Unmasking Anomalies in Road-Scene Segmentation (Supplementary Material)

Shyam Nandan Rai[1], Fabio Cermelli[1,2], Dario Fontanel[1], Carlo Masone[1], Barbara Caputo[1]
[1]Politecnico di Torino, [2]Italian Institute of Technology
first.last@polito.it

Summary: This supplementary material contains additional method explanations, experiments, and results of Mask2Anomaly. These include:

- explanation of anomaly segmentation evaluation metrics;
- Mask2Anomaly results on validation sets;
- outlier loss comparison and analysis;
- training loss functions of Mask2Anomaly ;
- an analysis of various inference techniques applied to a Mask2Anomaly;
- performance stability of Mask2Anomaly;
- additional results and supplementary video.

A. Evaluation Metrics

Pixel-Level: For pixel-wise evaluation, let $Y \in \{Y_a, Y_{na}\}$ be the pixel level annotated ground truth labels for image χ containing anomaly. Y_a and Y_{na} represents the anomalous and non-anomalous labels in the ground-truth. Assume, $\hat{Y}(\gamma)$ is the model prediction obtained at thresholding $f(x)$ at γ. Then, we can write precision and recall equations as:

$$\text{precision}(\gamma) = \frac{|Y_a \cap \hat{Y}_a(\gamma)|}{|\hat{Y}_a(\gamma)|} \tag{1}$$

$$\text{recall}(\gamma) = \frac{|Y_a \cap \hat{Y}_a(\gamma)|}{|Y_a|} \tag{2}$$

and, AuPRC can be approximated as:

$$\text{AuPRC} = \int_{\gamma} \text{precision}(\gamma)\text{recall}(\gamma) \tag{3}$$

The AuPRC works well for unbalanced datasets making it particularly suitable for anomaly segmentation since all the datasets are significantly skewed. Next, we consider the False Positive Rate at a true positive rate of 95% (FPR$_{95}$), an important criterion for safety-critical applications that is calculated as:

$$\text{FPR}_{95} = \frac{|\hat{Y}_a(\gamma^*) \cap Y_{na}|}{|Y_{na}|} \tag{4}$$

where γ^* is a threshold when the true positive rate is 95%.

Component-Level: SMIYC [2] introduced a few component-level evaluation metrics that solely focus on detecting anomalous objects regardless of their size. These metrics are important to be considered because pixel-level metrics may not penalize a model for missing a small anomaly, even though such a small anomaly may be important to be detected. In order to have a component-level assessment of the detected anomalies, the quantities to be considered are the component-wise true-positives (TP), false-negatives (FN), and false-positives (FP). These component-wise quantities can be measured by considering the anomalies as the positive class. From these quantities, we can use three metrics to evaluate the component-wise segmentation of anomalies: sIoU, PPV, and F1*. Here we provide the details of how these metrics are computed, using the notation $\mathcal{K}$ to denote the set of ground truth components, and $\hat{\mathcal{K}}$ to denote the set of predicted components.

The *sIoU* metric used in SMIYC [2] is a modified version of the component-wise intersection over union proposed in [17], which considers the ground-truth components in the computation of the TP and FN. Namely, it is computed as

$$\text{sIoU}(k) = \frac{|k \cap \hat{K}(k)|}{|k \cap \hat{K}(k) \backslash \mathcal{A}(k)|}, \qquad \hat{K}(k) = \bigcup_{\hat{k} \in \hat{\mathcal{K}},\, \hat{k} \cap k \neq \emptyset} \hat{k} \tag{5}$$

where $\mathcal{A}(k)$ is an adjustment term that excludes from the union those pixels that correctly intersect with another ground-truth component different from k. We refer the reader to [2] for more details on this term. Given a threshold $\tau \in [0, 1]$, a target $k \in \mathcal{K}$ is considered a TP if $sIoU(k) > \tau$, and a FN otherwise.

The positive predictive value (*PPV*) is a metric that measures the FP for a predicted component $\hat{k} \in \hat{\mathcal{K}}$, and it is computed as

$$\text{PPV}(\hat{k}) = \frac{|\hat{k} \cap \hat{K}(k)|}{|\hat{k}|} \tag{6}$$

A predicted component $\hat{k} \in \hat{\mathcal{K}}$ is considered a FP if $PPV(\hat{k}) \leq \tau$. Finally, the $F1^*$ summarizes all the

Methods	FS L&F		FS static	
	AuPRC↑	FPR$_{95}$ ↓	AuPRC↑	FPR$_{95}$ ↓
Max Softmax [8]	4.59	40.59	19.09	23.99
Max Logit [8]	14.59	42.21	38.64	18.26
Entropy [8]	10.36	40.34	26.77	23.31
Energy [14]	25.79	32.26	31.66	37.32
SynthCP [20]	6.54	45.95	23.22	34.02
SynBoost [6]	40.99	34.47	48.44	47.71
SML [9]	36.55	14.53	48.67	16.75
Deep Gambler [15]	39.77	12.41	67.69	15.39
Dense Hybrid [7]	63.80	**6.10**	60.20	4.90
PEBEL [19]	59.83	6.49	82.73	6.81
Mask2Anomaly (Ours)	**69.41**	9.46	**90.54**	**1.98**

Table 1: **Fishyscapes Validation Results:** The best and second best results are **bold** and underlined, respectively.

