

Matteo Sodano¹ and Federico Magistri¹ and Jens Behley¹ and Cyrill Stachniss^{1,2}

Abstract

Robots need to be able to understand their surroundings in order to operate safely and robustly, and to interact with the surrounding environment. Robots deployed in unconstrained real-world scenarios must additionally be able to deal with novel situations and objects that have never been seen before. In this article, we tackle the problem of open-world panoptic segmentation, i.e., the task of discovering new semantic categories and new object instances at test time, while enforcing consistency among the categories that we incrementally discover. We present Con2MAV, a general method for open-world panoptic segmentation. Experiments across a wide range of datasets, from road scenes to underwater environments, highlight its compelling capabilities in open-world segmentation and its competitive performance on known classes. We will open-source the implementation of our approach upon acceptance. In addition, we propose PANIC (**P**anoptic **A**nomalies **I**n **C**ontext), a benchmark for evaluating open-world segmentation tasks in autonomous driving scenarios. This dataset, recorded with a multi-modal sensor suite mounted on a car, and then manually annotated, provides high-quality, pixel-wise annotations of anomalous objects at both semantic and instance level. PANIC contains 800 images, more than 50 unknown classes, i.e., classes that do not appear in the training set, and over 4,000 object instances, providing a comprehensive benchmark for evaluating open-world segmentation methods in autonomous driving scenarios. We provide competitions for multiple open-world segmentation tasks on a hidden test set. Our dataset and competitions are available at <https://www.ipb.uni-bonn.de/data/panic>.

Keywords

open-world segmentation, semantic scene understanding, datasets and benchmarks

1 Introduction

Interpreting sensor data is a key capability needed by robots operating in unconstrained real-world scenarios. Robots, in fact, must be able to understand their surroundings to operate safely and robustly. For this reason, we witnessed enormous progress in scene interpretation tasks such as object detection (Girshick 2015; Ren et al. 2016), semantic segmentation (Milioto and Stachniss 2019; Oliveira et al. 2018), instance segmentation (He et al. 2017; Marks et al. 2023), and panoptic segmentation (Cheng et al. 2022; Kirillov et al. 2019a; Roggiolani et al. 2023; Sodano et al. 2023). However, these tasks in their conventional setting share the limiting assumption that all classes and objects that appear at inference time have already been seen at training time. This so-called *closed-world* assumption greatly limits the deployment of such algorithms in the real world. A central challenge for learning-based systems is the ability to deal with something that they have never seen before at training time. For example, autonomous cars navigating in the real world will eventually experience situations and objects that they have not seen before, and need to be able to identify them in order to operate safely and not cause any harm. Thus, autonomous vision systems should be able to deal with the *open-world* scenario, i.e., the fact that not everything can be covered in the training data and novel objects and categories should be identified at test time.

Recently, deep learning-based methods have achieved remarkable performance in semantic scene understanding (Cheng et al. 2020, 2022; Tang et al. 2021). However,

they typically operate under a closed-world assumption, predicting only within a fixed set of categories defined during training. As a result, they cannot recognize objects that fall outside this predefined taxonomy and often misclassify unknown instances with high confidence. Open-vocabulary, also called zero-shot segmentation, approaches (Carion et al. 2025; Ghiasi et al. 2022) attempt to relax this constraint by leveraging text embeddings from large-scale vision-language models. However, they still require all relevant categories to be anticipated and explicitly included in the vocabulary via prompts before segmenting an image. This fundamentally limits their applicability in dynamic, unpredictable environments where novel objects may appear without any prior semantic description.

Open-world panoptic segmentation enables a level of scene understanding that goes beyond both standard panoptic segmentation and open-vocabulary approaches. Unlike closed-world panoptic segmentation, which is restricted to a fixed set of predefined classes, the open-world setting requires the model to discover and segment previously unseen categories and their individual object instances at test time. And unlike open-vocabulary panoptic segmentation,

¹Center for Robotics, University of Bonn, Germany

²Lamarr Institute for Machine Learning and Artificial Intelligence, Germany

Corresponding author:

Matteo Sodano, Center for Robotics, University of Bonn, Germany.
Email: matteo.sodano7@gmail.com

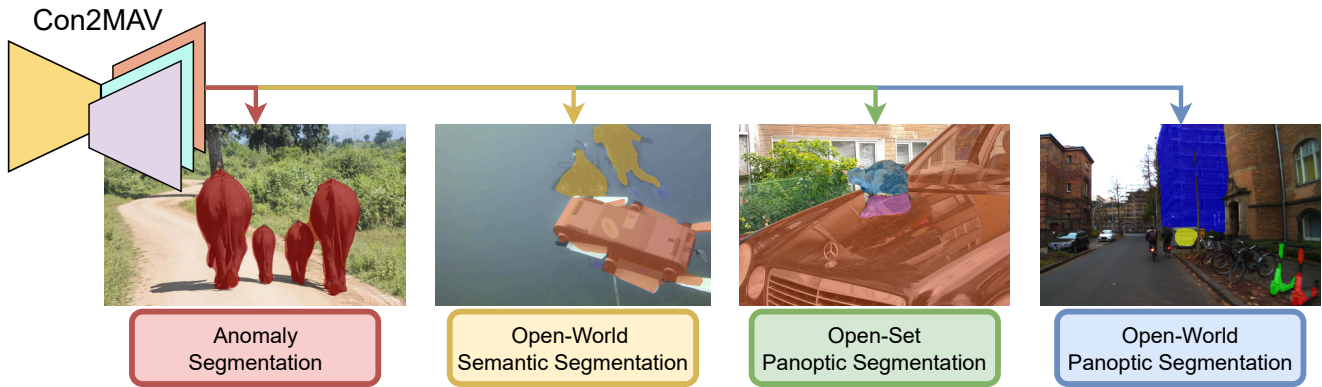


Figure 1. Our proposed approach, Con2MAV, is able to tackle multiple open-world tasks and segment unknown objects and categories. In the figure, we show predictions on (left to right) SegmentMelfYouCan (Chan et al. 2021), SUIM (Islam et al. 2020), COCO (Lin et al. 2014), and PANIC (ours).

which can only recognize categories that are explicitly provided through text prompts or belong to the vocabulary of a vision-language model, open-world panoptic segmentation makes no assumptions about which and how many classes may appear. This capability is essential for real-world autonomous systems, which must operate robustly in unpredictable, dynamic, and open environments where new objects or categories emerge at any time.

Thus, in this article, we address open-world panoptic segmentation. We introduce Con2MAV, a fully convolutional method that leverages a new loss function to construct robust pre-logit class descriptors, enabling meaningful representations even when only few training categories are available. Beyond segmenting unknown objects, Con2MAV incrementally discovers novel semantic categories at test time and maintains consistent category identities across images through an online descriptor update mechanism. Our formulation is architecture-agnostic and can be applied to any backbone. Con2MAV achieves compelling performance on anomaly segmentation, open-world semantic segmentation, and open-world panoptic segmentation across several benchmarks, including SegmentMelfYouCan (Chan et al. 2021), COCO (Lin et al. 2014), BDDAnomaly (Hendrycks et al. 2022), and SUIM (Islam et al. 2020). Illustrative predictions are shown in Fig. 1.

Additionally, motivated by recent advances in open-world segmentation (Delić et al. 2024; Grcić et al. 2023; Nayal et al. 2023; Rai et al. 2024; Sodano et al. 2024), we introduce the PANIC (Panoptic Anomalies In Context) benchmark. Recorded in Bonn, Germany, with our sensor suite (Vizzo et al. 2023) mounted on a car, PANIC includes 800 images with semantic and instance labels for all anomalous objects. We consider objects that do not appear in Cityscapes (Cordts et al. 2016) as anomalies. They can appear anywhere in the image and have any size. We split the dataset in validation and test images, and no unknown class that appears in the test set is present in the validation set, making the test set truly hidden. Examples from the dataset are shown in Fig. 2.

Lastly, we provide an extended discussion on open-world segmentation tasks. Beside anomaly segmentation, which is quite common in the literature, other open-world tasks suffer from an ambiguous task nomenclature, and it is not rare to encounter different works addressing the same task but using a different name. Here, we try to overcome this limitation

and aim to unify the nomenclature for open-world tasks, hoping that this would make open-world segmentation more common and accessible in the future.

In summary, our contributions are threefold. First, we propose a clear and unified nomenclature for open-world segmentation, resolving the inconsistencies that arise when different works use different terms to refer to the same task. Second, we introduce a fully convolutional neural network that, through a novel pre-logit loss formulation, achieves compelling performance on open-world segmentation tasks while retaining strong closed-world performance. In contrast to prior segmentation-based category discovery methods that operate offline, our approach incrementally discovers novel categories at test time and maintains cross-image category consistency through online updates of semantic descriptors. Third, we present PANIC, a dataset of 800 autonomous-driving images with pixel-wise annotations covering more than 50 anomalous categories and over 4000 object instances, accompanied by public challenges for multiple open-world segmentation tasks.

2 Related Work

Recently, semantic and panoptic segmentation have been among the most prominent perception tasks in different domains, such as autonomous driving (Borse et al. 2021; Roynard et al. 2018; Wang et al. 2023), indoor navigation (Hu et al. 2021; Kundu et al. 2020; Schwarz et al. 2018; Sodano et al. 2023), and agricultural robotics (Guclu et al. 2025; Halstead et al. 2024; Marks et al. 2023; Roggiolani et al. 2023). However, they mostly operate in closed-world settings, while the open-world domain has been relatively under-explored.

Open-World Detection and Segmentation. The first approaches that relaxed the closed-world assumption targeted anomaly detection and open-set recognition. Their primary goal was to identify and reject previously unseen samples. Early methods adapted standard closed-world segmentation pipelines by introducing dedicated anomaly classes (Blum et al. 2021; Mor and Wolf 2018; Vaze et al. 2021), modifying or thresholding the softmax output (Bendale and Boulton 2016; Cardoso et al. 2017; Hendrycks and Gimpel 2017), or using model ensembles (Ji et al. 2025; Lakshminarayanan et al. 2017; Vyas et al. 2018).

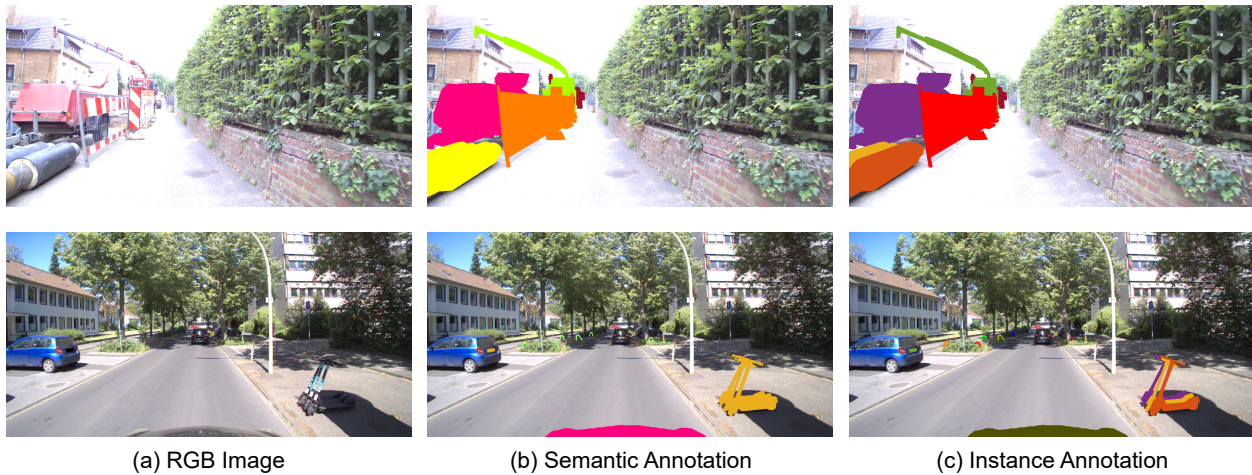


Figure 2. Our dataset, PANIC, provides pixel-wise annotations of unknown semantic categories and object instances of RGB images. The images have been recorded with our own inhouse sensor suite (Vizzo et al. 2023) mounted on our vehicle driving in Bonn, Germany. The dataset consists of images collected at different times of day over the span of more than a year.

Although influential, those approaches often struggled to reliably identify unknown samples because deep networks tend to produce overconfident predictions even for out-of-distribution inputs (Nguyen et al. 2015; Tommasi et al. 2017).

To address this limitation, follow-up works explored alternative signals for anomaly detection, including entropy maximization (Dhamija et al. 2018), energy-based scores (Liu et al. 2020), and uncertainty estimation through Bayesian deep learning (Gal and Ghahramani 2016; Lakshminarayanan et al. 2017; Sapkota and Yu 2022) or gradient-based measures (Liu et al. 2018; Maag and Riedlinger 2024). Other methods leverage examples of unknown objects during training to better separate known and unknown regions at test time (Hwang et al. 2021; Wong et al. 2020). Generative models have also been explored for open-world perception. While some approaches detect unknowns through reconstruction errors (Zhao 2023), others, such as OpenGAN (Kong and Ramanan 2021), generate synthetic open-set examples to improve the discrimination between known and unknown samples. More recently, attention has shifted toward open-set panoptic segmentation, where anomalous regions are further decomposed into object instances using class-agnostic grouping strategies (Gasperini et al. 2023; Nayal et al. 2023; Nekrasov et al. 2023; Rai et al. 2024; Schmidt et al. 2025). The ability to segment previously unseen objects is closely related to the literature on object proposals and class-agnostic instance segmentation. DeepMask (Pinheiro et al. 2015) learns category-independent object proposals by exploiting generic visual cues such as object boundaries and shape regularities, while Recurrent Pixel Embedding (Kong and Fowlkes 2018) groups pixels into instances through learned embedding representations without relying on category-specific predictions. These works suggest that object-centric cues can generalize beyond the training taxonomy and provide an important inductive bias for reasoning about unseen objects. Our approach builds on a similar intuition, but addresses the more challenging open-world panoptic segmentation setting, where the model must not only identify and segment unknown instances, but also discover and maintain semantic category identities across images. To encourage such object-centric behavior,

we further incorporate geometric priors through divergence- and curl-based regularization of the learned offset field.

