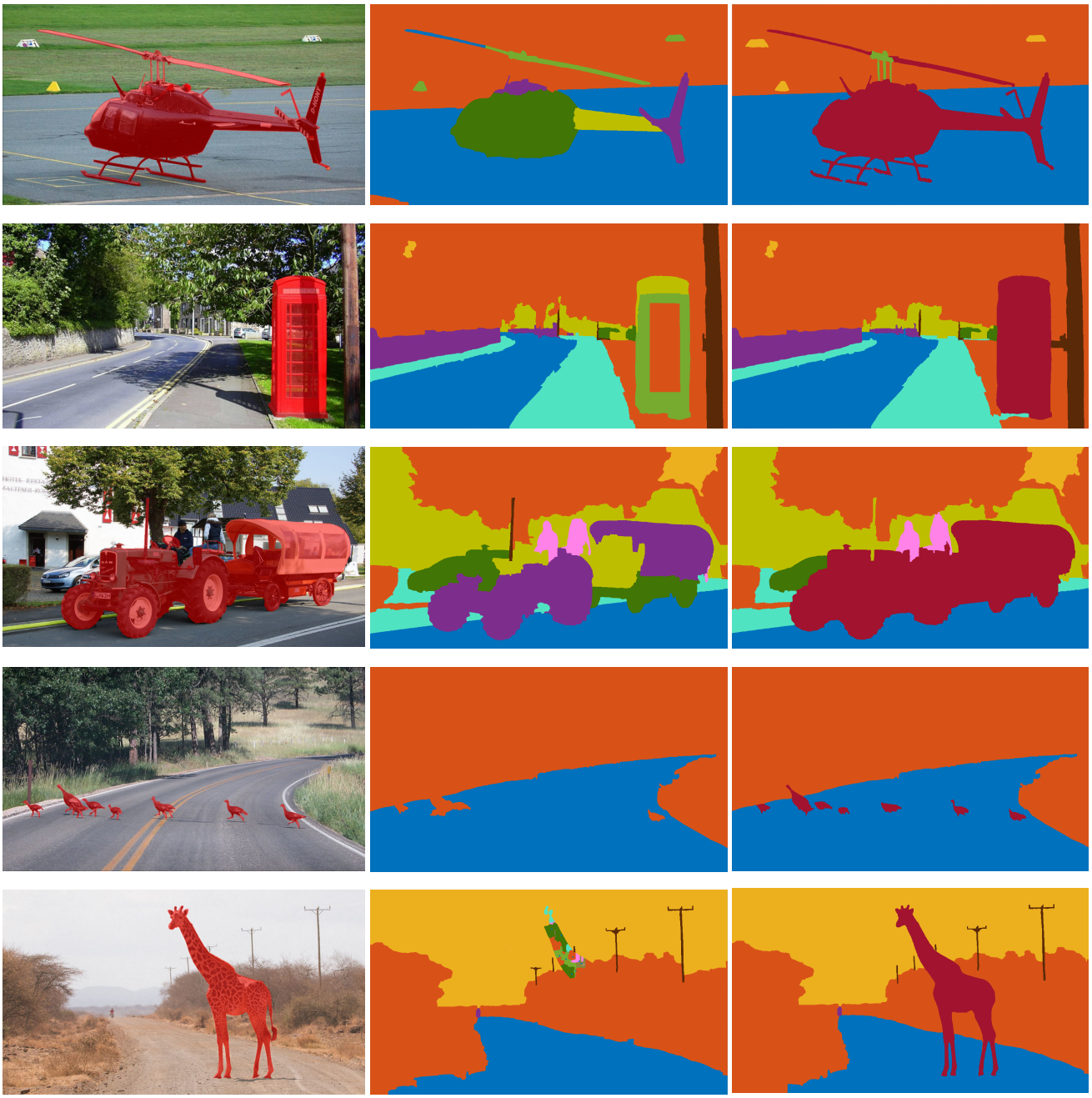
As pointed out in Sec. 3.1, we designed our neural network in order to be lightweight and faster at inference time. The architecture design choices explained in Sec. 3.1 allow inference on an image at 10 Hz. Additionally, we report the number of parameters and the GFLOPs of our model together with two state-of-the-art models from the SegmentMeIfYouCan public benchmark with code available in Tab. 12. We show that our architecture is competitive and performs very well in terms of efficiency.

D. Qualitative Results

We provide further qualitative results on the validation set of SegmentMeIfYouCan and the test set of BDDAnomaly in Fig. 5 and Fig. 6, respectively. Additionally, we report qualitative results on the test set of BDDAnomaly for class similarity in Fig. 7.