Methods	SMIYC-RA21		SMIYC-RO21	
	AuPRC↑	FPR$_{95}$ ↓	AuPRC↑	FPR$_{95}$ ↓
Max Softmax [8]	40.4	60.2	43.4	3.8
ODIN [12]	46.3	61.5	46.6	4.0
Mahalanobis [11]	22.5	86.4	25.9	26.1
MC Dropout [16]	29.2	77.9	7.9	43.8
Ensemble [10]	16.0	80.0	4.7	98.3
Void Classifier [1]	39.3	66.1	9.8	43.6
Learning Embedding [1]	51.9	60.0	1.5	56.7
Image Resynthesis [13]	76.4	20.5	70.3	1.3
SynBoost [6]	68.8	30.9	81.4	2.8
Maximized Entropy [3]	80.7	17.4	**94.4**	0.4
Mask2Anomaly (Ours)	**94.5**	**3.3**	88.6	**0.3**

Table 2: **SMIYC Validation Results:** The best and second best results are **bold** and underlined, respectively.

component-wise TP, FN, and FP quantities by the following formula:

$$F1^*(\tau) = \frac{2TP(\tau)}{2TP(\tau) + FN(\tau) + FP(\tau)} \qquad (7)$$

B. Results on Fishyscapes and SMIYC validation sets

To provide a comprehensive evaluation, we have benchmarked Mask2Anomaly results on the Fishyscapes and SMIYC validation sets as presented in Tab. 1 and Tab. 2, respectively. We can observe that Mask2Anomaly outperforms all the prior methods by a large margin on both benchmarks. Interestingly, maximized entropy and dense hybrid show the best AuPRC for SMIYC-RO21 and FPR$_{95}$ for FS L&F, respectively. However, overall Mask2Anomaly gives the best performance on all the benchmarks. This suggests that mask-based architecture offers better generalizability in comparison to per-pixel architecture due to its intrinsic property of encouraging objectness.

C. Outlier Loss Comparision

We now empirically demonstrate why the proposed mask contrastive loss, a margin-based loss, performs better at anomaly segmentation than binary cross-entropy loss. We train Mask2Anomaly with M_{OOD} using binary-cross entropy. The new loss based on the binary cross entropy can be written as:

$$L_{BCE} = M_{OOD}\log(l_N) + (1 - M_{OOD})\log(1 - l_N) \quad (8)$$

$$\text{where, } l_N = -\max_{k=1}^{K}\left(\text{softmax}(C)^T \cdot \text{sigmoid}(M)\right) \quad (9)$$

l_N is the negative likelihood of in-distribution classes calculated using the class scores C and class masks M. Figure 1 illustrates the anomaly segmentation performance comparison on FS L&F validation dataset between the Mask2Anomaly when trained with the binary cross entropy loss and mask contrastive loss, respectively. We can observe that the mask contrastive loss achieves a wider margin between out-of-distribution(anomaly) and in-distribution prediction while maintaining significantly lower false positives.

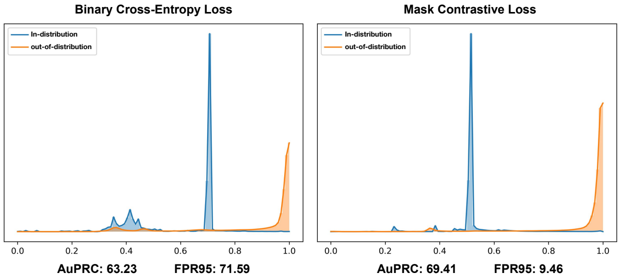


Figure 1: **Outlier Loss Comparision:** To train Mask2Anomaly on the outlier set, we find that mask contrastive loss, which is a margin-based loss shows better performance compared to the binary cross-entropy loss. Both experiments are done on the FS L&F validation set.