A related but less explored problem is the discovery of multiple novel categories. In semantic segmentation, NCDSS (Zhao et al. 2022) discovers novel semantic classes from unlabeled images containing the categories to be learned, assuming that the number of unknown categories is known beforehand. At the instance level, Weng et al. (2021) and RNCDL (Fomenko et al. 2022) jointly perform unknown instance segmentation and category discovery, producing cross-image consistent labels for novel categories. However, these methods operate in an offline setting, where the entire dataset is available before category discovery takes place.

The online counterpart of this problem has been studied in image classification through online category discovery. Methods such as OCD (Du et al. 2023) and PHE (Zheng et al. 2024) process samples sequentially and decide whether each incoming observation belongs to a previously known class or to a newly discovered category. Open-world semantic segmentation extends this idea to dense prediction by incrementally discovering multiple semantic categories while maintaining category consistency across images (Sodano et al. 2024). Open-world panoptic segmentation further extends this setting by additionally requiring instance-level segmentation. As a result, it combines semantic category discovery, unknown instance segmentation, and online category formation within a single framework.

Another related research direction is open-vocabulary segmentation, also known as zero-shot segmentation. Those methods aim to recognize and segment categories that were not seen during training by leveraging large-scale vision-language models such as CLIP (Radford et al. 2021). Most approaches rely on text prompts or predefined vocabularies to guide the segmentation process (Ghiasi et al. 2022; Liang et al. 2023; Qin et al. 2023; Xu et al. 2023; Zhang et al. 2023), requiring potential object categories to be specified beforehand. More recently, promptless VLM-based approaches such as Zero-Guidance Segmentation (Rewatbowornwong et al. 2023) and AutoSeg (Ülger et al. 2025) have explored automatic category labeling without user-provided prompts by combining class-agnostic segmentation

with vision-language representations. While these methods remove the need for predefined vocabularies, the resulting labels remain image-dependent and do not explicitly maintain category identities across images. Consequently, semantically equivalent objects may receive different textual descriptions or be represented at different levels of granularity. In contrast, open-world panoptic segmentation focuses on discovering and maintaining consistent category identities across images while simultaneously providing semantic and instance-level segmentation, without relying on language supervision at inference time.

In this paper, we present a fully convolutional neural network for open-world panoptic segmentation. Our goal is to discover multiple unknown semantic categories while simultaneously segmenting their individual object instances. Unlike prior approaches, we do not rely on unknown objects during training, predefined vocabularies, or vision-language supervision. To the best of our knowledge, our approach is the first method specifically designed for open-world panoptic segmentation, jointly addressing online category discovery and unknown instance segmentation.

Datasets. While large-scale datasets exist for closed-world segmentation in different domains (Behley et al. 2021; Cordts et al. 2016; Lin et al. 2014; Sethuraman et al. 2025; Vidanapathirana et al. 2025; Weyler et al. 2024b), datasets for open-world segmentation are comparably rare. The WildDash dataset (Zendel et al. 2018) provides anomalous images where full-image anomalies are present. Another dataset is MVTec (Bergmann et al. 2019), which mostly targets the industrial scenario. BrainMRI & HeadCT (Salehi et al. 2021) are two datasets targeting the medical domain for detecting lesions on different organs. ALLO (Leveugle et al. 2024) is a synthetic photorealistic dataset with full-image anomalies for robot operations in lunar orbit. Among the most common datasets for open-world segmentation in the autonomous driving domain are the Fishyscapes-LostAndFound benchmark (Blum et al. 2019; Pinggera et al. 2016), which is based on the same setup as Cityscapes and presents anomalous objects in the middle of the street, the CAOS benchmark (Hendrycks et al. 2022) that originates from the BDD100K (Yu et al. 2020) dataset to create an open-world test set, and TAO-OW (Liu et al. 2022) for open-world instance tracking. These datasets are, however, characterized by a low diversity of anomalies. Chan et al. (2021) introduced the SegmentMeIfYouCan benchmark, which is a test set that relies on the known classes from Cityscapes and introduces anomalies of various kinds and sizes. However, it provides a limited number of images, and no semantic or instance annotation, but only a binary segmentation mask between known and unknown. Recently, Nekrasov et al. (2024) extended SegmentMeIfYouCan by adding instance information, even though semantic information is still missing.

In this work, we propose a dataset for open-world panoptic segmentation with more than 50 unknown classes and more than 4000 instances. We assume everything that appears in Cityscapes to be known, and provide pixel-wise annotations for the unknowns. We also provide public competitions for several open-world segmentation task, spanning from anomaly segmentation to open-world panoptic segmentation.

3 Task Description

Open-world segmentation remains a relatively under-explored area, having gained momentum only in recent years with the introduction of dedicated benchmarks (Blum et al. 2019; Chan et al. 2021; Nekrasov et al. 2024) and with adaptations of closed-world datasets for open-world evaluation (Hendrycks et al. 2022; Lin et al. 2014). This emerging status has led to terminology that is often ambiguous or inconsistently used across the literature. For example, the task of discovering multiple novel semantic categories has been referred to as both open-set and open-world semantic segmentation (Rai et al. 2024; Sodano et al. 2024). Likewise, the task of identifying a single “unknown” super-category and segmenting all object instances within it has appeared under names such as open-world instance segmentation (Nunes et al. 2022), open-set panoptic segmentation (Rai et al. 2024), and open-world panoptic segmentation (Chakravarthy et al. 2025; Schmidt et al. 2025). Additionally, some works use the term open-set semantic segmentation (Kong and Ramanan 2021) to describe the task of assigning pixels either to one of the known semantic classes or to a generic unknown category. In our formulation, this setting is equivalent to anomaly segmentation, since all tasks considered build upon closed-world semantic segmentation and therefore always predict the known classes alongside any unknown regions.

Thus, in this section, we propose a clear and unified terminology for all open-world segmentation tasks and use it throughout the paper. In Fig. 3, we show a visual example of the four tasks on an exemplary image that has been manually labeled for illustration purposes. All tasks build on top of closed-world segmentation, i.e., they output prediction masks for known classes as well. Additionally, none of these tasks relies on a known vocabulary of categories to be discovered, as open-vocabulary approaches are not covered.

3.1 Anomaly Segmentation

SegmentMeIfYouCan (Chan et al. 2021) first formalized the task of *anomaly segmentation*: given a model trained on a certain number of known categories, anomaly segmentation segments the areas belonging to none of them. There is no concept of unknown category or object. The input image is segmented into two areas: one belonging to everything that is known (i.e., seen at training time), and one including everything else.

Given an image $I \in \mathbb{R}^{H \times W}$, the prediction for anomaly segmentation is a mask $M_a \in \{0, 1\}^{H \times W}$, where 0 indicates known and 1 indicates unknown.

3.2 Open-World Semantic Segmentation

Open-world semantic segmentation (Sodano et al. 2024) segments the anomalous area into different semantic categories. As for standard semantic segmentation, these categories stay consistent (i.e., areas belonging to the same category across different images are assigned to the same semantic label). This task has no object instances.

Given an image $I \in \mathbb{R}^{H \times W}$, the prediction for this task is a mask $M_s \in \{0, \dots, K\}^{H \times W}$, with 0 for known and $1, \dots, K$ for the K new categories.

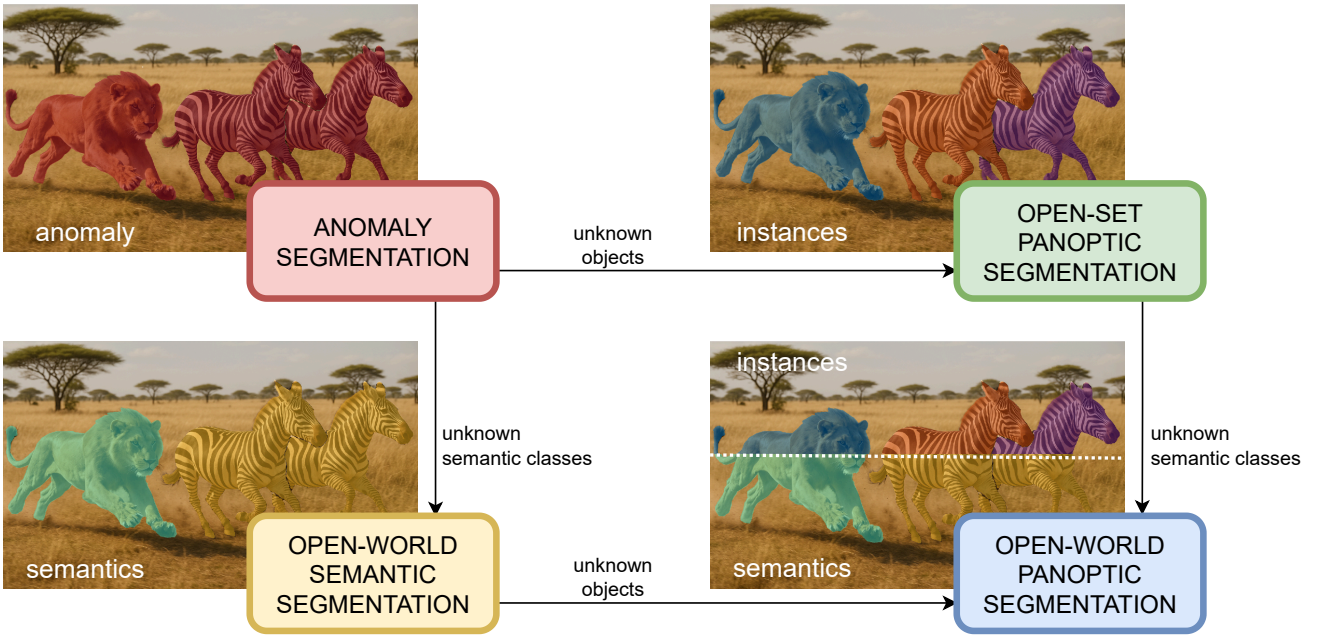


Figure 3. A visual breakdown of the four open-world segmentation tasks. Anomaly segmentation segment all anomalous areas as unknown (zebras and lion together). Open-world semantic segmentation separates classes but has no objects (zebras segmented together, lion separate). Open-set panoptic segmentation segments separate objects but has no category information. Open-world panoptic segmentation has both, classes and object information. Exemplary RGB image is generated with perplexity.ai (Perplexity Deep Research 2025).

3.3 Open-Set Panoptic Segmentation

Open-set panoptic segmentation is more common in the literature (Hwang et al. 2021; Rai et al. 2024), and consists of segmenting individual objects within the anomaly segmentation mask. This task does not look at finding new semantic classes, but rather new objects. Generally speaking, this task performs class-agnostic instance segmentation within the anomalous area.

Given an image $I \in \mathbb{R}^{H \times W}$, the prediction for this task is a mask $M_a \in \{0, 1\}^{H \times W}$, where 0 indicates known and 1 indicates unknown, and a mask $M_i \in \{0, \dots, N\}^{H \times W}$ where 0 indicates both known and no object areas, and $1, \dots, N$ indicate the N object instances discovered.

3.4 Open-World Panoptic Segmentation

Open-world panoptic segmentation builds on open-world semantic segmentation and aims to discover not only novel and consistent semantic classes, but also individual objects within these classes. It goes a step further from open-set panoptic segmentation as well, since the latter does not aim to discover new semantic classes. The task of open-world panoptic segmentation gives a holistic understanding of the scene, since every pixel belongs to a semantic class and will have an instance ID (when applicable), irrespectively of it being *known* or *unknown*.

Given an image $I \in \mathbb{R}^{H \times W}$, the prediction for open-world panoptic segmentation is a mask $M_s \in \{0, \dots, K\}^{H \times W}$, where 0 indicates known and $1, \dots, K$ indicate the K categories discovered, and a mask $M_i \in \{0, \dots, N\}^{H \times W}$, where 0 indicates both known and no object areas, and $1, \dots, N$ indicate the N object instances discovered.

Open-world panoptic segmentation provides a richer scene representation than its related variants because it

jointly discovers previously unseen semantic categories and their individual object instances, without requiring prior prompts or textual descriptions. Unlike open-world semantic segmentation, which identifies unknown regions but does not separate individual objects, open-world panoptic segmentation enables reasoning at the instance level. This allows an autonomous system to estimate how many unknown objects are present and to distinguish them individually. Compared to open-set panoptic segmentation, which groups all unknown pixels into a single category, it further enables the discovery of meaningful structure among previously unseen objects.

