Prior2Former - Evidential Modeling of Mask Transformers for Assumption-Free Open-World Panoptic Segmentation

Supplementary Material

Sebastian Schmidt[1,2,*] Julius Körner[1,2,*] Dominik Fuchsgruber[1]
Stefano Gasperini[1,3,5] Federico Tombari[1,4] Stephan Günnemann[1,5]

[1] Technical University of Munich [2] BMW Group [3] Visualais
[4] Google [5] Munich Center for Machine Learning

A. Experimental Details

In this section, we provide details about the individual experimental setup and baselines, as well as the implementation details.

A.1. Anomaly Segmentation

For evaluating anomaly segmentation, we compare P2F against several baselines based on M2F [12] and U3HS [20] on the SMIYC [4] benchmark. The datasets used include SMIYC Road Anomaly [5], FS [3], and FS L&F [3]. All these benchmarks employ Cityscapes [14] as their in-distribution data source. The Cityscapes dataset consists of 19 classes — 8 categorized as “things” and 11 as “stuff” — captured from various cities across Germany. We employ the official SMIYC evaluation script with the addition of the FS dataset. Following the official evaluation, void-labeled regions are ignored for computing results for all three datasets. For FS L&F, only the region of interest, typically the road in front of the vehicle, is evaluated. The official metric requires an image-shaped array with anomaly scores and dynamically selects the best-fitting threshold per image. As *avoiding prior knowledge* is a fundamental aspect of our work, we primarily focus on metric-based baselines without OOD usage. Specifically, we employ the baselines **RbA** [41], **EAM** [22] and **M2A** [46] based on M2F and implemented the general benchmarks **SML** [29] and **MM** as mask variant of **MSP** [26] on M2F. While **RbA** reports an OOD-free version, we use the OOD-free versions of **EAM** and **M2A** presented in their ablation study. Further, we use the prior knowledge-free U3HS [20], which is based on DeeplabV3+ [9].

In contrast to how the respective papers evaluate these methods, EAM, RbA, and M2A are trained on the panoptic segmentation task, not on semantic segmentation. We use the same ResNet50 (R50) [24] backbone architecture for all methods to enable comparability between M2F-based approaches and U3HS. Additionally, EAM, RbA, and M2A train on OOD data to supervise anomaly detection. We omit this to mitigate assumptions about OOD data to enable a fair comparison of the uncertainty metric. Therefore, all presented M2F-based baselines are primarily post-processing functions applied to the output of a standard M2F model. For M2A, a global mask attention mechanism complements the “local” mask attention during training and inference.

For SML, we apply the maximum logit standardization to the logits $L(x)$ of the M2F model introduced in Sec. 3:

$$L(x)_c = \sum_{i=1}^{N_m} p_i(c) \cdot m_i[h, w], \tag{12}$$

Let $\mathbb{T}$ be the set of all training images and let $L(x)_c \in \mathbb{R}^{H \times W}$ be the pixel-wise logits of class c. We compute the mean and standard deviation $\mu, \sigma \in \mathbb{R}^K$ of these maximum logits over all classes:

$$\mu_c = \text{MEAN}(\{L_{\hat{c}(x)}[h, w] \mid 1 \leq h \leq H, 1 \leq w \leq W, x \in \mathcal{X}_{\text{train}}\}) \tag{13}$$

$$\sigma_c = \text{STD}(\{L_{\hat{c}(x)}[h, w] \mid 1 \leq h \leq H, 1 \leq w \leq W, x \in \mathcal{X}_{\text{train}}\}) \tag{14}$$

where

$$\hat{c}(x) = \underset{c \in \{1,\ldots,K\}}{\arg\max} \; L_c(x)[h, w].$$

Hence,

$$U^{\text{SML}}[h, w] = -\frac{L_{\hat{c}}[h, w] - \mu_{\hat{c}}}{\sigma_{\hat{c}}}.$$

To obtain the SML uncertainty $U^{\text{SML}}[h, w]$ for a pixel at position h, w of image x, we standardize the maximum logit $L_{\hat{c}}^{(x)}[h, w]$ using the mean and standard deviation computed from all pixels $[h, w]$ over all images x in the training set, where the class $\hat{c}$ has the greatest logit.