D. Training Loss

Mask2Anomaly gives two sets of outputs: class scores (C) and class masks (M). To train M, we first pad the ground truth mask M^{gt} with "no object" masks denoted by ϕ. Since we assume $M \geq M^{gt}$, padding the ground truth masks allow us one-to-one matching. Now, we use bipartite matching to match the ground truth and the predicted masks, and the assignment cost is given by:

$$L_{masks} = \lambda_{bce}L_{bce} + \lambda_{dice}L_{dice} \qquad (10)$$

where L_{bce} and L_{dice} are the binary cross entropy loss and the dice loss calculated between the matched masks. λ_{bce} and λ_{dice} are the loss weights that are both set to 5.0. To train C, which indicates the semantic class of a mask, we used the cross-entropy loss L_{ce}. The total training loss is given by:

$$L = L_{masks} + \lambda_{ce}L_{ce} \qquad (11)$$

with λ_{ce} set to 2.0 for the prediction that matched with

			SMIYC-RA21		SMIYC-RO21		FS L&F		FS Static		Road Anomaly		Average	
C	M	$f(C).f(M)$	AuPRC ↑	$\text{FPR}_{95}\downarrow$	AuPRC ↑	$\text{FPR}_{95}\downarrow$	AuPRC ↑	$\text{FPR}_{95}\downarrow$	AuPRC ↑	$\text{FPR}_{95}\downarrow$	AuPRC ↑	$\text{FPR}_{95}\downarrow$	AuPRC ↑	$\text{FPR}_{95}\downarrow$
I	I	I	9.47	95.16	4.44	73.45	2.53	92.16	1.18	99.97	65.59	97.56	16.64	91.66
Softmax	Softmax	I	44.73	38.27	3.16	95.72	4.82	47.98	10.34	52.04	42.74	55.73	21.13	57.94
Sigmoid	Sigmoid	I	25.04	93.14	83.14	1.24	14.55	43.83	45.67	96.87	28.1	91.63	39.3	65.34
Sigmoid	Softmax	I	29.29	39.01	7.48	98.01	0.42	48.23	6.37	52.16	25.61	55.78	13.83	58.63
Softmax	Sigmoid	I	95.48	2.41	92.89	0.15	69.41	9.46	90.54	1.98	79.7	13.45	**85.56**	**5.51**
Softmax	Sigmoid	Softmax	94.55	3.31	88.59	0.36	70.8	32.66	88.96	2.22	78.3	15.54	84.24	10.81

Table 3: **Mask2Anomaly Inference**: we show various inference techniques on Mask2Anomaly for anomaly segmentation. $f(.)$ represents the function applied to class scores or masks. I is the identity function. The best results are in bold.

	SMIYC-RA21		SMIYC-RO21		FS L&F		FS Static		Average σ	
Methods	AuPRC ↑	$\text{FPR}_{95}\downarrow$	AuPRC ↑	$\text{FPR}_{95}\downarrow$	AuPRC ↑	$\text{FPR}_{95}\downarrow$	AuPRC ↑	$\text{FPR}_{95}\downarrow$	AuPRC	FPR_{95}
Mask2Anomaly-S1	95.48	2.41	92.89	0.15	69.41	9.46	90.54	1.98	-	-
Mask2Anomaly-S2	92.03	3.22	92.3	0.27	69.19	13.47	85.63	5.06	-	-
σ(Mask2Anomaly)	± 2.44	±0.57	±0.42	±0.08	±0.16	±2.84	±3.47	±2.18	±1.62	±1.41
Dense Hybrid-S1	52.99	38.87	66.91	1.91	56.89	8.92	52.58	6.03	-	-
Dense Hybrid-S2	60.59	32.14	79.64	1.01	47.97	18.35	54.22	5.24	-	-
σ(Dense Hybrid)	±5.37	±4.76	±9.00	±0.64	±6.31	±6.67	±1.16	±0.56	±5.46	±3.15

Table 4: **Performance stability in Mask2Former:** we can observe that the average deviation in the performance of the dense hybrid is significantly higher than Mask2Anomaly. σ denotes the standard deviation.

ground truth and 0.1 for ϕ, *i.e.* for no object. After training the Mask2Anomaly for 90K iterations, we fine-tune the network with the mask contrastive loss L_{CL}. The new training loss is written as:

$$L_{M2A} = L + L_{CL} \tag{12}$$

We perform all the training and inference on a single Nvidia Titan RTX with 24GB memory.

E. Mask2Anomaly Inference

The per-pixel classification networks have a straightforward inference as the network outputs a pixel-wise anomaly map. However, in the case of a mask architecture, we get a set of class scores C and a set of binary mask M. So, we test various inference techniques on Mask2Anomaly for anomaly segmentation, as shown in Table 3. We find that the marginalization over class scores obtained after the softmax and taking the sigmoid of the mask yields the best results. Also, we observe that applying a softmax after the marginalization to perform max-softmax [8] does not give good results.

F. Performance stability on different outlier sets

Employing an outlier set to train an anomaly segmentation model presents a challenge because the model's performance can vary significantly across different sets of outliers. Here, we show that Mask2Anomaly performs similarly when trained on different outlier sets.