While binary anomaly segmentation may be sufficient for certain safety-critical behaviors such as triggering a stop or slowdown, many downstream robotics tasks are inherently object-centric. Multi-object tracking, behavior prediction, and planning often require reasoning about individual agents rather than undifferentiated anomaly regions. In this context, distinguishing and grouping unknown objects may provide additional information that can support future decision-making processes. More generally, we view open-world panoptic segmentation as a step toward holistic scene understanding in environments where not all object categories are known during training, extending the panoptic segmentation paradigm to open-world settings.

In this article, we propose a novel method and provide a benchmark for open-world panoptic segmentation.

4 Our Methodology

We address the problem of open-world panoptic segmentation. Our approach, illustrated in Fig. 4, uses a fully convolutional neural network with a single encoder and three decoders. Since our contributions focus primarily on the design of loss functions and on the strategy used to discover

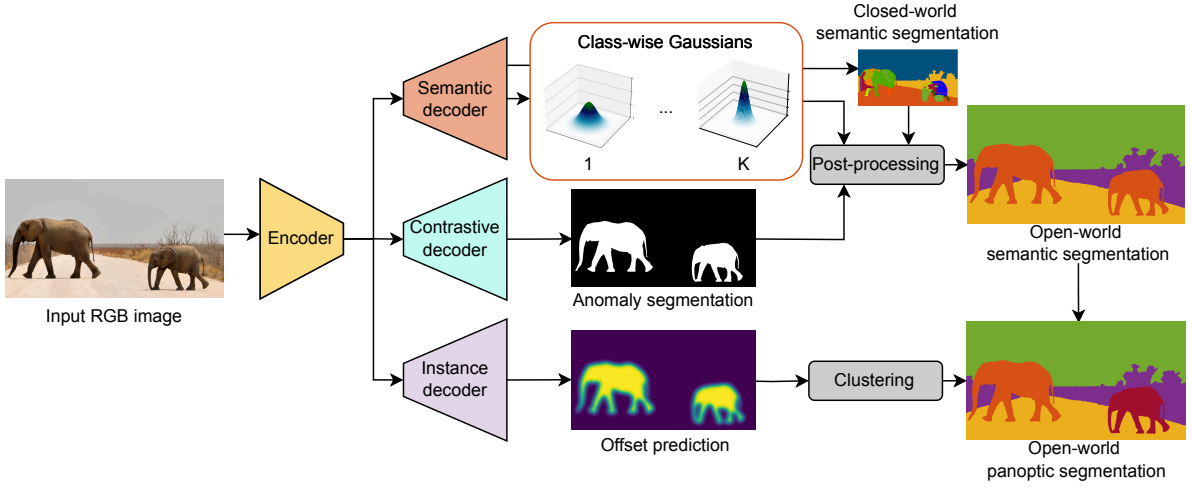


Figure 4. Our network processes an RGB image via an encoder and three decoders, for semantic segmentation, anomaly segmentation and class-agnostic instance segmentation. The semantic segmentation decoder also builds class descriptors for the known categories. Results are post-processed and yield the final open-world panoptic segmentation result.

novel categories and objects, the method is not tied to this particular architectural choice. In fact, it can be integrated into any other backbone network without modification.

4.1 Encoder Architecture

We use a lightweight fully-convolutional neural network, to facilitate future deployment in the real world. Our CNN is composed of one encoder and three decoders. We use a ResNet34 (He et al. 2016) encoder, where we replaced the standard ResNet block with the NonBottleneck-1D block (Romera et al. 2018). This choice is motivated solely from the fact that we aim to make our architecture lighter, and the NonBottleneck-1D block replaces all 3×3 convolution by a sequence of 3×1 and 1×3 convolutions with a ReLU in between. Additionally, we incorporate contextual information by means of a pyramid pooling module (Zhao et al. 2017) after the encoder.

4.2 Decoder Architectures

Our three decoders have the same structure and are composed of three SwiftNet modules (Orsic et al. 2019) with NonBottleneck-1D blocks, and two final upsampling modules based on nearest-neighbor and depth-wise convolutions (Chollet 2017), that have the major advantage of substantially reducing the computations needed. Following the standard U-Net (Ronneberger et al. 2015) design, we also employ encoder-decoder skip connections after each downsampling stage of the encoder to propagate fine-grained features to the three decoders.

The first decoder, called “semantic decoder”, targets semantic segmentation and additionally builds a class descriptor for each known category. At inference time, we compare the final features to the known descriptors to discriminate between known and unknown classes, building descriptors for unknown classes as well. The second decoder, called “contrastive decoder”, targets anomaly segmentation by mapping the features of pixels belonging to known classes on a unit hypersphere, and the ones belonging to unknown classes to 0. The third decoder, called “instance decoder” targets class-agnostic instance segmentation by means of vector fields-inspired loss functions (Weyler et al. 2024a).

The outputs of the decoders go through different post-processing phases to obtain the final open-world panoptic segmentation result.

Consider an image $I \in \mathbb{R}^{H \times W}$. We refer to the set of pixels in the image as $\Omega = \{(1, 1), \dots, (H, W)\}$. Additionally, we call $Y \in \{1, \dots, K\}^{H \times W}$ the ground truth semantic mask of I , where K is the number of known categories used at training time. In contrast, we refer to the predicted (closed-world) semantic mask as $\hat{Y} \in \{1, \dots, K\}^{H \times W}$. Similarly, we call Z the ground truth and $\hat{Z}$ the predicted instance mask, respectively. Finally, we denote with $\Omega_k = \{p \in \Omega \mid Y_p = k\}$ the set of pixels in the image whose ground truth label is k , and with $\hat{\Omega}_k = \{p \in \Omega \mid \hat{Y}_p = Y_p = k\}$ the set of pixels whose predicted label $\hat{Y}_p$ and ground truth label Y_p are both equal to k , i.e., the set of true positives for class k . We do not use bold notation for the pixel tuple p for the sake of clarity.

Semantic Decoder. The semantic decoder targets semantic segmentation and it leverages the standard weighted cross-entropy loss:

$$\mathcal{L}_{\text{sem}} = -\frac{1}{|\Omega|} \sum_{p \in \Omega} \omega_k t_p^\top \log(\sigma(f_p)), \quad (1)$$

where ω_k is a class-wise weight computed via the inverse frequency of each class in the dataset, $t \in \mathbb{R}^{H \times W \times K}$ is the one-hot encoded ground truth annotation of the image, $t_p \in \mathbb{R}^K$ is the one-hot encoded ground truth annotation at pixel location p , $\sigma(\cdot)$ denotes the softmax operation, and f_p denotes the pre-softmax feature predicted at pixel p .

Unlike approaches that operate on pre-softmax features (Sodano et al. 2024), whose dimensionality is tied to the number of known classes and can therefore limit performance in few known classes scenarios, we apply our loss directly to the higher-dimensional, class-agnostic pre-logit features ℓ_p , where p indicates the pixel. This allows us to learn more expressive and discriminative class prototypes.

To build the class descriptors, during training we accumulate the pre-logits of all true positives for each class, where a true positive is a pixel that is correctly segmented. By doing this, we can build a running average class descriptor $\mu_k \in \mathbb{R}^D$ for each class $k \in \{1, \dots, K\}$

that appears at training time:

$$\mu_k = \frac{1}{|\hat{\Omega}_k|} \sum_{p \in \hat{\Omega}_k} \ell_p. \quad (2)$$

Together with the mean, we also compute the per-class variance $\sigma_k^2 \in \mathbb{R}^D$ via the sum of squares:

$$\sigma_k^2 = \frac{1}{|\hat{\Omega}_k|} \sum_{p \in \hat{\Omega}_k} (\ell_p - \mu_k) \odot (\ell_p - \mu_k), \quad (3)$$

where $\odot$ indicates the element-wise product (Hadamard product).

At the beginning of epoch e , each category has a mean μ_k^{e-1} and a variance σ_k^{e-1} , computed at the previous epoch $e-1$. The goal of having a unique class descriptor for each known class is that the prediction at each pixel belonging to that class must be equal to the descriptor. Thus, we use the feature loss function $\mathcal{L}_{\text{feat}}$ at the pre-logit level given by

$$\mathcal{L}_{\text{feat}} = \frac{1}{|\Omega|} \sum_{k=1}^K \sum_{p \in \Omega_k} \frac{\|\ell_p - \mu_k^{e-1}\|^2}{(\sigma_k^{e-1})^\top (\sigma_k^{e-1})}. \quad (4)$$

The idea is that true positives help build class descriptors, as specified in Eq. (2), which are then used as supervision in the next epoch. This loss formulation extends our previous work (Sodano et al. 2024) by operating on the pre-logit representation rather than the pre-softmax feature. This removes the dependence of the feature dimensionality on the number of known training classes, making the representation more suitable for category discovery when only few known categories are available.

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

$$\mathcal{L}_{\text{sdec}} = w_1 \mathcal{L}_{\text{sem}} + w_2 \mathcal{L}_{\text{feat}}. \quad (5)$$

Contrastive Decoder. The contrastive decoder targets anomaly segmentation, and aims to predict features whose norm is close to 1 for pixels belonging to known classes, and to 0 for pixels belonging to unknown classes.

This decoder also operates at the pre-logit level and combines the contrastive loss (Chen et al. 2020) and the objectsphere loss (Dhamija et al. 2018). To do so, we first compute the mean pre-logit $\bar{\ell}_k \in \mathbb{R}^D$ for each class k in the current image from the pre-logit predicted by this decoder. We compute the contrastive loss $\mathcal{L}_{\text{cont}}$ such that $\bar{\ell}_k$ approximates the normalized mean representation $\bar{\mu}_k^{e-1}$ of the corresponding class in the previous epoch μ_k^{e-1} and gets dissimilar from the other classes mean representation:

$$\mathcal{L}_{\text{cont}} = - \sum_{k=1}^K \log \frac{\exp(\bar{\ell}_k^\top \bar{\mu}_k^{e-1} / \tau)}{\sum_{i=1}^K \exp(\bar{\ell}_k^\top \bar{\mu}_i^{e-1} / \tau)}, \quad (6)$$

where τ is a temperature parameter. The goal is to have $\bar{\ell}_k$ approximate the corresponding vector in the feature bank, i.e., the normalized class descriptors $\bar{\mu}_k^{e-1}$. To achieve this, we changed from the pre-softmax features to the pre-logit in this loss function, in order to have dimensionally-consistent vectors. The outcome of the contrastive loss is to scatter the class-specific pre-logits on the unit hypersphere.

Additionally, we use the objectsphere loss function

$$\mathcal{L}_{\text{obj}} = \begin{cases} \max(1 - \|\ell_p\|^2, 0) & , \text{if } p \in \Omega_k \\ \|\ell_p\|^2 & , \text{otherwise} \end{cases}, \quad (7)$$

where Ω_k is the set of pixels in the image belonging to known classes, while the others belong to unlabeled (void) areas of the image and are treated as unknown. In practice, these void regions correspond to pixels annotated as “void” in Cityscapes and are used as known-unknown examples during training. The loss pushes known pre-logits to have norm ≥ 1 and unknowns to have norm 0. Although other values could be used, norm 1 aligns with the contrastive loss in Eq. (6), which maps features to the unit sphere. Using a different norm would conflict with this goal, especially values larger than 1.

We optimize the contrastive decoder with a weighted sum of the loss functions introduced above:

$$\mathcal{L}_{\text{cdec}} = w_3 \mathcal{L}_{\text{cont}} + w_4 \mathcal{L}_{\text{obj}}. \quad (8)$$

The contrastive decoder is used for anomaly segmentation and could, in principle, be replaced by alternative architectures serving the same purpose. We adopt this design because it is lightweight, architecturally consistent with the semantic decoder, and operates in the same pre-logit representation space used for category discovery.

Instance Decoder. The instance decoder addresses class-agnostic instance segmentation, where we predict a 2D offset for each pixel, so that pixels belonging to the same object instance point to the same area in the image and can be clustered together.

Given an instance $\mathcal{C}_j \subset \Omega$, defined as the set of pixels of the images belonging to the j -th object instance, with a centroid $\mathbf{c}_j$, we want to obtain offset vectors for all pixels belonging to $\mathcal{C}_j$ that point toward $\mathbf{c}_j$, while the remaining offset vectors should not. We achieve this by means of the Lovász Hinge loss (Neven et al. 2019)

$$\mathcal{L}_{\text{off}} = \frac{1}{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{C}|} \text{Lovász}(\mathbf{F}_{\mathcal{C}_j}, \mathbf{G}_{\mathcal{C}_j}). \quad (9)$$

In this equation, $\mathcal{C}$ is the set of instances in the image, $\mathbf{F}_{\mathcal{C}_j} \in \mathbb{R}^{H \times W}$ is a soft-mask obtained from the offset predictions for optimization purposes, while $\mathbf{G}_{\mathcal{C}_j} \in \{0, 1\}^{H \times W}$ denotes the binary ground truth mask of the j -th instance. The soft mask $\mathbf{F}_{\mathcal{C}_j}$ is not directly predicted by the network; instead, it is an intermediate quantity derived from the predicted offsets and used exclusively within the Lovász-Hinge loss. Specifically, each pixel receives a score based on how close the location pointed to by its predicted offset is to the centroid of instance $\mathcal{C}_j$. This is formalized as

$$f_{\mathcal{C}_j} = \exp\left(-\frac{\|e_p - \mathbf{c}_j\|^2}{2\eta^2}\right), \quad (10)$$

where $e_p = p$ indicates the pixel location pointed by the prediction at pixel p , and η is a hyperparameter that defines an isotropic clustering region around the centroid. This loss function encourages the network to predict, for all pixels associated with a specific instance, an offset vector that points toward their corresponding centroid. Simultaneously,

it penalizes offset vectors that point towards a different centroid than their own. This forms distinct clusters in 2D space, even if the centroid lies outside the instance.