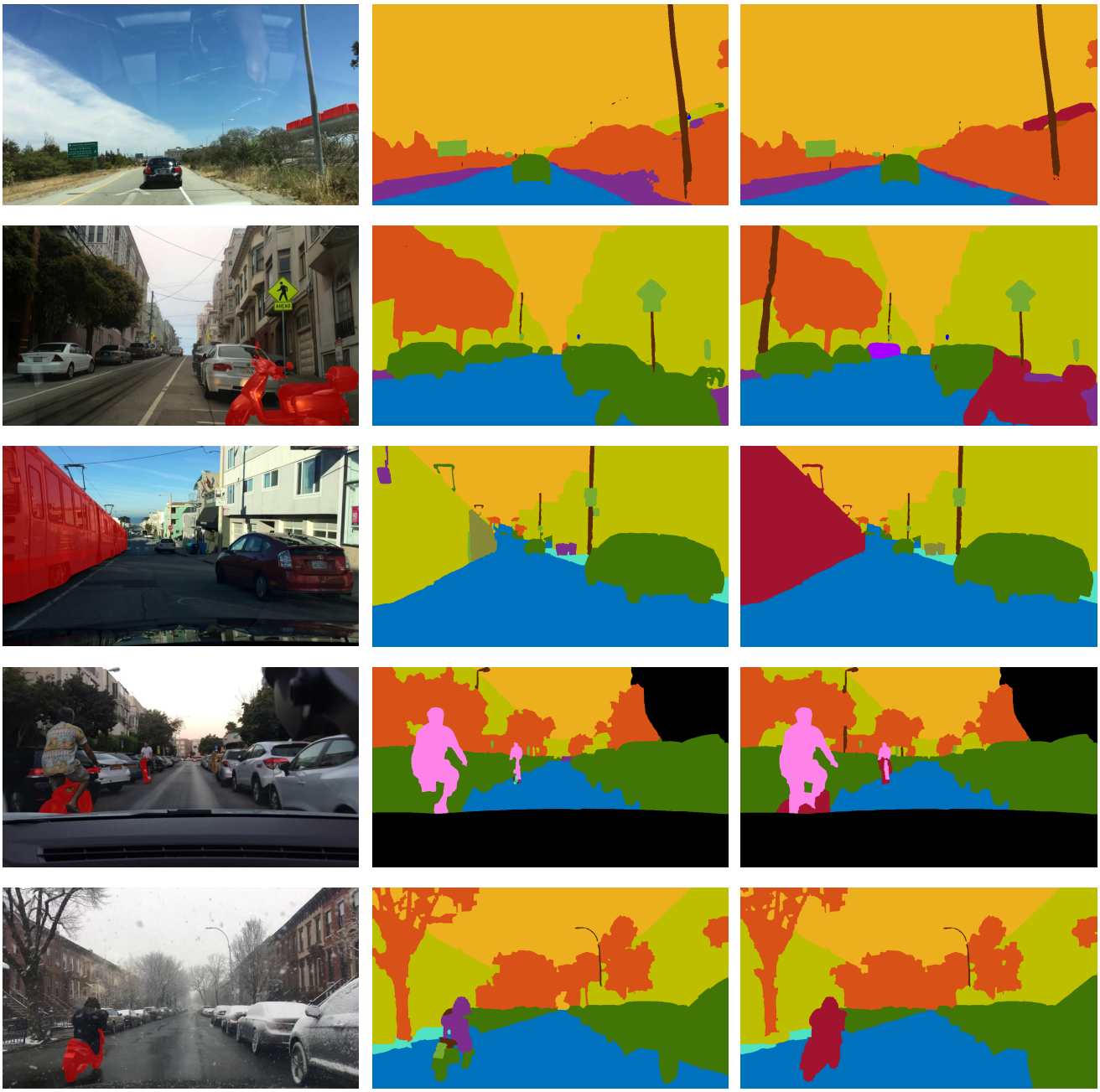
E. Limitations and Future Works

As shown in the various experiments, our approach achieves state-of-the-art results on different datasets on both, anomaly segmentation and novel class discovery. Still, our approach presents some limitations which offer interesting avenues for future work in order to make the approach more robust and performing. In particular, the semantic decoder builds a mean activation vector, or average class prototype, for each class and the dimension of this descriptor is equal to the number of known classes. When not many classes are available at training time, this descriptor collapses to a few dimensions, which might be not descriptive enough for ensuring a reliable novel class discovery where many new classes can be found. The contrastive decoder instead leverages the unlabeled portions of the image as unknowns available at training time to train the objectosphere loss (basically following the concept of "known unknowns" introuced by Bendale *et al*. [2]), and would suffer from a fully labeled dataset where no pixel is left with no ground truth annotation. Additionally, we provide open-world semantic segmentation (*i.e.*, anomaly segmentation and novel class discovery) only, but no instance are segmented. An interesting research avenue is to extend this work in the direction of open-world panoptic segmentation.