We randomly chose two subsets of 300 MS-COCO images (S1, S2) as our outlier dataset for training Mask2Anomaly and DenseHybrid. Table 4 shows the performance of Mask2Anomaly and Dense Hybrid trained on S1 and S2 outlier sets, along with the standard deviation(σ) in the performance. We can observe that the variation in performance for the dense hybrid is significantly higher than Mask2Anomaly. Specifically, in dense hybrid, the average deviation in AuPRC is greater than 300%, and the average variation in FPR_{95} is more than 200% compared to Mask2Anomaly.

G. Additional Results

Segmentation results: In Tab. 5 and Fig. 2, we show the segmentation results for Mask2Anomaly and Mask2Former. We can qualitatively and qualitatively infer that Mask2Anomaly performs better than Mask2Former.

Qualitative anomaly segmentation: In Fig. 3, we show the qualitative comparison of Mask2Anomaly with best-existing anomaly segmentation methods: Maximized Entropy [3] and Dense Hybrid [7]. We observe that these per-pixel classification architectures suffer from large false positives, whereas Mask2Anomaly, a mask-transformer, shows confident results across all datasets.

Attention comparison: Figure 4 shows the anomaly segmentation results obtained using various attention mechanisms, and the global mask attention clearly exhibits the best performance.

Qualitative ablation study: We show a component-wise qualitative ablation of Mask2Anomaly in Fig. 5 by progressively adding each components. We can observe that each proposed component improves anomaly segmentation and complements the others.

Supplementary video: Shows the performance of Mask2Anomaly on the sequence of images of small

Methods	road	s. walk	building	wall	fence	pole	t. light	t. sign	veg.	terrain	sky	person	rider	car	truck	bus	train	mbike	bicycle	mIoU
Mask2Former	98.4	87.0	92.7	46.1	59.9	69.5	75.3	82.2	92.9	63.8	95.2	84.9	69.3	95.6	58.7	77.0	79.9	62.7	80.0	77.4
Mask2Anomaly	98.5	86.3	91.5	53.9	60.2	67.5	74.3	88.1	93.1	62.6	96	84.1	62.7	95.7	79.6	80.3	77.1	70.1	77.1	**78.8**

Table 5: Class-wise semantic segmentation results comparison between Mask2Former and Mask2Anomaly on Cityscapes validation set.