Following the idea formulated by [Weyler et al. \(2024a\)](#), we adopt divergence and curl loss functions to refine offset predictions. These two loss functions are inspired by the concept of vector field in physics, and have so far only been applied for closed-world segmentation in the case of partial annotations. Here, we show their effectiveness for open-world segmentation, and provide a quantitative analysis of each component in the ablation study.

Divergence captures how pixels “flow” toward or away from a point, aligning well with the idea of pixels pointing to a centroid. Curl measures rotation in the field, helping to regularize the vector pattern. For formalizing the divergence and curl loss functions, we need a simplified vector field defined by a continuous multivariable function $\mathbf{O}(h, w) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ given by:

$$\mathbf{O}(h, w) = \begin{bmatrix} o_h(h, w) \\ o_w(h, w) \end{bmatrix} = \begin{bmatrix} -h + c_h \\ -w + c_w \end{bmatrix}, \quad (11)$$

that behaves as a perfect sink, i.e., all offset vectors point toward $\mathbf{c} = [c_h, c_w]^\top$. In 2D, divergence is defined as

$$\text{div } \mathbf{O} := \frac{\partial o_h(h, w)}{\partial h} + \frac{\partial o_w(h, w)}{\partial w}. \quad (12)$$

Thus $\frac{\partial o_h(h, w)}{\partial h} = \frac{\partial o_w(h, w)}{\partial w} = -1$ and $\text{div } \mathbf{O} = -2$. We enforce that divergence is equal to -2 and its two components must be equal with two regression loss functions:

$$\begin{aligned} \mathcal{L}_{\text{div}} &= \frac{1}{|\Omega|} \sum_{p \in \Omega} \rho \left((\text{div } \mathbf{O})_p - (-2) \right) \\ \mathcal{L}_{\text{div}}^{\text{aux}} &= \frac{1}{|\Omega|} \sum_{(h, w) \in \Omega} \rho \left(\frac{\partial o_h(h, w)}{\partial h} - \frac{\partial o_w(h, w)}{\partial w} \right), \end{aligned} \quad (13)$$

where $\rho(\cdot)$ denotes the Geman-McClure loss ([Barron 2019](#)), and (h, w) denote the image coordinates of pixel p .

The curl of a vector field $\mathbf{O}$ is defined as

$$\text{curl } \mathbf{O} := \frac{\partial o_h(h, w)}{\partial w} - \frac{\partial o_w(h, w)}{\partial h}. \quad (14)$$

When applying the curl to the vector field in Eq. (11), we obtain $\frac{\partial o_h(h, w)}{\partial w} = \frac{\partial o_w(h, w)}{\partial h} = 0$, which imply $\text{curl } \mathbf{O} = 0$. Concretely speaking, this means we want no rotational behavior in our vector field. We formalize two loss functions to have the curl equal to zero and the two partial derivatives to be equal to each other:

$$\begin{aligned} \mathcal{L}_{\text{curl}} &= \frac{1}{|\Omega|} \sum_{p \in \Omega} \rho \left((\text{curl } \mathbf{O})_p \right) \\ \mathcal{L}_{\text{curl}}^{\text{aux}} &= \frac{1}{|\Omega|} \sum_{(h, w) \in \Omega} \rho \left(\frac{\partial o_h(h, w)}{\partial w} - \frac{\partial o_w(h, w)}{\partial h} \right). \end{aligned} \quad (15)$$

The instance decoder is optimized with the weighted sum of all losses introduced so far

$$\begin{aligned} \mathcal{L}_{\text{ins}} &= w_5 \mathcal{L}_{\text{off}} + w_6 \mathcal{L}_{\text{div}} + w_7 \mathcal{L}_{\text{div}}^{\text{aux}} + \\ &\quad w_8 \mathcal{L}_{\text{curl}} + w_9 \mathcal{L}_{\text{curl}}^{\text{aux}}. \end{aligned} \quad (16)$$

4.3 Post-Processing

Given the outputs of the three decoders, we now describe the generation of the corresponding masks.

Post-Processing for Anomaly Segmentation. We obtain anomaly segmentation results by fusing the outputs of the semantic and contrastive decoders. The semantic decoder provides class descriptors (mean μ_k and variance σ_k^2) in the pre-logit space for each known class, allowing us to construct a multi-variate normal distribution $\mathcal{N}_k(\mu_k, \Sigma_k)$, where $\Sigma_k = \text{diag}(\sigma_k^2)$ is the covariance matrix, which we obtain from the variance assuming that all classes are independent. In this way, for each predicted pre-logit ℓ_p at pixel p , we can compute its fitting score to each known category by means of the squared exponential kernel

$$s_k(\ell_p) = \exp \left(-\frac{1}{2} (\ell_p - \mu_k)^\top \Sigma_k^{-1} (\ell_p - \mu_k) \right). \quad (17)$$

Then, we take the highest score

$$s(p) = \max_k s_k(\ell_p), \quad (18)$$

which corresponds to the closest known category to the predicted pixel. At test time, this process replaces the standard procedure for closed-world semantic segmentation, i.e., $\text{argmax}(\cdot)$ on pre-softmax outputs.

To determine whether a pixel belongs to a known or unknown region, we apply a simple post-processing rule to the outputs of the first two decoders. For the semantic decoder, we compare each pre-logit feature to the distribution of its corresponding known class: if the feature lies within 1σ from the class mean, the pixel is considered known; otherwise, it is marked as unknown, yielding a binary score $s_{\text{unk}, p}^{\text{sem}} \in \{0, 1\}$. The contrastive decoder provides a complementary viewpoint by enforcing known-class features to have norm close to 1 and unknown-class features to have norm near 0. Under the assumption that unknown features follow a standard normal distribution $\mathcal{N}(0, 1)$, known features naturally emerge as outliers. Applying the same 1σ threshold to the feature norms produces a second boolean score $s_{\text{unk}, p}^{\text{con}} \in \{0, 1\}$. Together, these two signals provide a robust estimation of unknown regions.

Finally, we consider a pixel to be unknown only if both $s_{\text{unk}, p}^{\text{sem}} = 1$ and $s_{\text{unk}, p}^{\text{con}} = 1$.

Post-Processing for Open-World Semantic Segmentation. At test time, we have K pre-logit descriptors ℓ_k for the known classes. For each pixel, we compare its predicted pre-logit descriptor $\ell_p \in \mathbb{R}^D$ against all existing descriptors using Eq. (18). If the closest descriptor belongs to a known class and satisfies the similarity criterion, the pixel is assigned to that class. If no existing descriptor is sufficiently similar, the pixel is considered to belong to a newly discovered category and its descriptor is added to the database as a new unknown class. As described in the open-world semantic segmentation task, we aim to maintain category consistency across images. Therefore, when a pixel matches the descriptor of a previously discovered unknown category, it is assigned to that category rather than creating a new one. In this way, our database contains $\bar{K} = K + K_{\text{unk}}$ descriptors, where K_{unk} is the number of unknown categories discovered so far, effectively treating discovered categories as “newly known” classes.

Dataset	Images		Semantic Classes	Instances	Hidden Test Set
	Val	Test			
Fishyscapes Lost-and-Found (Blum et al. 2019)	373	1203	N.A.	1864	✓
CAOS BDDAnomaly (Hendrycks et al. 2022)	0	810	3	1231	✗
RoadObstacle21 (Chan et al. 2021)	0	327	N.A.	388	✓
SegmentMeIfYouCan (Chan et al. 2021)	10	100	N.A.	262	✓
PANIC (ours)	131	679	58	4029	✓

Table 1. Comparison of open-world segmentation datasets. In the semantic classes, “N.A.” means that there is no label.



Figure 5. Sensor setup we used for recording data for the PANIC dataset. The setup includes four cameras, one GNSS/IMU device, and two 3D LiDARs. For further details, please refer to Vizzo et al. (2023).

While descriptors of known classes remain fixed after training, descriptors of discovered unknown categories are updated online using a running mean and variance. This follows the same principle used during training for known classes, where class prototypes are updated from correctly classified samples. The difference is that, at test time, ground-truth labels are unavailable, and therefore the update is performed using all pixels assigned to the corresponding discovered category. This online update allows the category descriptors to progressively adapt as additional instances of the same unknown class are observed.

Post-Processing for Open-World Panoptic Segmentation. We obtain instance predictions by clustering the raw offset predictions with HDBSCAN (McInnes et al. 2017). Unlike the soft masks used during training, no thresholding or intermediate score computation is performed at inference time. We execute the clustering only in the “thing” areas, which means all the known “thing” classes, and all the unknown categories. To ensure consistency, we filter the clustering results using semantic predictions, preventing pixels from different semantic classes from being grouped into the same instance. In essence, we prioritize semantic consistency over instance grouping.

5 The PANIC Dataset

In this article, we also introduce PANIC (Panoptic Anomalies In Context), an open-world panoptic segmentation test set in the autonomous driving context. Datasets for anomaly and open-world segmentation are rare compared to closed-world segmentation ones. PANIC includes 800 images, and contains 58 unknown categories. Similarly

to SegmentMeIfYouCan (Chan et al. 2021), we consider Cityscapes (Cordts et al. 2016) as the standard training set with known classes. The 58 categories are divided into validation and test classes. The validation set includes classes that do not have a label to optimize for during training, but still appear at training time in the void/unlabeled area (such as garbage bins, parkimeters, monuments, etc.). In the literature, they are referred to as *known unknown* (Bendale and Boult 2016). The test set contains the *unknown unknowns*, more challenging objects such as electric scooters, recumbent bicycles, forklifts, etc., that are completely absent from the training set.

We release image and ground truths for the validation set, and only images for the hidden test set which is used for the competitions mentioned below.

A comparison with existing datasets is shown in Tab. 1. LostAndFound (Blum et al. 2019) contains 9 different types of anomalies, misclassifying common Cityscapes classes like children and bicycles as anomalies. The CAOS BDD-Anomaly benchmark (Hendrycks et al. 2022) suffers from a similar low-diversity issue, featuring just 3 unknown classes sampled from BDD100K (Yu et al. 2020). Both, Fishyscapes and CAOS, try to mitigate this low diversity by introducing synthetic data in the dataset, but this lacks real-world realism. RoadObstacle21 and SegmentMeIfYouCan (Chan et al. 2021) offer more diverse anomalies, but only focus on anomaly segmentation. Recently, Nekrasov et al. (2024) provided instance-level annotation for these datasets, but without semantic labels, falling under open-set panoptic segmentation, where instances are detected without category differentiation.

In PANIC, anomalies can appear anywhere in the image. They can have any size, and are not limited to active traffic participants. We provide pixel-wise annotations of semantic classes and object instances.

5.1 Data Collection and Labeling

We recorded all images of PANIC in Bonn, Germany, with a car equipped with commonly used automotive sensors, such as cameras and LiDARs (Vizzo et al. 2023), shown in Fig. 5. The acquisition platform additionally includes synchronized and calibrated LiDARs, IMUs, GPS receivers, and other sensors, as detailed by Vizzo et al. (2023). However, PANIC focuses exclusively on image-based open-world panoptic segmentation and therefore only utilizes RGB images acquired by the front camera, a Basler Ace acA2040-35gc. While the additional sensor modalities are not used in the present benchmark, they may provide valuable geometric information for future research on multimodal open-world perception and object discovery.

After recording, we sampled a subset of the recorded images in order to avoid duplicates and filter out images that did not contain any anomalous object, and labeled them using the online tool [segments.ai](#). The 19 Cityscapes evaluation classes (Cordts et al. 2016) on which most segmentation models for autonomous driving are trained, serve as basis to determine anomalies. Everything that belongs to any of those 19 classes is labeled as “not anomaly”. For all the rest, we provide a pixel-wise semantic and instance annotation. To comply with privacy restrictions, we anonymize all faces, license plates, and windows of private buildings when necessary. The provided images have those areas in black. This does not bias any learning algorithm, since the dataset we propose is supposed to be an evaluation-only dataset, and the training happens on Cityscapes, which provides the known classes. Additionally, we label an “ignore” class that will not be included in the evaluation, as well as the anonymized areas.

We provide more details on PANIC, including the categories we labeled, in Sec. A of the supplementary material.

5.2 Benchmarks and Metrics

We propose four benchmarks accompanied by public Codabench competitions (Xu et al. 2022), one for each of the segmentation tasks described in Sec. 3. Each has its own evaluation pipeline and metrics.

5.2.1 Anomaly Segmentation

Following SegmentMeIfYouCan (Chan et al. 2021), we use two groups of metrics for anomaly segmentation: pixel-level and component-level.

Pixel-level metrics include the area under the precision-recall curve (AUPR), which evaluates pixel-wise anomaly score separability and favors the minority class, and the false positive rate at 95% true positive rate (FPR95), that indicates how many false positive predictions are needed to reach 95% true positive rate. Pixel-level metrics, however, fail to properly evaluate small anomalies. Component-level metrics better capture all anomaly regions irrespectively of their size. SegmentMeIfYouCan (Chan et al. 2021) introduces three metrics for component-level evaluation: the segment-wise intersection over union (sIoU) for evaluating true positive and false negatives, the positive predicted value (PPV, or component-wise precision) for false positives, and the component-wise F1-score which summarizes true positive, false positive, and false negative. For further details, insights, and considerations, we refer to the original paper (Chan et al. 2021).

