Open-World Semantic Segmentation Including Class Similarity

Supplementary Material

A. Further Details on Experiments

A.1. Datasets and Metrics

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

A.2. Experiments on Hyperparameters

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A.3. Further Details on Anomaly Segmentation

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

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

A.4. Further Details on Class Similarity

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

Table 10. Anomaly segmentation results on BDDAnomaly*.

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

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

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

A.5. Architectural Choices

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

B. Further Details on the Contrastive Decoder

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

Table 12. Architectural Efficiency

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

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

C. Architectural Efficiency

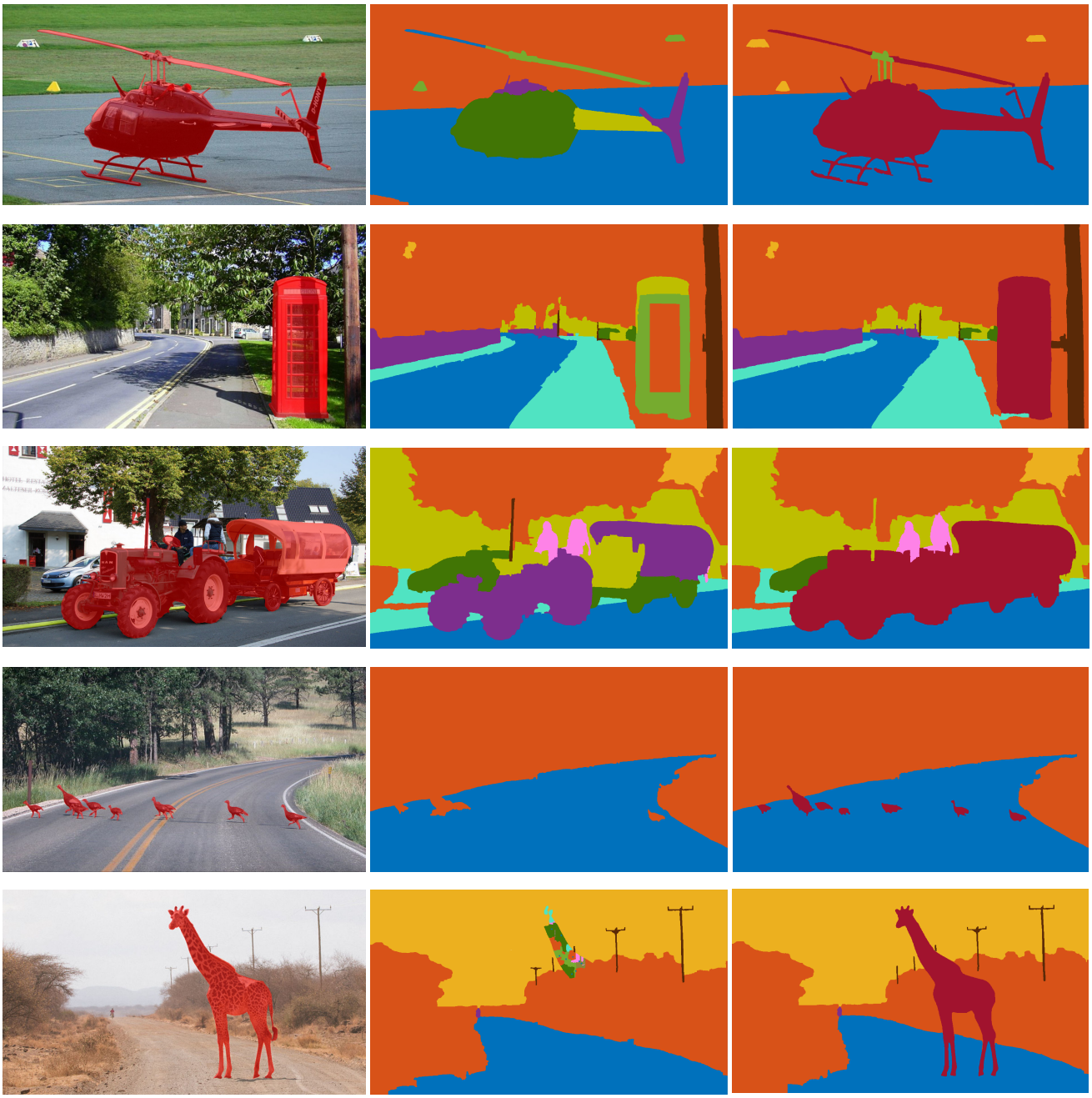
As pointed out in Sec. 3.1, we designed our neural network in order to be lightweight and faster at inference time. The architecture design choices explained in Sec. 3.1 allow inference on an image at 10 Hz. Additionally, we report the number of parameters and the GFLOPs of our model together with two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