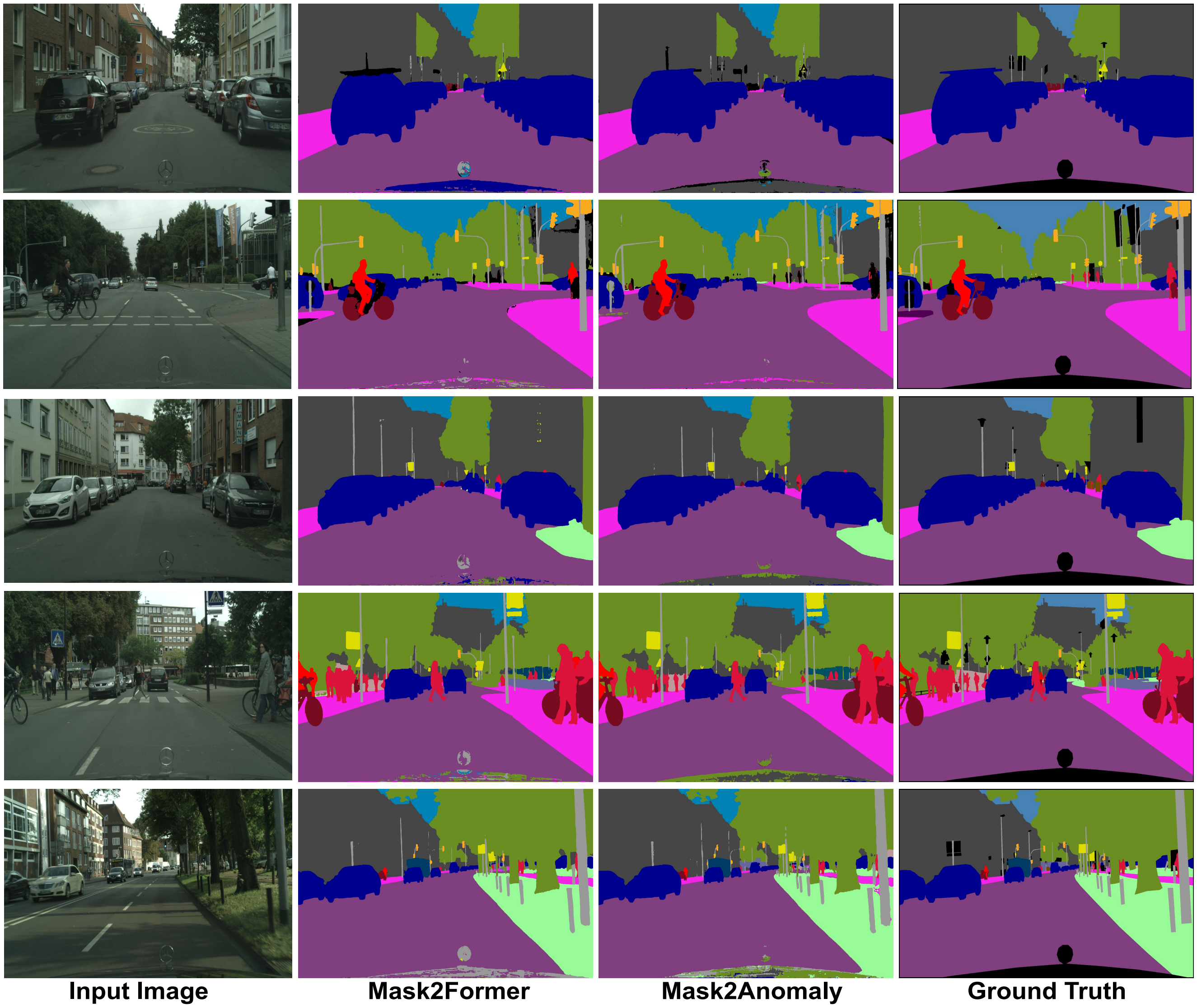


Figure 2: **Semantic Segmentation Results:** We can visually infer that Mask2Anomaly shows similar segmentation results when compared with Mask2Former [4].

obstacle dataset [18]. Mask2Anomaly displays an impressive performance in segmenting wildlife on the road and anomalies in low-light conditions.

Failure cases: Fig. 6 shows that Mask2Anomaly struggles to segment tiny anomalies and falsely detects road potholes as anomalies.