We decided to not use the area under the ROC curve (AUROC), because recently several papers showed its limitations (Halligan et al. 2015; Humblot-Renaux et al. 2023; Wang et al. 2022), as two models with the same performance may differ substantially 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.

5.2.2 Open-World Semantic Segmentation

Open-world semantic segmentation is still under-explored, thus a proper evaluation pipeline does not exist. Since

the standard evaluation of semantic segmentation relies on intersection over union, or in short IoU (Everingham et al. 2010), we extend it to the open-world setting.

Closed-world IoU is typically computed using a confusion matrix $C \in \mathbb{R}^{K \times K}$, where K is the number of known classes. Each row i counts predictions for class i , and a perfect IoU is achieved when only diagonal entries (i, i) are non-zero, and matched the number of pixels belonging to ground truth class i . In the following, we indicate as $\text{row}_i \in \mathbb{R}^K$ the i -th row of a confusion matrix. Similarly, $\text{column}_k \in \mathbb{R}^K$ is the k -th column of the confusion matrix. The closed-world IoU concept can easily be extended to the open-world case, with two major modifications:

1. The open-world confusion matrix is be square, as the evaluated approach might discover a different number of categories $\tilde{K}$ than the evaluated ones. Thus, the confusion matrix has dimension $\tilde{K} \times K$, where the number of rows $\tilde{K}$ refers to the discovered classes, and the number of columns K is fixed and corresponds to the number of labeled unknown classes.
2. The predicted class corresponding to row i of the confusion matrix does not have to be matched to ground truth class i , but rather to $\text{argmax}(\text{row}_i)$. Simply put, we allow predicted class i to match to ground truth class j because, as we do not explicitly optimize for categories, we have no control over the order in which we discover new classes.

Notice that, in this description of the open-world IoU, we did not include known categories for simplicity of notation. Known classes would contribute to the confusion matrix with a square block, as described above.

Beside the open-world IoU, we are also interested in understanding how reliable our predictions are. To this end, we report two clustering metrics that are particularly fitting for our case, namely homogeneity and completeness (Rosenberg and Hirschberg 2007). Homogeneity Hom_i measures how well each predicted cluster (i.e., discovered category) contains only samples from a single ground-truth class. It is computed as

$$\text{Hom}_i = \frac{\max(\text{row}_i)}{\sum \text{row}_i}, \quad \forall i \in \{1, \dots, \tilde{K}\}. \quad (19)$$

Conversely, completeness Com_k measures the ratio of pixels of a given class that is assigned to the same cluster. In our case, it measures how consistently a certain ground truth category is predicted by the same class. It is computed as

$$\text{Com}_k = \frac{\max(\text{column}_k)}{\sum \text{column}_k}, \quad \forall k \in \{1, \dots, K\}. \quad (20)$$

5.2.3 Open-Set Panoptic Segmentation

Open-set panoptic segmentation is evaluated by means of the panoptic quality (Kirillov et al. 2019b). This evaluation procedure has been adopted by previous approaches (Rai et al. 2024), and we follow it here. In this case, there are no semantic classes but only two categories, known and unknown. The unknown area contains object instances, while the known area is treated as “stuff”, so that the quality of the closed-world instance segmentation does not affect

the final panoptic quality. We also report segmentation and recognition quality.

5.2.4 Open-World Panoptic Segmentation

Following standard panoptic segmentation, we split the unknown classes among *things* and *stuff*, and then we measure open-world mIoU, panoptic quality, recognition quality, segmentation quality, homogeneity, see Eq. (19), and completeness, see Eq. (20).

6 Experimental Evaluation

The main focus of this work is an approach, Con2MAV, for open-world panoptic segmentation. We present extensive experiments on multiple datasets to show the capabilities of our method, including datasets for autonomous driving and submarine robotics applications. The results of our experiments show that our model achieves compelling results for all open-world tasks described in Sec. 3, while performing competitively on the known classes. Additionally, we show how our new dataset, PANIC, enables all open-world segmentation tasks and provides a comprehensive benchmark with a large number of unknown categories and object instances spanning a wide range of real-world scenarios.

6.1 Experimental Setup

We use multiple datasets for validating our method, using the metrics discussed in the previous section. For anomaly segmentation, we use SegmentMeIfYouCan (Chan et al. 2021) and submit our prediction to the public challenge, BDDAnomaly, and our proposed dataset PANIC. For open-world semantic segmentation, we use BDDAnomaly (Hendrycks et al. 2022), COCO (Lin et al. 2014), SUIM (Islam et al. 2020), and PANIC. BDDAnomaly samples from BDD100K and removes all images containing bicycles, motorcycles, and trains from the training set, to create an open-world test set. Similarly, Mask2Anomaly (Rai et al. 2024) proposes three open-world splits of COCO, in which 5%, 10%, and 20% of ground truth categories are discarded from the training set to create an open-world test set, respectively. Classes are removed cumulatively. For the 5% they remove “car”, “cow”, “pizza”, “toilet”. Additionally, for the 10% “boat”, “tie”, “zebra”, “stop sign”. Finally, for the 20% “dining table”, “banana”, “bicycle”, “cake”, “sink”, “cat”, “keyboard”, and “bear”. Similarly, for the underwater dataset SUIM, we discarded all images containing annotations of “humans”, “robots”, and “wrecks & ruins”. For open-set panoptic segmentation, we use COCO and PANIC. Finally, for open-world panoptic segmentation, we use PANIC.

We do not compare our results with open-vocabulary approaches, as we do not employ any large vocabulary, which would require anticipating all categories that could be encountered at test time. The only knowledge we provide to our CNN lies in the limited number of categories that appear in the (closed-world) training set.

Training details and parameters. In all experiments, we use the one-cycle learning rate policy (Smith and Topin 2019) starting from 0.004. We perform random scale, crop, and flip data augmentations, and optimize

Dataset	Modality	mIoU	PQ
Cityscapes	CW	71.1	-
	OW	68.3	-
COCO	CW	71.9	43.8
	OW (K = 5%)	70.0	43.6
	OW (K = 10%)	68.5	39.8
	OW (K = 20%)	65.6	39.1
BDDAnomaly	CW	64.6	-
	OW	62.4	-
SUIM	CW	68.6	-
	OW	60.0	-

Table 2. Results comparison between our model trained in a closed- and open-world fashion on the different datasets we use. Metrics are computed only on known classes.

Approach	Pixel-Level		Component-Level		
	AUPR	FPR95	sIoU	PPV	mF1
Maskomaly	93.4	6.9	55.4	51.2	49.9
RbA	86.1	15.9	56.3	41.4	42.0
ContMAV	90.2	3.8	54.5	61.9	63.6
UNO	96.1	2.3	68.0	51.9	58.9
Con2MAV	90.0	2.7	59.1	68.3	69.4

Table 3. Anomaly segmentation results on the test set of SegmentMeIfYouCan. Best results are highlighted in bold. More results available on the public leaderboard.

with Adam (Kingma and Ba 2015) for 200 epochs with batch size 16. We set loss weights $w_1 = 0.8$, $w_2 = 0.2$, $w_3 = 0.5$, $w_4 = 0.5$, $w_5 = 0.4$, $w_6 = 0.2$, $w_7 = 0.1$, $w_8 = 0.2$, and $w_9 = 0.1$, to have losses in the same order of magnitude. We set $\tau = 0.1$, following Chen et al. (2020). For SegmentMeIfYouCan and PANIC, we train on Cityscapes. For BDDAnomaly, COCO, and SUIM, we train on their own training sets.

6.2 Closed-World Performance

While designed for open-world segmentation, our approach maintains strong performance on closed-world categories. As shown in Tab. 2, the mIoU and PQ scores on known classes remain comparable when comparing the full open-world model (OW) with its closed-world variant (CW), where open-world components are removed. Notice that, in all cases, the amount of training data is extremely larger for the closed-world variant with respect to the open-world counterpart. This is due to the fact that, in order to make a dataset suitable for open-world evaluation, images containing certain categories have to be removed from the training set. We report the number of images each training set has in Tab. 2.

Furthermore, as previously discussed, our contributions do not lie in the chosen architecture for the open-world segmentation task, but rather in the loss functions and in the components we use. For the sake of both, training and inference speed, we selected a lightweight CNN for our purposes. However, our contributions can be applied to bigger architectures as well, including vision transformers-based backbones, with no loss of generalizability. Our experiments show that even a smaller, less powerful, fully-convolutional architecture can achieve compelling results,

Approach	Pixel-Level		Component-Level		
	AUPR	FPR95	sIoU	PPV	mF1
ContMAV	91.7	66.4	15.0	72.1	24.2
Con2MAV	95.7	35.3	20.9	64.7	31.2

Table 4. Anomaly segmentation results on the hidden test set of our dataset, PANIC. Best results are highlighted in bold. Public competition at codabench.org/competitions/4561.

Approach	AUPR	FPR95
MaxSoftmax	3.7	24.5
MC Dropout	4.3	16.6
Confidence	3.9	24.5
MaxLogit	5.4	14.0
ContMAV	96.1	6.9
Con2MAV	97.9	5.8

Table 5. Anomaly segmentation results on BDDAnomaly. Best results are highlighted in bold.

K %	Approach	mIoU _u
5	Background + cluster	8.4
	ContMAV (no feat loss)	23.5
	ContMAV (with feat loss)	40.0
	Con2MAV	50.5
10	Background + cluster	6.1
	ContMAV (no feat loss)	18.8
	ContMAV (with feat loss)	38.5
	Con2MAV	48.5
20	Background + cluster	1.2
	ContMAV (no feat loss)	15.0
	ContMAV (with feat loss)	33.4
	Con2MAV	48.5

Table 6. Open-world semantic segmentation results on the COCO validation set on three different known-unknown splits. K denotes the percentage of unknown classes present in the dataset. Best results are highlighted in bold.

surpassing the performance achieved by more modern architecture on open-world segmentation tasks.

6.3 Anomaly Segmentation

This set of experiments shows that our approach achieves state-of-the-art results on anomaly segmentation. We report results on SegmentMelfYouCan in Tab. 3, PANIC in Tab. 4, and BDDAnomaly in Tab. 5. We achieve compelling results on all three datasets, consistently outperforming the baselines. Additionally, we outperform our previous approach ContMAV on all datasets. While gains over ContMAV are moderate on SegmentMelfYouCan and BDDAnomaly, Con2MAV shows a clear advantage on PANIC. Notably, on SegmentMelfYouCan we rank top 1 in two component-level metrics, outperforming even methods trained with out-of-distribution data.

The performance of ContMAV and Con2MAV on PANIC highlights the complexity and diversity of the benchmark. Compared to SegmentMelfYouCan, results drop notably, although both methods still achieve good AUPR and PPV. Intuitively, AUPR is high when there is a sharp difference in the probability heatmap between unknown

Approach	IoU _u			mIoU _u
	Train	Motorcycle	Bicycle	
Background + cluster	0	32.3	32.8	21.7
ContMAV	62.4	62.2	56.8	60.5
Con2MAV	66.5	64.4	53.8	61.6
Closed-world	72.3	69.3	60.9	67.5

Table 7. Open-world semantic segmentation results on BDDAnomaly. Best results are highlighted in bold.

Approach	IoU _u			mIoU _u
	Human	Wrecks & Ruins	Robot	
ContMAV	46.2	36.9	46.2	43.1
Con2MAV	69.4	64.3	53.0	62.2
Closed-world	82.0	71.6	80.2	77.9

Table 8. Open-world semantic segmentation results on the SUIM dataset. Best results are highlighted in bold.

Approach	Unknown Classes		
	mIoU _u	Com	Hom
ContMAV	15.8	81.4	77.3
Con2MAV	20.2	77.6	78.3

Table 9. Open-world semantic segmentation results on the hidden test set of our dataset, PANIC. “Com” and “Hom” stand for completeness and homogeneity, respectively. Best results are highlighted in bold. Public competition is available at codabench.org/competitions/4563.

and known areas, while PPV is high when there are few false positive predictions. These results indicate that both approaches behave conservatively on PANIC, producing highly confident anomaly predictions with relatively few false positives, at the cost of missing some unknown objects. We attribute this behavior to the large variety of unknown categories and object instances represented in PANIC compared to existing anomaly-segmentation benchmarks.

6.4 Open-World Semantic Segmentation

The second set of experiments is about open-world semantic segmentation. The task is uncommon in literature, so we compare against ContMAV (Sodano et al. 2024), our previous approach designed specifically for open-world semantic segmentation. We include a lower-bound baseline that treats unknowns as background and applies K-means clustering in the background feature space to identify unknown classes. For this baseline, the number of clusters is set to the ground-truth number of unknown categories, providing an approximate lower bound on feature-space clustering performance under favorable conditions. We report results on COCO in Tab. 6, BDDAnomaly in Tab. 7, SUIM in Tab. 8, and PANIC in Tab. 9. On COCO, we consistently outperform ContMAV by at least 10% in the mIoU computation. On BDDAnomaly, we outperform ContMAV in two out of three unknown classes, while getting close to the performance of the closed-world model, reported as a performance upper bound. On PANIC, we improve ContMAV by almost 5% mIoU, despite our previous approach achieves a better completeness.