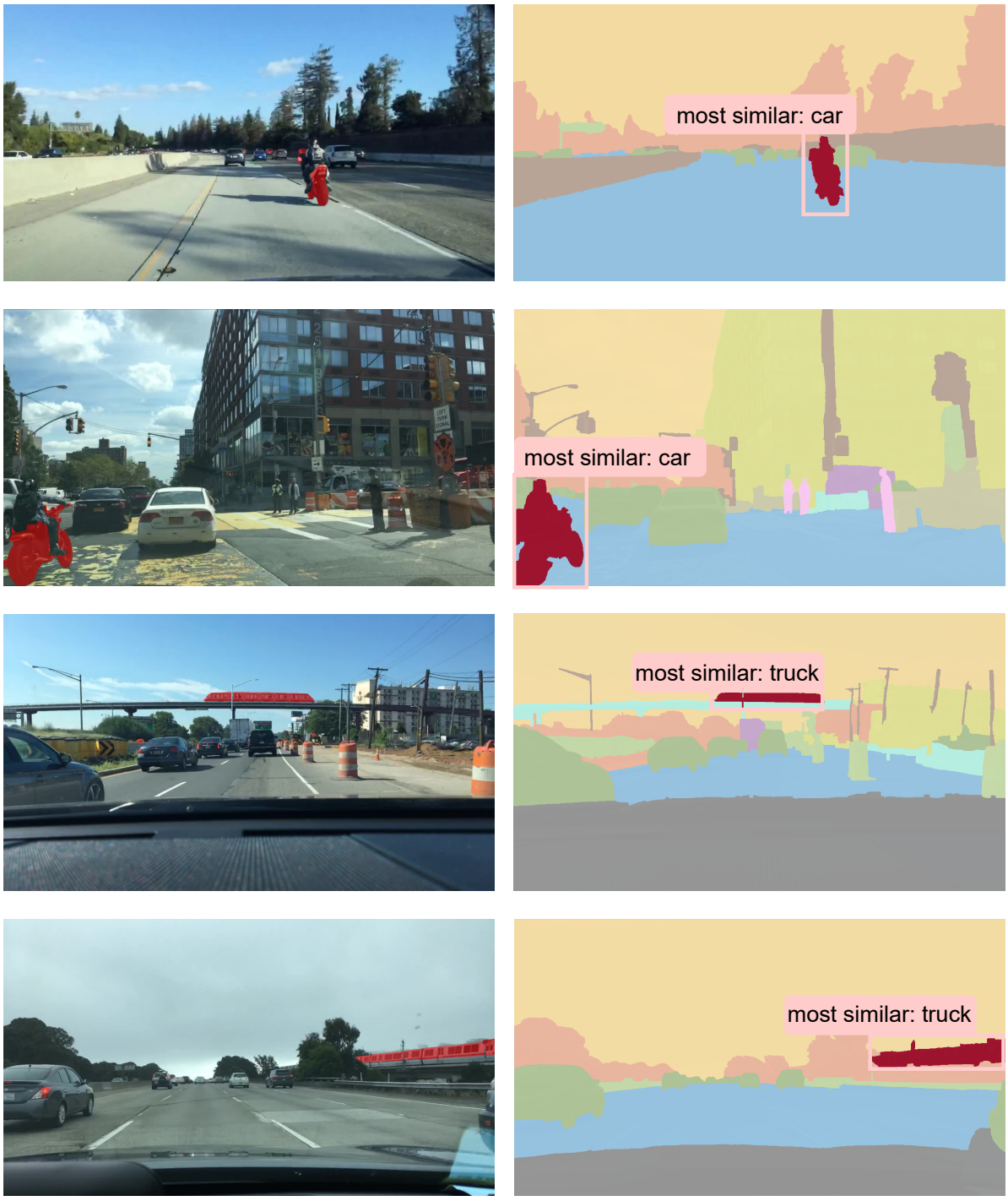
(a) Input RGB (b) Closed-world prediction (c) Open-world prediction

Figure 5. Anomaly segmentation results from the validation set of SegmentMeIfYouCan. We show the input RGB overlayed with the ground truth unknown mask (a), the prediction of our closed-world model (b), and the prediction of our approach for open-world segmentation (c). In the open-world prediction, the unknown class is shown in red. Notice how the two models, that are both trained on CityScapes, perform similarly on known classes, demonstrating that our approach does not degrade closed-world performance.