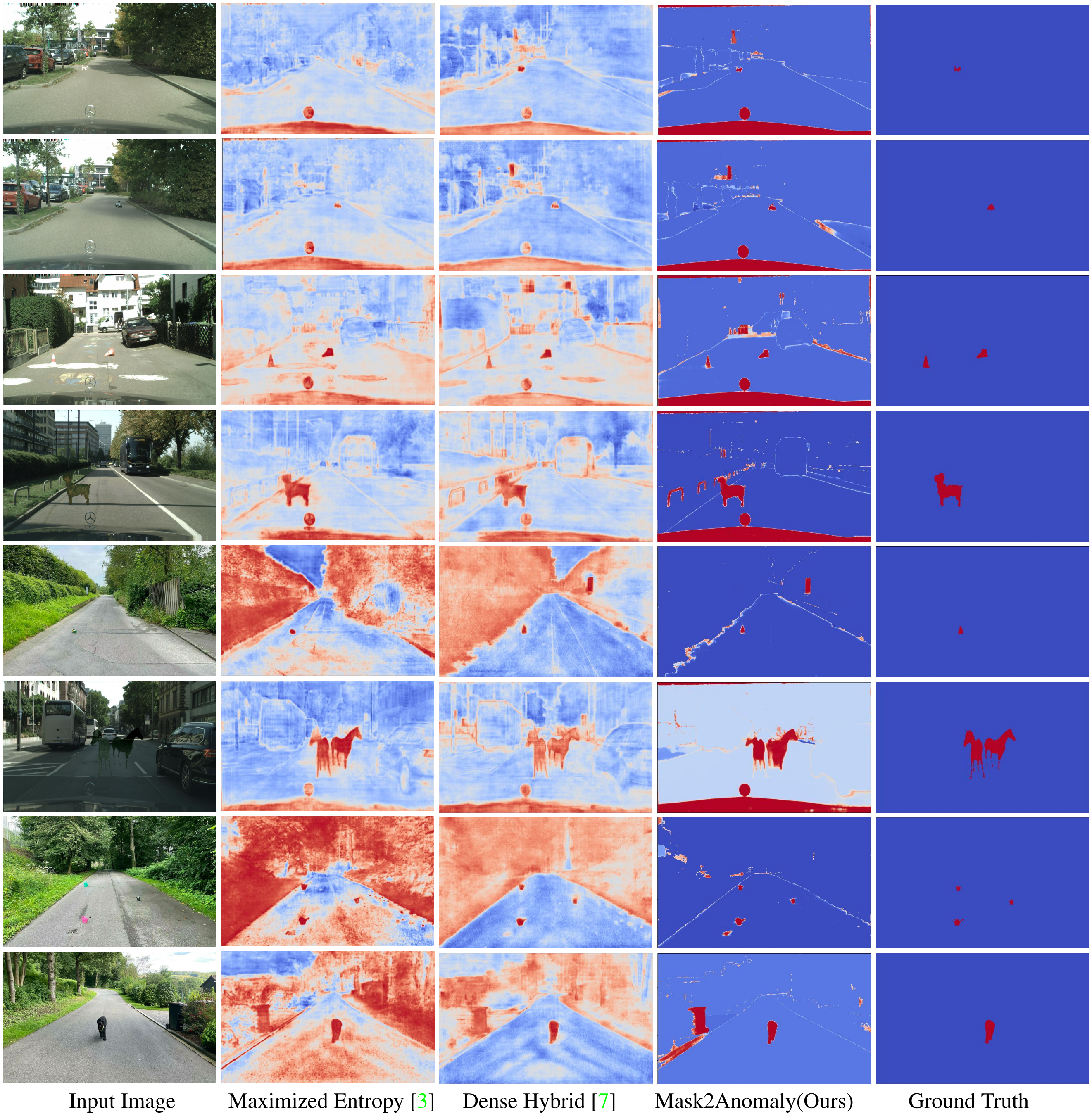
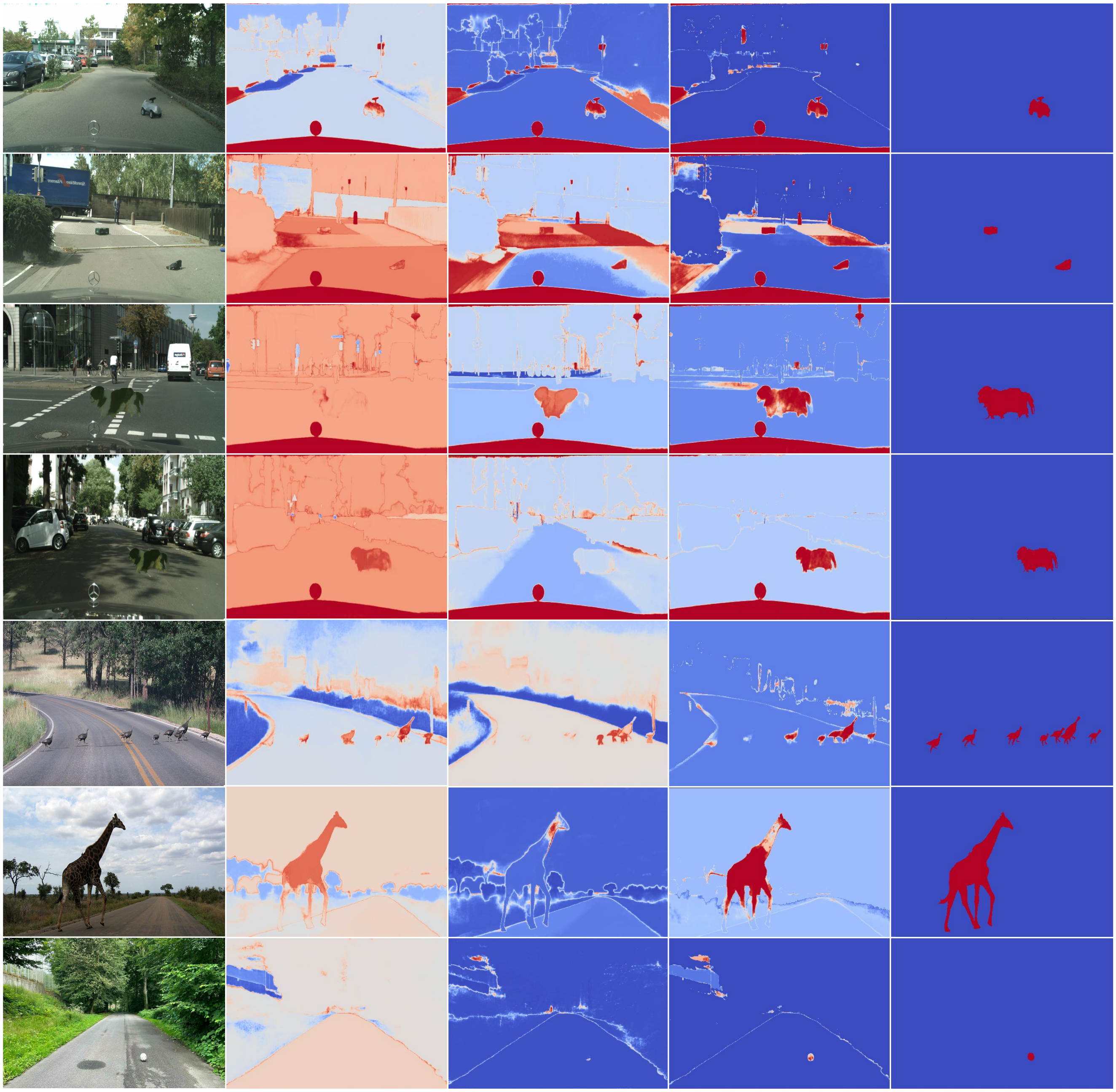


Figure 3: **Qualitative Results**: We observe that per-pixel classification architecture: Maximized Entropy and Dense Hybrid suffer from large false positives, whereas Mask2Anomaly which is a mask-transformer, show confident results across all datasets. Anomalies are represented in red.