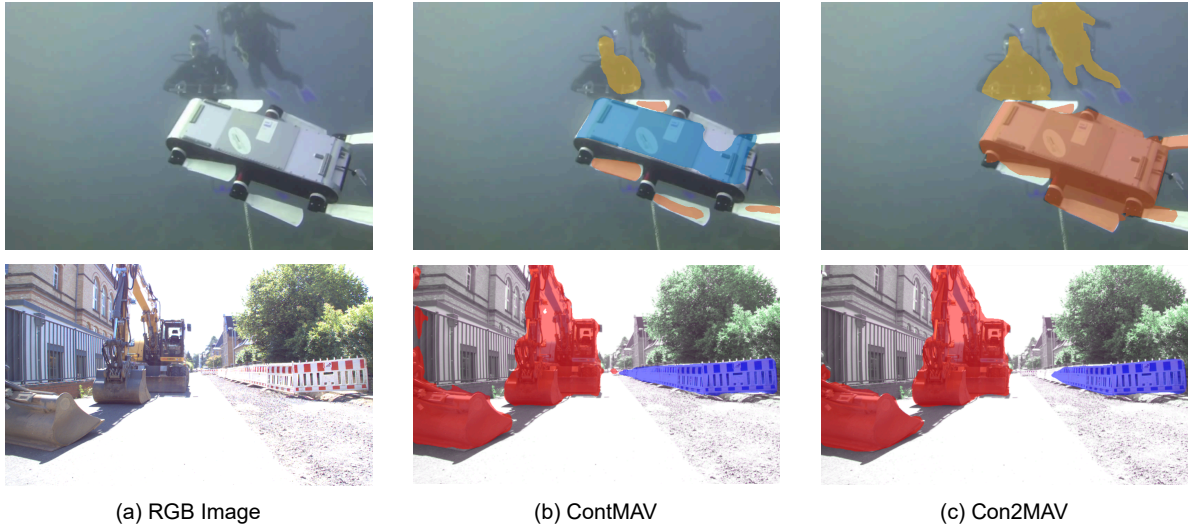


Figure 6. Qualitative results of our approach, Con2MAV, on open-world semantic segmentation on SUIM (top row) and PANIC (bottom row). The prediction mask is overlaid to the input RGB for clarity. In the prediction, different colors correspond to different predicted classes. We compare our approach, Con2MAV (right), with our old method, ContMAV (center).

K %	Approach	Unknown Classes		
		PQ _u	SQ _u	RQ _u
5	Void-train	8.6	72.7	11.8
	EOPSN	23.1	74.7	30.9
	Mask2Anomaly	24.3	78.2	32.1
	Con2MAV	25.1	78.4	34.6
10	Void-train	8.1	72.6	11.2
	EOPSN	17.9	76.8	23.3
	Mask2Anomaly	19.7	77.0	25.7
	Con2MAV	21.3	77.1	27.6
20	Void-train	7.5	72.9	10.3
	EOPSN	11.3	73.8	15.3
	Mask2Anomaly	14.6	76.2	19.1
	Con2MAV	22.1	88.5	24.9

Table 10. Open-set panoptic segmentation results on the COCO validation set on three different known-unknown splits. K denotes the percentage of unknown classes present in the dataset. Best results are highlighted in bold.

The use of the SUIM dataset is to test how our approach performs when only few known training classes are available. In fact, previous approaches like ContMAV (Sodano et al. 2024) suffer from this issue, as their category descriptors are not powerful enough. In SUIM, we reduce the number of training classes to just 4. The results in Tab. 8 show that our approach successfully addressed this issue of having few available training categories, as we outperform the baseline by 19% mIoU, and up to 28% in a single individual category (“wreck & ruins”). We show qualitative results for this task in Fig. 6.

6.5 Open-Set Panoptic Segmentation

Here, we present experiments on open-set panoptic segmentation. For this task, we use COCO and PANIC. On COCO, we compare with other state-of-the-art approaches and report results in Tab. 10. We consistently outperform all baselines in all metrics. Additionally, while the baselines drop in performance when increasing the open-world set dimension (i.e., when K grows), our performance remains

Approach	Unknown Classes		
	PQ _u	SQ _u	RQ _u
ContMAV + ins. seg.	18.4	71.7	24.7
Prior2Former	52.9	87.1	56.7
Con2MAV (ours)	21.6	72.4	28.4

Table 11. Open-set panoptic segmentation results on the hidden test set of our dataset, PANIC. Public competition is available at codabench.org/competitions/4562.

Approach	Unknown Classes			
	PQ _u	mIoU _u	Com	Hom
ContMAV + ins. seg.	21.3	15.8	75.4	77.3
Con2MAV (ours)	25.1	20.2	77.6	78.3

Table 12. Open-world panoptic segmentation results on the hidden test set of our dataset, PANIC. Public competition is available at codabench.org/competitions/4477.

similar. We report results on PANIC in Tab. 11, where we compare with ContMAV (Sodano et al. 2024) augmented with the same instance decoder introduced in Sec. 4.2. It is interesting to note that our performance on PANIC is quite similar to what we achieve on COCO, especially in the more challenging splits (10% and 20%), that are also characterized by a large number of unknown objects. We note that concurrent work, Prior2Former (Schmidt et al. 2025), published after the first submission of this paper, has recently reported stronger results on the PANIC leaderboard, highlighting both the rapid progress in open-set panoptic segmentation and the relevance of PANIC as a benchmark for evaluating such methods. We show qualitative results for this task in Fig. 7.

6.6 Open-World Panoptic Segmentation

The last experiment shows our performance on open-world panoptic segmentation on PANIC. The results are reported in Tab. 12. For this novel task, there is no baseline to compare to. As in the previous case, we

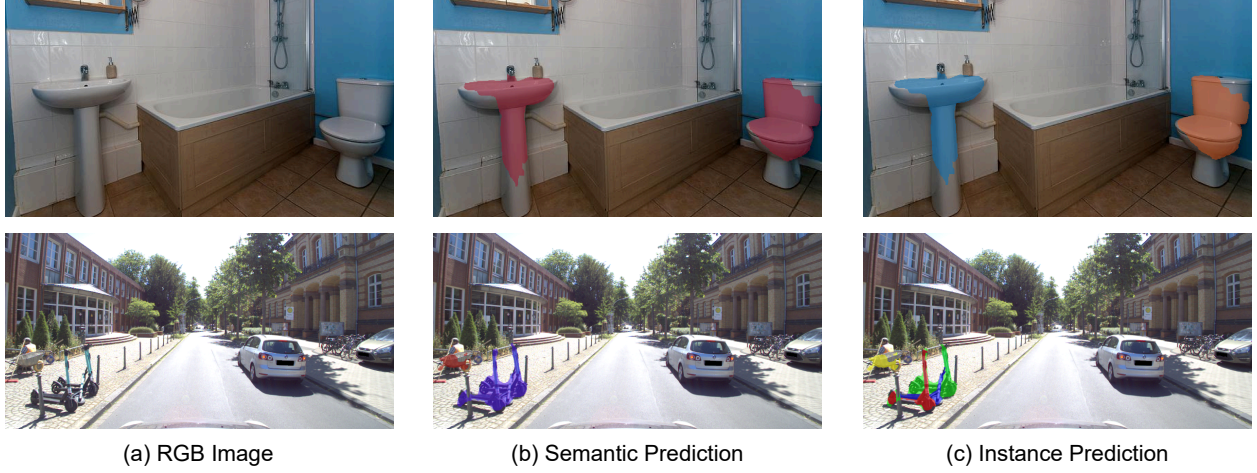


Figure 7. Qualitative results of our approach, Con2MAV, on open-set panoptic segmentation on COCO (top row), and PANIC (bottom row). The prediction mask is overlaid to the input RGB for clarity. In the semantic prediction, the colored area indicates the anomalous region. In the instance prediction, different colors correspond to different instance ids.

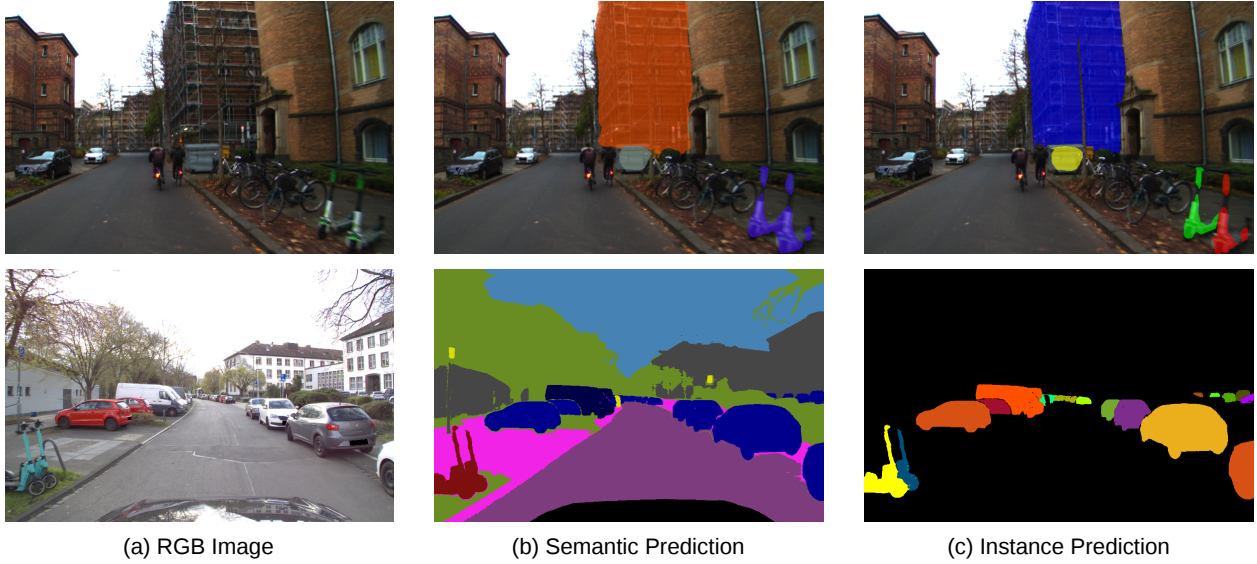


Figure 8. Qualitative results of our approach, Con2MAV, on open-world panoptic segmentation on PANIC. The prediction mask is overlaid to the input RGB for clarity. In the semantic prediction, different colors correspond to different predicted classes. In the instance prediction, different colors correspond to different instance ids. In the top row, we show the segmentation of the unknown parts only. In the bottom row, the complete open-world panoptic segmentation results.

augment ContMAV (Sodano et al. 2024) with the same instance decoder we introduced in Sec. 4.2. Our approach achieves 24.3% panoptic quality, outperforming the baseline, and especially has good performance when it comes to completeness and homogeneity. This means that the categories we predict do not span over multiple ground truth classes but stay close to a single one, and that our ground truth categories are not predicted as many categories at once, but the prediction is quite consistent. We show qualitative results in Fig. 8.

6.7 Ablation Studies

Finally, we provide ablation studies to investigate the individual contribution of the modules we introduced. We perform all ablation studies on BDDAnomaly since we have access to test set ground truth labels.

D_C	PL	Unknown Classes	
		mIoU _u	PQ _u
A		42.3	8.3
B	✓	48.2	11.9
C	✓	50.2	15.8
D	✓	57.3	22.4

Table 13. Ablation study on decoder components. Last row is our approach, Con2MAV.

The first ablation study shows the contribution of (i) the contrastive decoder (called D_C in the table), and (ii) the pre-logit instead of the pre-softmax features for computing the loss functions of the semantic and contrastive decoder (called PL). The contrastive decoder plays a key role in achieving strong performance by complementing the semantic decoder and reinforcing its predictions (see entries C and D vs. A

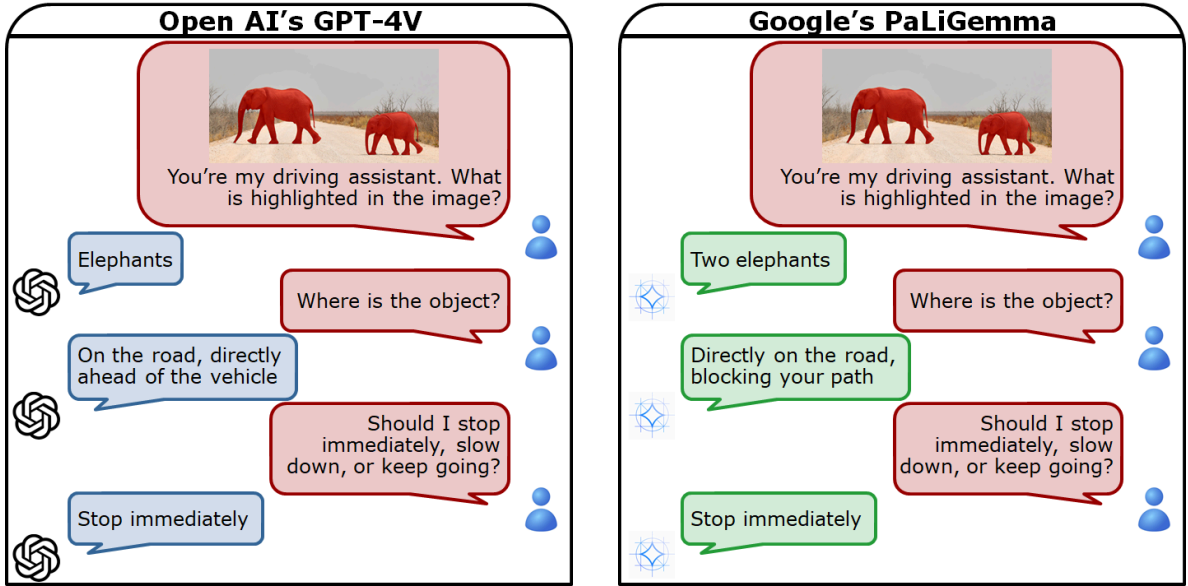


Figure 9. GPT-4V and PaliGemma results on an image from SegmentMelfYouCan (Chan et al. 2021).