D. Qualitative Results

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

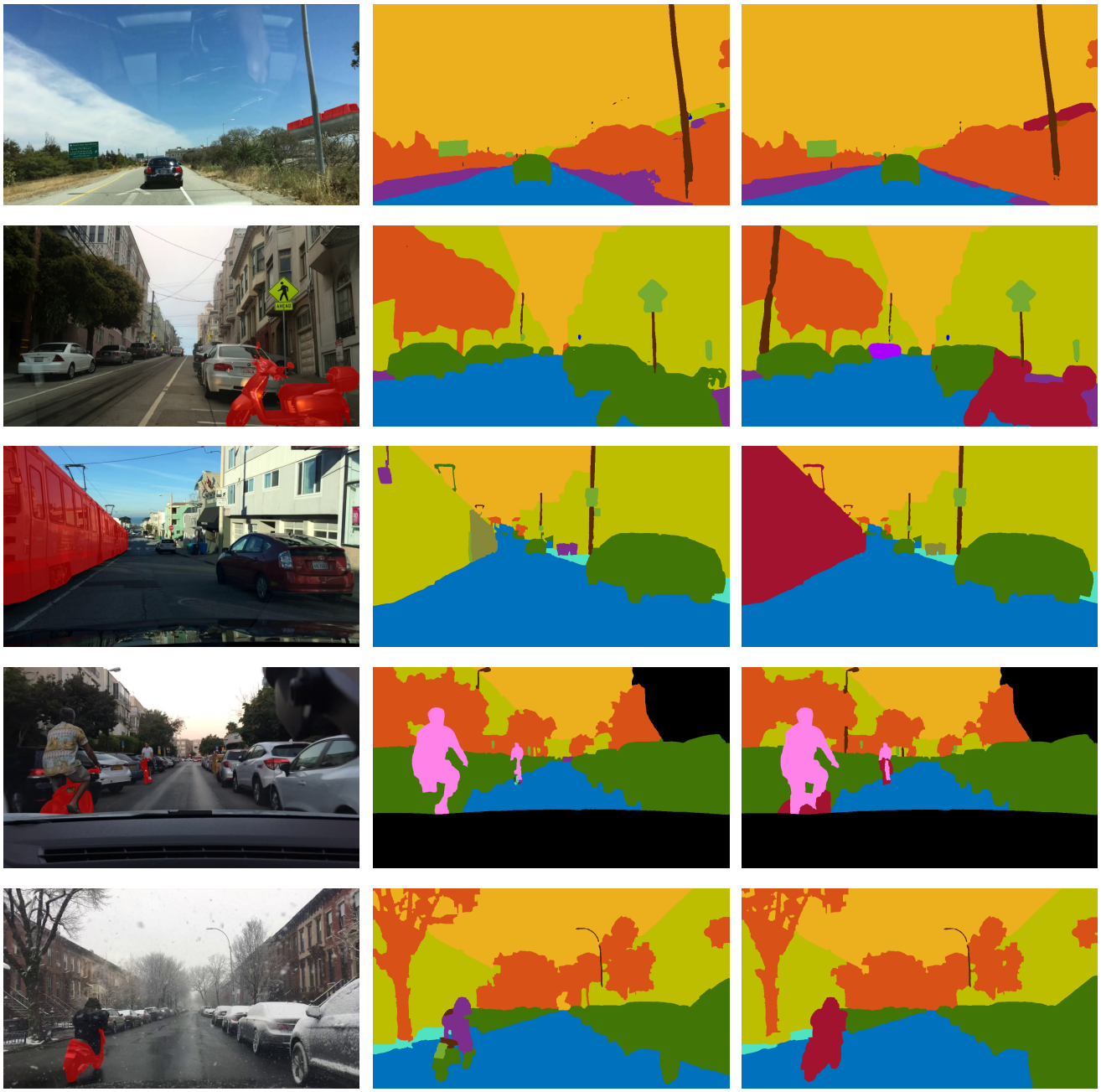
E. Limitations and Future Works

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



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

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



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

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

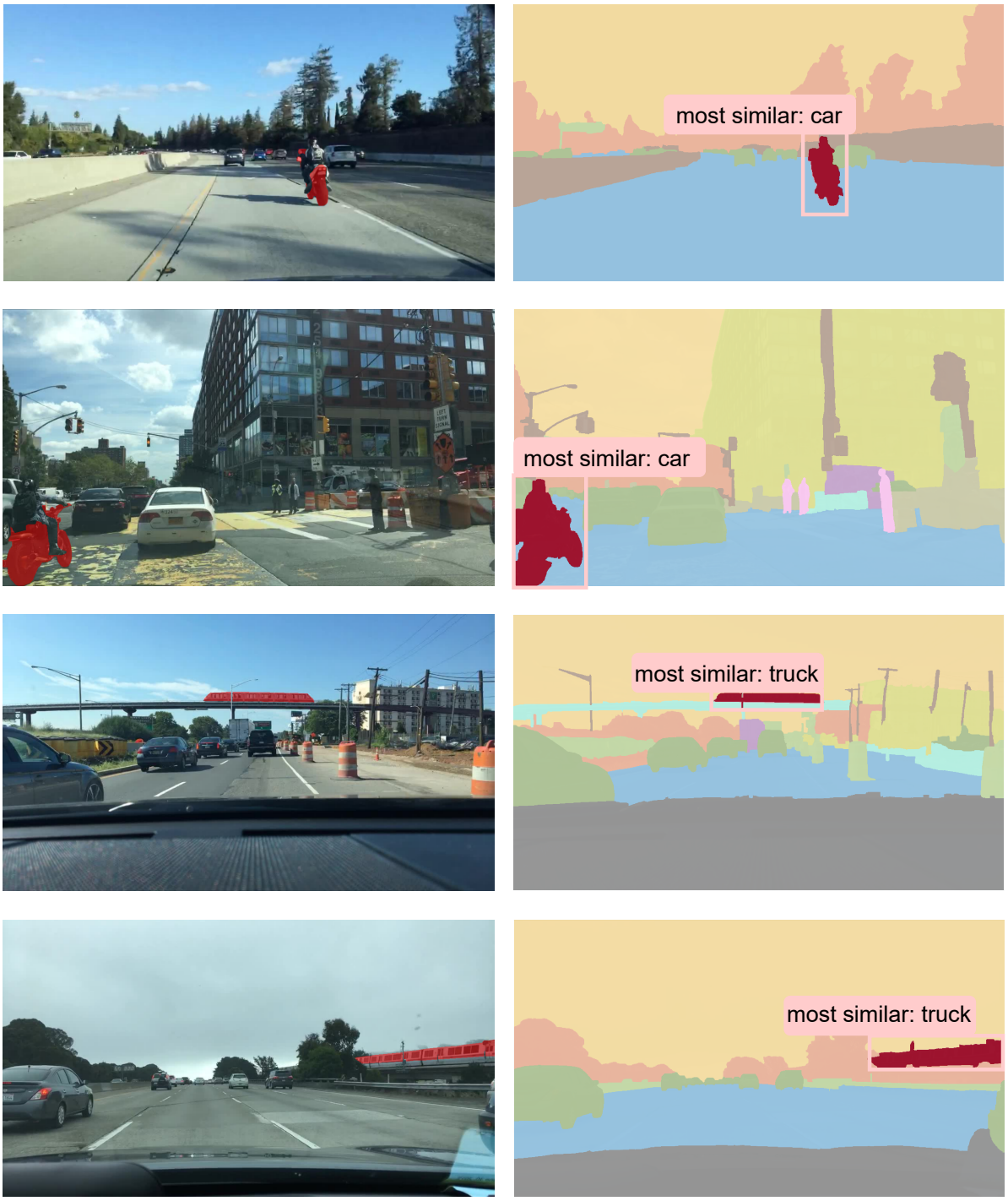
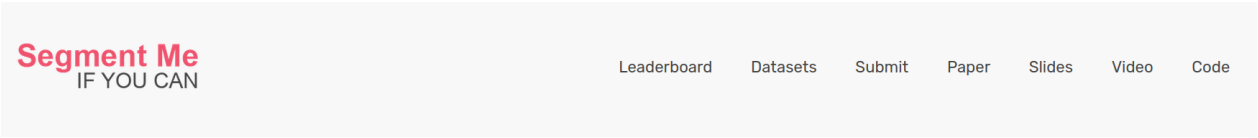


Figure 7. Class similarity results from the test set of BDDAnomaly. We show the input RGB overlayed with the ground truth unknown mask (a) and the prediction of our class similarity pipeline (b). In the open-world prediction, the unknown class is shown in red, and the overall semantic segmentation is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



LEADERBOARD

Evaluation Metrics

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

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

Anomaly Track

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

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