Input Image Cross-Attention [5] Mask-Attention [4] Global Mask Attention (Ours) Ground Truth

Figure 4: **Attention Comparison**: We observe that the proposed global mask attention can better segment anomaly among the compared attention mechanism. Anomalies are represented in red.

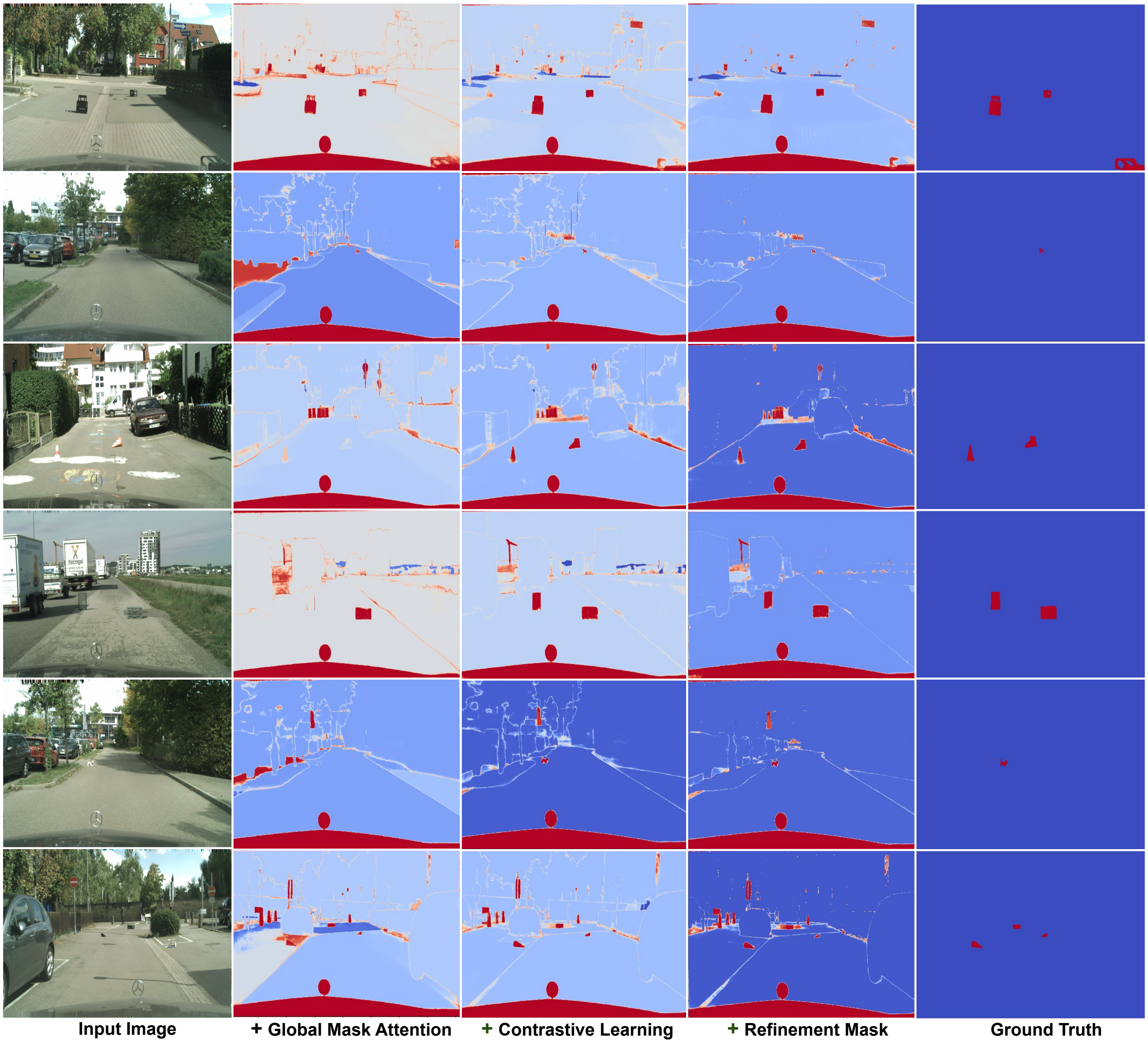


Figure 5: **Mask2Anomaly Qualitative Ablation**: shows the performance gain by progressively adding (left to right) proposed components. Anomalies are represented in red.