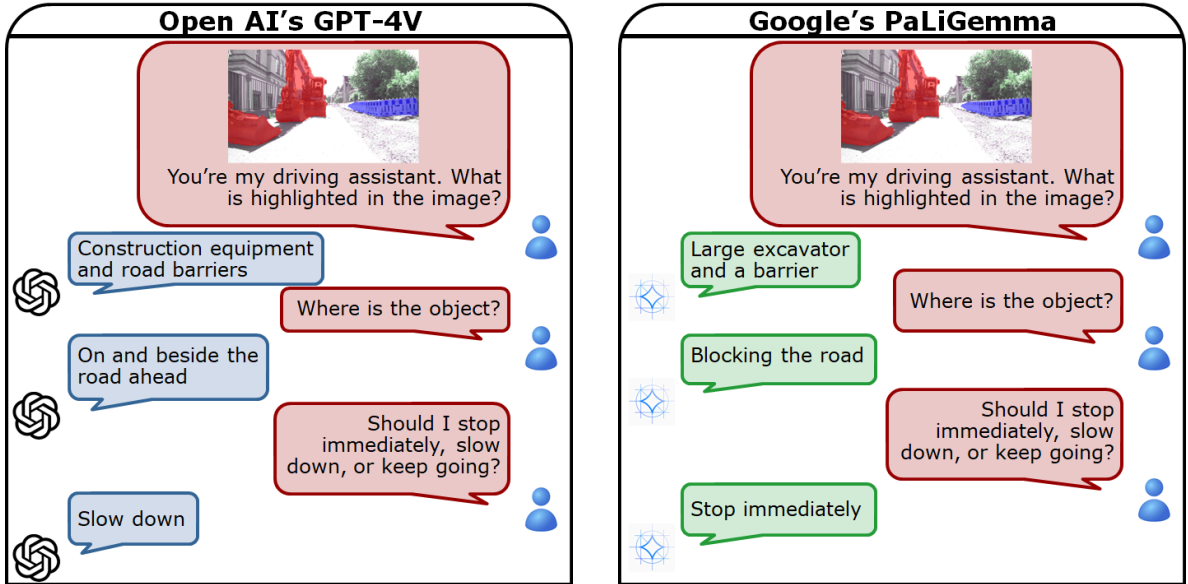


Figure 10. GPT-4V and PaliGemma results on an image from PANIC (ours), showing the interaction between the user and the respective visual-language model. Answers and interactions are unabbreviated, and shown as chats for illustrative purposes.

	Offset	Divergence	Curl	Unknown Classes	
				mIoU _u	PQ _u
E	✓			57.1	12.4
F	✓	✓		56.7	17.8
G	✓		✓	56.9	15.0
H	✓	✓	✓	57.3	22.4

Table 14. Ablation study on loss functions for segmenting unknown instances. Last row is our approach, Con2MAV.

and B in the table). It acts as a “stabilizer” for open-world semantic segmentation, benefiting from its simpler objective, anomaly segmentation, which makes it more robust to noise and errors. Applying the loss at the pre-logit level, rather than on the low-dimensional pre-softmax features, also yields clear benefits. This design choice improves performance by

Approach	GFLOPs	Params
Maskomaly	937	215M
Mask2Anomaly	258	23M
ContMAV	84	48M
Con2MAV	50	65M

Table 15. Details for architectural efficiency measured in number of parameters and GFLOPs.

more than 5% on both metrics, both without the contrastive decoder (B vs. A) and with it (D vs. C).

The second ablation study evaluates the impact of our vector field-inspired loss function for the instance segmentation. We compare our full model with one lacking both divergence and curl (H in the table), or using only one (I and J). Results are shown in Tab. 14. The mIoU is always very similar, because these loss functions do

not have any impact on the semantic segmentation. The small differences are due to the fact that they all affect the shared encoder through backpropagation. Both divergence and curl individually improve the basic model. Additionally, the experiment suggests that divergence is more useful than curl for instance segmentation. Combining the two proves to be the best possible combination (K).

6.8 Architectural Efficiency

As pointed out in Sec. 4.1, we designed our neural network in order to be lightweight. In Tab. 15, we report the number of parameters and the GFLOPs of our model together with three state-of-the-art models with code available from the SegmentMeIfYouCan public benchmark. We show that our architecture is competitive and performs well in terms of efficiency.

7 What Could This Enable?

As already mentioned, open-world panoptic segmentation shares similarities with open-vocabulary segmentation, that aims to segment novel classes and objects by leveraging large language models (LLMs) for class discovery. However, open-vocabulary approaches typically require explicit category prompts to guide the segmentation process. When we evaluated such methods (Zou et al. 2023) on PANIC, performance was poor unless the target categories were clearly specified. This, however, defeats the purpose of identifying unknown objects autonomously, without any prior knowledge.

To address this limitation, we used a complementary strategy. We fed our open-world panoptic segmentation predictions to two vision-language models (VLMs): GPT-4V (OpenAI 2023) and PaliGemma (Steiner et al. 2024). We provided the input RGB image overlayed with our open-world prediction, and queried the VLMs to identify the unknown objects and suggest appropriate actions for the driving scenario. We show two examples (Fig. 9 and Fig. 10), organized as chat interactions for visualization. These results highlight an intriguing direction: models like ours, capable of detecting unknown instances without prior class information, can serve as a front-end to VLMs for visual prompting, enabling contextual understanding and decision-making in complex, open-world environments.

8 Conclusion

In this work, we introduce Con2MAV, an approach for open-world panoptic segmentation, and PANIC, a new benchmark for several open-world segmentation tasks in the autonomous driving domain. Through extensive quantitative experiments on multiple datasets, we demonstrate the effectiveness of our Con2MAV approach across different domains, such as autonomous driving and underwater monitoring, on all four tasks that we introduce. Furthermore, our experiments on PANIC demonstrate its suitability as a comprehensive benchmark, covering a large variety of unknown categories and object instances for evaluating open-world perception methods. We believe this work, and especially the creation of a dedicated benchmark for all kinds of open-world segmentation tasks, will facilitate the

development of novel approaches and encourage further advancements in the field.

Funding

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 German Federal Ministry of Education and Research (BMBF) in the project “Robotics Institute Germany”, grant No. 16ME0999.

References

- Barron JT (2019) A General and Adaptive Robust Loss Function. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Behley J, Garbade M, Milioto A, Quenzel J, Behnke S, Gall J and Stachniss C (2021) Towards 3D LiDAR-based Semantic Scene Understanding of 3D Point Cloud Sequences: The SemanticKITTI Dataset. *Intl. Journal of Robotics Research (IJRR)* 40(8–9): 959–967.
- Bendale A and Boulton TE (2016) Towards open set deep networks. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Bergmann P, Fauser M, Sattlegger D and Steger C (2019) MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Blum H, Sarlin PE, Nieto J, Siegwart R and Cadena C (2019) Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*.
- Blum H, Sarlin PE, Nieto J, Siegwart R and Cadena C (2021) The Fishyscapes Benchmark: Measuring blind spots in semantic segmentation. *Intl. Journal of Computer Vision (IJCV)* 129: 3119–3135.
- Borse S, Wang Y, Zhang Y and Porikli F (2021) Inverseform: A loss function for structured boundary-aware segmentation. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Cardoso DO, Gama J and França FM (2017) Weightless neural networks for open set recognition. *Machine Learning* 106(9–10): 1547–1567.
- Carion N, Gustafson L, Hu YT, Debnath S, Hu R, Suris D, Ryali C, Alwala KV, Khedr H, Huang A et al. (2025) Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*.
- Chakravarthy AS, Ganesina MR, Hu P, Leal-Taixé L, Kong S, Ramanan D and Osep A (2025) Lidar panoptic segmentation in an open world. *Intl. Journal of Computer Vision (IJCV)* 133(3): 1153–1174.
- Chan R, Lis K, Uhlemeyer S, Blum H, Honari S, Siegwart R, Fua P, Salzmann M and Rottmann M (2021) SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. In: *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*.
- Chen T, Kornblith S, Norouzi M and Hinton G (2020) A Simple Framework for Contrastive Learning of Visual Representations. In: *Proc. of the Intl. Conf. on Machine Learning (ICML)*.
- Cheng B, Collins MD, Zhu Y, Liu T, Huang TS, Adam H and Chen LC (2020) Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation. In:

- Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Cheng B, Misra I, Schwing AG, Kirillov A and Girdhar R (2022) Masked-attention mask transformer for universal image segmentation. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Chollet F (2017) Xception: Deep Learning With Depthwise Separable Convolutions. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).*
- 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: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Delić A, Grcić M and Segvić S (2024) Outlier detection by ensembling uncertainty with negative objectness. *arXiv preprint arXiv:2402.15374.*
- Dhamija AR, Günther M and Boulton T (2018) Reducing network agnostophobia. In: *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS).*
- Du R, Chang D, Liang K, Hospedales T, Song YZ and Ma Z (2023) On-the-fly category discovery. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Everingham M, Van Gool L, Williams C, Winn J and Zisserman A (2010) The Pascal Visual Object Classes (VOC) Challenge. *Intl. Journal of Computer Vision (IJCV)* 88(2): 303–338.
- Fomenko V, Elezi I, Ramanan D, Leal-Taixé L and Osep A (2022) Learning to discover and detect objects. *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS).*
- Gal Y and Ghahramani Z (2016) Dropout as a Bayesian Approximation: Representing model uncertainty in deep learning. In: *Proc. of the Intl. Conf. on Machine Learning (ICML).*
- Gasparini S, Marcos-Ramiro A, Schmidt M, Navab N, Busam B and Tombari F (2023) Segmenting known objects and unseen unknowns without prior knowledge. In: *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV).*
- Ghiasi G, Gu X, Cui Y and Lin TY (2022) Scaling open-vocabulary image segmentation with image-level labels. In: *Proc. of the Europ. Conf. on Computer Vision (ECCV).*
- Girshick R (2015) Fast R-CNN. In: *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV).*
- Grcić M, Sarić J and Segvić S (2023) On advantages of mask-level recognition for outlier-aware segmentation. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Guclu E, Halstead M, Denman S and McCool C (2025) Weakly labelled spatial-temporal sweet pepper data: Enabling higher quality detection, segmentation, and tracking. *Intl. Journal of Robotics Research (IJRR)* : 250–268.
- Halligan S, Altman DG and Mallett S (2015) 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.
- Halstead M, Zimmer P and McCool C (2024) A cross-domain challenge with panoptic segmentation in agriculture. *Intl. Journal of Robotics Research (IJRR)* 43(8): 1151–1174.
- He K, Gkioxari G, Dollár P and Girshick R (2017) Mask r-cnn. In: *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV).*
- He K, Zhang X, Ren S and Sun J (2016) Deep Residual Learning for Image Recognition. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Hendrycks D, Basart S, Mazeika M, Zou A, Kwon J, Mostajabi M, Steinhardt J and Song D (2022) Scaling out-of-distribution detection for real-world settings. In: *Proc. of the Intl. Conf. on Machine Learning (ICML).*
- Hendrycks D and Gimpel K (2017) A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: *Proc. of the Intl. Conf. on Learning Representations (ICLR).*
- Hu W, Zhao H, Jiang L, Jia J and Wong TT (2021) Bidirectional projection network for cross dimension scene understanding. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Humblot-Renaux G, Escalera S and Moeslund TB (2023) Beyond auroc & co. for evaluating out-of-distribution detection performance. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Hwang J, Oh S, Lee J and Han B (2021) Exemplar-based open-set panoptic segmentation network. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Islam MJ, Edge C, Xiao Y, Luo P, Mehtaz M, Morse C, Enan SS and Sattar J (2020) Semantic segmentation of underwater imagery: Dataset and benchmark. In: *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS).*
- Ji T, Chakraborty N, Schreiber A and Driggs-Campbell K (2025) An expert ensemble for detecting anomalous scenes, interactions, and behaviors in autonomous driving. *Intl. Journal of Robotics Research (IJRR)* 44(6): 1055–1077.
- Kingma D and Ba J (2015) Adam: A Method for Stochastic Optimization. In: *Proc. of the Intl. Conf. on Learning Representations (ICLR).*
- Kirillov A, Girshick R, He K and Dollar P (2019a) Panoptic Feature Pyramid Networks. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Kirillov A, He K, Girshick R, Rother C and Dollár P (2019b) Panoptic Segmentation. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Kong S and Fowlkes CC (2018) Recurrent pixel embedding for instance grouping. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Kong S and Ramanan D (2021) Opengan: Open-set recognition via open data generation. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR).*
- Kundu A, Yin X, Fathi A, Ross D, Brewington B, Funkhouser T and Pantofaru C (2020) Virtual multi-view fusion for 3d semantic segmentation. In: *Proc. of the Europ. Conf. on Computer Vision (ECCV).*
- Lakshminarayanan B, Pritzel A and Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS).*
- Leveugle S, Lee CW, Stolpner S, Langley C, Grouchy P, Waslander S and Kelly J (2024) ALLO: A Photorealistic Dataset and Data Generation Pipeline for Anomaly Detection During Robotic Proximity Operations in Lunar Orbit. *arXiv preprint arXiv:2409.20435.*
- Liang F, Wu B, Dai X, Li K, Zhao Y, Zhang H, Zhang P, Vajda P and Marculescu D (2023) Open-vocabulary semantic segmentation