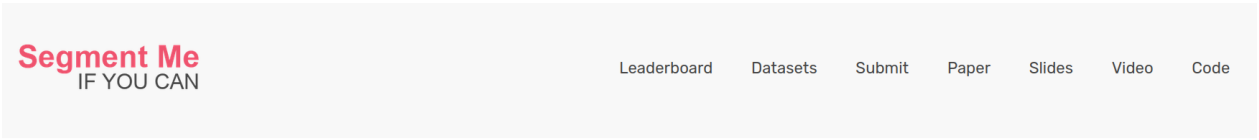
(a) Input RGB (b) Closed-world prediction (c) Open-world prediction

Figure 6. Anomaly segmentation results from the test set of BDDAnomaly. We show the input RGB overlayed with the ground truth unknown mask (a), the prediction of our closed-world model (b), and the prediction of our approach for open-world segmentation (c). In the open-world prediction, the unknown class is shown in red. Notice how the two models, that are both trained on BDDAnomaly, perform similarly on known classes, demonstrating that our approach does not degrade closed-world performance.



(a) Input RGB (b) Prediction of Class Similarity

Figure 7. Class similarity results from the test set of BDDAnomaly. We show the input RGB overlayed with the ground truth unknown mask (a) and the prediction of our class similarity pipeline (b). In the open-world prediction, the unknown class is shown in red, and the overall semantic segmentation is shown in transparency.



LEADERBOARD

Evaluation Metrics

- **AUPR :** pixel-wise Area Under Precision Recall curve
- **FPR_{95} :** pixel-wise False Positive Rate at a true positive rate of 95%
- **sIoU gt :** adjusted Intersection over Union averaged over all ground truth segmentation components
- **PPV :** predictive positive value (or precision) averaged over all predicted segmentation components
- **mean F1 :** component-wise F1-score averaged over different detection thresholds

For a more detailed explanation of the metrics, we refer to our paper.

Anomaly Track

Method	OoD Data	Pixel Level		Component Level		
		AUPR	FPR_{95}	sIoU gt	PPV	mean F1
ContMAV		90.20%	3.83%	54.55%	61.86%	63.64%
EAM [paper]	✓	93.75%	4.09%	67.09%	53.77%	60.86%
RbA [paper]	✓	94.46%	4.60%	64.93%	47.51%	51.87%
Maskomaly [paper] [code]	✗	93.35%	6.87%	55.43%	51.46%	49.90%
CSL	✗	80.08%	7.16%	46.46%	50.02%	50.39%
DenseHybrid [paper] [code]	✓	77.96%	9.81%	54.17%	24.13%	31.08%
cDNP [paper] [code]	✗	88.90%	11.42%	50.44%	29.04%	28.12%
RPL+CoroCL [paper] [code]	✓	83.49%	11.68%	49.77%	29.96%	30.16%
Mask2Anomaly [paper] [code]	✓	88.72%	14.63%	55.28%	51.68%	47.16%
Maximized Entropy [paper] [code]	✓	85.47%	15.00%	49.21%	39.51%	28.72%
RbA [paper]	✗	86.13%	15.94%	56.26%	41.35%	42.04%
Image Resynthesis [paper] [code]	✗	52.28%	25.93%	39.68%	10.95%	12.51%

Figure 8. Screenshot of the top methods in the public leaderboard of SegmentMeIfYouCan, taken on November 21st 2023. Our method, ContMAV, is the top approach for FPR95, PPV, and mean F1. To preserve anonymity, paper and code will be attached to the benchmark submission upon acceptance.

Certificate of Reproducibility

The authors of this publication declare that:

1. The software related to this publication is distributed in the hope that it will be useful, support open research, and simplify the reproducability of the results but it comes without any warrenty and without even the implied warranty of merchantability or fitness for a particular purpose.

2. *Matteo Sodano* primarily developed the implementation related to this paper. This was done on Ubuntu 20.04.

3. *Federico Magistri* verified that the code can be executed on a machine that follows the software specification given in the Git repository available at:

   `https://github.com/PRBonn/ContMAV`

4. *Federico Magistri* verified that the experimental results presented in this publication can be reproduced using the implementation used at submission, which is labeled with a tag in the Git repository and can be retrieved using the command:

   ```
   git checkout CVPR
   ```