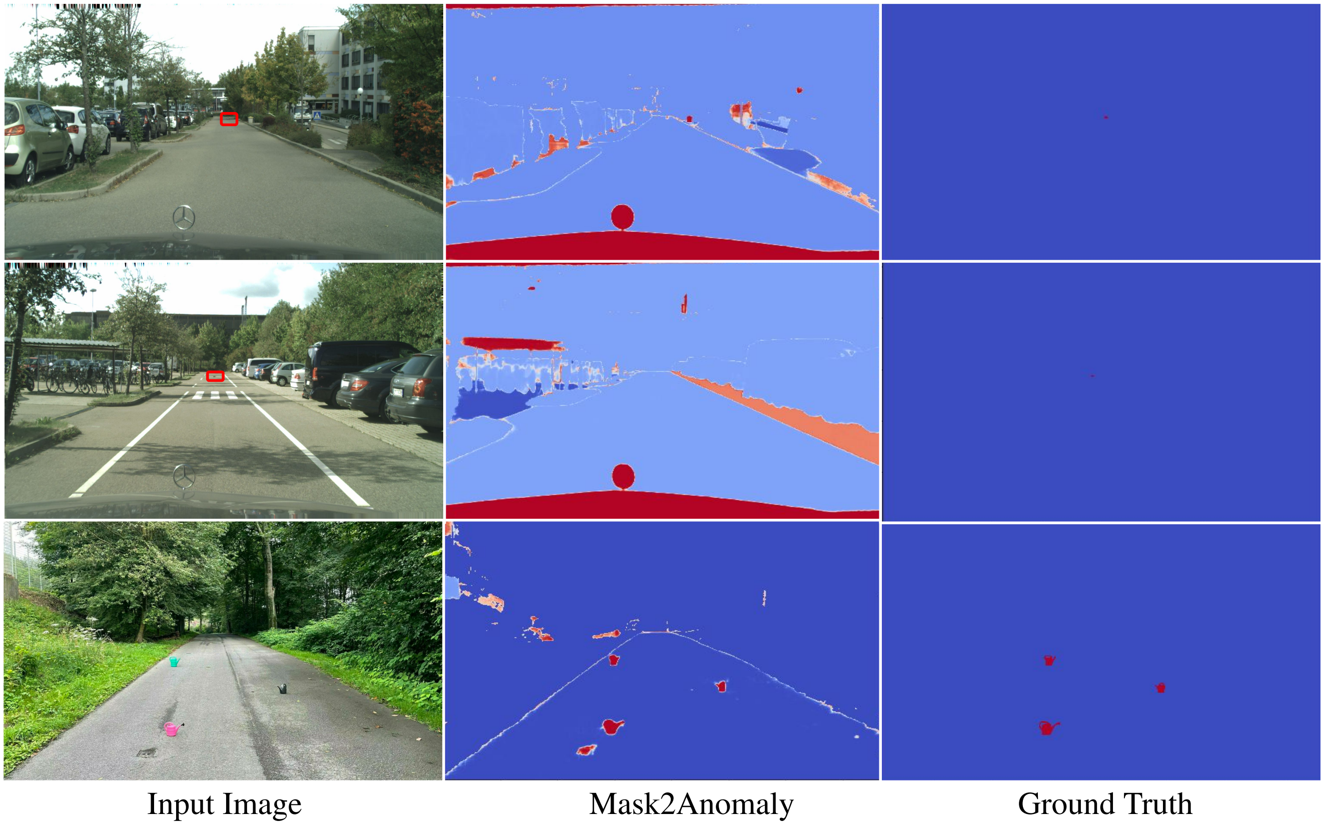


Figure 6: **Failure Cases:** Row (1,2): We can observe that Mask2Anomaly is unable to segment tiny anomalies(inside red bounding boxes of input image). Please zoom in for better clarity. Row 3: Mask2Anomaly falsely segments the pothole on the road as an anomaly. Anomalies are indicated in red in the ground truth.

References

[1] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, 129(11):3119–3135, 2021. 2

[2] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Mathieu Salzmann, Pascal Fua, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. *arXiv preprint arXiv:2104.14812*, 2021. 1

[3] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 5128–5137, 2021. 2, 3, 5

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 4, 6

[5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 6

[6] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16918–16927, 2021. 2

[7] Matej Grcić, Petra Bevandić, and Siniša Šegvić. Densehybrid: Hybrid anomaly detection for dense open-set recognition. In *European Conference on Computer Vision*, pages 500–517. Springer, 2022. 2, 3, 5

[8] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 2, 3

[9] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15425–15434, 2021. 2

[10] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2

[11] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 2

[12] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 2

[13] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2152–2161, 2019. 2

[14] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020. 2

[15] Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[16] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 2

[17] Matthias Rottmann, Pascal Colling, Thomas Paul Hack, Robin Chan, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020. 1

[18] Aasheesh Singh, Aditya Kamireddypalli, Vineet Gandhi, and K Madhava Krishna. Lidar guided small obstacle segmentation. *arXiv preprint arXiv:2003.05970*, 2020. 4

[19] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *European Conference on Computer Vision*, pages 246–263. Springer, 2022. 2

[20] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2