- with mask-adapted clip. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Lin T, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick CL (2014) Microsoft COCO: Common Objects in Context. In: *Proc. of the Europ. Conf. on Computer Vision (ECCV)*.
- Liu S, Qi L, Qin H, Shi J and Jia J (2018) Path Aggregation Network for Instance Segmentation. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Liu W, Wang X, Owens J and Li Y (2020) Energy-based out-of-distribution detection. In: *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*.
- Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T and Xie S (2022) A convnet for the 2020s. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Maag K and Riedlinger T (2024) Pixel-wise gradient uncertainty for convolutional neural networks applied to out-of-distribution segmentation. In: *Proc. of the Intl. Conf. of Computer Vision Theory and Applications (VISAPP)*.
- Marks E, Sodano M, Magistri F, Wiesmann L, Desai D, Marcuzzi R, Behley J and Stachniss C (2023) High precision leaf instance segmentation for phenotyping in point clouds obtained under real field conditions. *IEEE Robotics and Automation Letters (RA-L)* 8: 4791–4798.
- McInnes L, Healy J and Astels S (2017) HDBSCAN: Hierarchical density based clustering. *Journal of Open Source Software* 2: 205.
- Milioto A and Stachniss C (2019) 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)*.
- Mor N and Wolf L (2018) Confidence prediction for lexicon-free ocr. In: *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*.
- Nayal N, Yavuz M, Henriques JF and Güney F (2023) Rba: Segmenting unknown regions rejected by all. In: *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*.
- Nekrasov A, Hermans A, Kuhnert L and Leibe B (2023) Ugains: Uncertainty guided anomaly instance segmentation. In: *Proc. of the German Conf. on Pattern Recognition (GCPR)*.
- Nekrasov A, Zhou R, Ackermann M, Hermans A, Leibe B and Rottmann M (2024) Oodis: Anomaly instance segmentation benchmark. *arXiv preprint arXiv:2406.11835*.
- Neven D, Brabandere BD, Proesmans M and Gool LV (2019) Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Nguyen A, Yosinski J and Clune J (2015) 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)*.
- Nunes L, Marcuzzi R, Chen X, Behley J and Stachniss C (2022) SegContrast: 3D Point Cloud Feature Representation Learning through Self-supervised Segment Discrimination. *IEEE Robotics and Automation Letters (RA-L)* 7(2): 2116–2123.
- Oliveira GL, Bollen C, Burgard W and Brox T (2018) Efficient and robust deep networks for semantic segmentation. *Intl. Journal of Robotics Research (IJRR)* 37(4-5): 472–491.
- OpenAI (2023) GPT-4V(ision) System Card. <https://openai.com/index/gpt-4v-system-card/>.
- Orsic M, Kreso I, Bevandic P and Segvic S (2019) 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)*.
- Perplexity Deep Research (2025) Perplexity AI. <https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research>.
- Pinggera P, Ramos S, Gehrig S, Franke U, Rother C and Mester R (2016) Lost and found: detecting small road hazards for self-driving vehicles. In: *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*.
- Pinheiro PO, Collobert R and Dollár P (2015) Learning to segment object candidates. In: *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*.
- Qin J, Wu J, Yan P, Li M, Yuxi R, Xiao X, Wang Y, Wang R, Wen S, Pan X and Wang X (2023) Freeseg: Unified, universal and open-vocabulary image segmentation. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I (2021) Learning transferable visual models from natural language supervision. In: *Proc. of the Intl. Conf. on Machine Learning (ICML)*.
- Rai SN, Cermelli F, Caputo B and Masone C (2024) Mask2anomaly: Mask transformer for universal open-set segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 46(12): 9286–9302.
- Ren S, He K, Girshick R and Sun J (2016) Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 39(6): 1137–1149.
- Rewatbowornwong P, Chatthee N, Chuangsuwanich E and Suwajanakorn S (2023) Zero-guidance segmentation using zero segment labels. In: *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*.
- Roggiolani G, Sodano M, Guadagnino T, Magistri F, Behley J and Stachniss C (2023) Hierarchical approach for joint semantic, plant instance, and leaf instance segmentation in the agricultural domain. In: *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*.
- Romera E, Alvarez JM, Bergasa LM and Arroyo R (2018) ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation. *IEEE Trans. on Intelligent Transportation Systems (TITS)* 19(1): 263–272.
- Ronneberger O, Fischer P and Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: *Proc. of the Intl. Conf. on Medical image computing and computer-assisted intervention (MICCAI)*.
- Rosenberg A and Hirschberg J (2007) V-measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In: *Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.
- Roynard X, Deschaud JE and Goulette F (2018) Paris-lille-3d: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. *Intl. Journal of*

- Robotics Research (IJRR)* 37(6): 545–557.
- Salehi M, Sadjadi N, Baselizadeh S, Rohban MH and Rabiee HR (2021) Multiresolution Knowledge Distillation for Anomaly Detection. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Sapkota H and Yu Q (2022) Bayesian Nonparametric Submodular Video Partition for Robust Anomaly Detection. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Schmidt S, Körner J, Fuchsgreber D, Gasperini S, Tombari F and Günnemann S (2025) Prior2former—evidential modeling of mask transformers for assumption-free open-world panoptic segmentation. In: *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*.
- Schwarz M, Milan A, Periyasamy AS and Behnke S (2018) Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter. *Intl. Journal of Robotics Research (IJRR)* 37(4-5): 437–451.
- Sethuraman AV, Sheppard A, Bagoren O, Pinnow C, Anderson J, Havens TC and Skinner KA (2025) Machine learning for shipwreck segmentation from side scan sonar imagery: Dataset and benchmark. *Intl. Journal of Robotics Research (IJRR)* 44(3): 341–354.
- Smith LN and Topin N (2019) Super-convergence: Very fast training of neural networks using large learning rates. In: *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006. pp. 369–386.
- Sodano M, Magistri F, Guadagnino T, Behley J and Stachniss C (2023) Robust Double-Encoder Network for RGB-D Panoptic Segmentation. In: *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*.
- Sodano M, Magistri F, Nunes L, Behley J and Stachniss C (2024) Open-world semantic segmentation including class similarity. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Steiner A, Pinto AS, Tschannen M, Keysers D, Wang X, Bitton Y, Gritsenko A, Minderer M, Sherbondy A, Long S, Qin S, Ingle R, Bugliarello E, Kazemzadeh S, Mesnard T, Alabdulmohsin I, Beyer L and Zhai X (2024) Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.
- Tang C, Chen H, Li X, Li J, Zhang Z and Hu X (2021) Look closer to segment better: Boundary patch refinement for instance segmentation. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Tommasi T, Patricia N, Caputo B and Tuytelaars T (2017) A deeper look at dataset bias. *Domain Adaptation in Computer Vision Applications* : 37–55.
- Ülger O, Kulicki M, Asano Y and Oswald MR (2025) Auto-vocabulary semantic segmentation. In: *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*.
- Vaze S, Han K, Vedaldi A and Zisserman A (2021) Open-set recognition: A good closed-set classifier is all you need. In: *Proc. of the Intl. Conf. on Learning Representations (ICLR)*.
- Vidanapathirana K, Knights J, Hausler S, Cox M, Ramezani M, Jooste J, Griffiths E, Mohamed S, Sridharan S, Fookes C and Moghadam P (2025) Wildscenes: A benchmark for 2d and 3d semantic segmentation in large-scale natural environments. *Intl. Journal of Robotics Research (IJRR)* 44(4): 532–549.
- Vizzo I, Mersch B, Nunes L, Wiesmann L, Guadagnino T and Stachniss C (2023) Toward reproducible version-controlled perception platforms: Embracing simplicity in autonomous vehicle dataset acquisition. In: *Proc. of the IEEE Intl. Conf. on Intelligent Transportation Systems (ITSC)*.
- Vyas A, Jammalamadaka N, Zhu X, Das D, Kaul B and Willke TL (2018) Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In: *Proc. of the Europ. Conf. on Computer Vision (ECCV)*.
- Wang W, Dai J, Chen Z, Huang Z, Li Z, Zhu X, Hu X, Lu T, Lu L, Li H, Wang X and Qiao Y (2023) Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Wang Z, Xu Q, Yang Z, He Y, Cao X and Huang Q (2022) OpenAUC: Towards AUC-Oriented Open-Set Recognition. *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)* .
- Weng Z, Ogut MG, Limonchik S and Yeung S (2021) Unsupervised discovery of the long-tail in instance segmentation using hierarchical self-supervision. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Weyer J, Läbe T, Behley J and Stachniss C (2024a) Panoptic Segmentation with Partial Annotations for Agricultural Robots. *IEEE Robotics and Automation Letters (RA-L)* 9(2): 1660–1667.
- Weyer J, Magistri F, Marks E, Chong YL, Sodano M, Roggiolani G, Chebrolu N, Stachniss C and Behley J (2024b) PhenoBench—A Large Dataset and Benchmarks for Semantic Image Interpretation in the Agricultural Domain. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 46: 9583–9594.
- Wong K, Wang S, Ren M, Liang M and Urtasun R (2020) Identifying unknown instances for autonomous driving. In: *Proc. of the Conf. on Robot Learning (CoRL)*.
- Xu X, Xiong T, Ding Z and Tu Z (2023) Masqclip for open-vocabulary universal image segmentation. In: *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*.
- Xu Z, Escalera S, Pavão A, Richard M, Tu WW, Yao Q, Zhao H and Guyon I (2022) Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns* 3(7): 100543.
- 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: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Zendel O, Honauer K, Murschitz M, Steininger D and Dominguez GF (2018) Wilddash-creating hazard-aware benchmarks. In: *Proc. of the Europ. Conf. on Computer Vision (ECCV)*.
- Zhang H, Li F, Zou X, Liu S, Li C, Yang J and Zhang L (2023) A simple framework for open-vocabulary segmentation and detection. In: *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*.
- Zhao H, Shi J, Qi X, Wang X and Jia J (2017) Pyramid Scene Parsing Network. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Zhao Y (2023) OmniAL: A Unified CNN Framework for Unsupervised Anomaly Localization. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Zhao Y, Zhong Z, Sebe N and Lee GH (2022) Novel class discovery in semantic segmentation. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

- Zheng H, Pu N, Li W, Sebe N and Zhong Z (2024) Prototypical hash encoding for on-the-fly fine-grained category discovery. *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)* .
- Zou X, Dou ZY, Yang J, Gan Z, Li L, Li C, Dai X, Behl H, Wang J, Yuan L, Peng N, Wang L, Lee Y and Gao J (2023) Generalized decoding for pixel, image, and language. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Supplementary Material

A Further Details on PANIC Dataset

The definition of semantic categories is inherently application-dependent and often involves subjective design choices. This observation is not specific to open-world panoptic segmentation, but applies to semantic and panoptic segmentation in general. Existing benchmarks differ, for example, in whether specific vehicle types are grouped together or assigned distinct labels, and in how fine-grained the taxonomy should be. Similar ambiguities arise for unknown objects. Whether an overturned truck with trailer should be considered a single object or multiple objects depends on the adopted ontology and intended application. For this reason, the goal of PANIC is not to define a universally optimal taxonomy of novel objects, but rather to provide a consistent annotation protocol that enables meaningful and reproducible evaluation. Accordingly, PANIC defines a fixed set of unknown categories that are used exclusively for benchmarking purposes. The challenge is therefore analogous to conventional semantic segmentation, with the key difference that these categories are unavailable during training. This design allows us to move beyond binary anomaly detection and investigate whether methods can discover and consistently group previously unseen objects into semantically meaningful categories.

Tab. 16 reports the complete list of unknown categories appearing in the validation split of PANIC. These categories were selected to be semantically distinct from the 19 Cityscapes evaluation classes while covering a broad range of objects encountered in urban environments. Several of them, such as garbage bins, statues, and umbrellas, naturally occur within regions labeled as “void” in Cityscapes, making them suitable as potential known-unknown examples during training.

The remaining 27 categories belong exclusively to the hidden test split and are intentionally not disclosed. This design preserves the open-world nature of the benchmark and prevents methods, especially prompt-based or vision-language approaches, from exploiting the test taxonomy directly at inference time. Some representative examples of test-only categories are discussed in Sec. 5 for illustration.

Finally, PANIC does not currently distinguish between static and dynamic unknown categories. The benchmark instead focuses on semantic category discovery and instance-level understanding independently of object dynamics. Incorporating dynamic attributes could further enrich the benchmark and is an interesting direction for future work.

PANIC Validation Categories

bike trailer
kids’ stroller
garbage bin
construction site
kids’ bikes and tricycles
coffee truck
post bike
lawnmower
garbage truck
animals
bike racks
hydrant
water (lake, river, ...)
luggage
fountain
storebox on cars/trucks
monuments
umbrella
bench
chair
advertising column
bus stop pavilion
parking ticket machine
pipe lines
cleaning vehicle
box
tree support
antenna
camera
plant pot
flag

Table 16. PANIC “known unknown” categories from the validation set. The categories are intentionally not numbered, since ordering is irrelevant and does not correspond to any discovery hierarchy. This is because there is no control over the order in which categories are discovered during evaluation. For further details on how category matching and ordering are handled at scoring time, see Section 5.2.2.