The Maximum Mask (MM) baseline presents a mask-based variant of MSP [26]:

$$U^{\text{MM}}[h, w] = -\max_{i \in \{1,\ldots,N_m\}} m_i[h, w].$$

[*] Equal Contribution

For MSP [26], the maximum softmax probability of the current pixel is taken as the confidence score, whereas for MM, the maximum sigmoid probability over all masks serves as the confidence score. The uncertainty estimates according to EAM [22], RbA [41], and M2A [46] follow the corresponding papers:

$$U^{(\text{EAM})}[h,w] = -\sum_{i=1}^{N_m} m_i[h,w] \cdot \left(\max_{c=1,\dots,K} (p_i(c)) \right), \tag{15}$$

$$U^{(\text{RBA})}[h,w] = -\sum_{c=1}^{K} \tanh(L_c[h,w]), \tag{16}$$

$$U^{(\text{M2A})}[h,w] = \left(1 - \max_{c=1,\dots,K} L_c[h,w] \right) \cdot R_M[h,w], \tag{17}$$

where the sigmoid function is applied elements-wise and R_M is the mask filter as presented in M2A [46], such that the uncertainty $U^{(M2A)}[h,w]$ is set to zero if there is no mask for pixel $[h,w]$ where the masks score $m_i[h,w] > 0.5$ and the predicted class is "road" or a "thing" class and the softmax confidence is greater than 0.95. We evaluate M2A by only using the pixel-wise filter $m_i[h,w] > 0.5$ since all other models do not include class-specific information for the detection of anomalies. Note that for the M2A uncertainty, the logits are obtained from a model trained with global mask attention.

Since the SMIYC metric expects one anomaly score per pixel on a given image, the logical "and" between thresholded distances and anomaly scores, as implemented by U3HS [20], can not be used for the evaluation metric. Hence, for U3HS, the Dirichlet strength from the semantic head is used as an anomaly score.

A.2. Anomaly Instance Segmentation

The task of Anomaly Instance Segmentation is evaluated using the official OoDIS benchmark code [43]. This benchmark assesses performance on three datasets: an unknown split of L&F [44], as well as the test splits of the SMIYC RoadAnomaly21 [5] and RoadObstacles21 [5]. The evaluation requires binary images containing the recognized anomalies and a confidence score for each binary image. An anomaly instance prediction with an Intersection over Union (IoU) greater than 0.5 and the highest confidence score is considered a positive prediction, contributing to the true positive and false positive rates. All other predictions, not counted as valid predictions for another anomaly, are considered false positives.

For submission and evaluation on the publicly available validation set containing only L&F [44] data, P2F, and U3HS [20] are trained on the Cityscapes dataset.

A.3. Closed-World and Open-World Panoptic Segmentation

Here, we detail the experimental setup for computing the PQ and mIoU metrics on the two datasets COCO [35] and BDD100k [65] for P2F and U3HS [20].

COCO: For the COCO [35] dataset, we preprocess the data by excluding all images containing any of the 20% (i.e., 16) least frequent classes in the training set, specifically: *baseball bat, bear, fire hydrant, frisbee, hairdryer, hot dog, keyboard, microwave, mouse, parking meter, refrigerator, scissors, snowboard, stop sign, toaster, and toothbrush*, which aligns with [20, 63]. The images in the validation set containing any held-out classes are placed into a separate set, termed the open-world validation set, while the remaining images form the closed-world validation set. The semantic labels of the known classes are retained, while the semantic labels of the held-out classes are merged into a single OOD class. We then evaluate the mIoU and PQ of the open-world validation set, treating all anomalies as belonging to the OOD class. We use COCO panopticapi[1] and torchmetrics[2] for PQ and mIoU evaluation, respectively. Following the training scheme presented in M2F [46] for COCO, we employ a random resize crop augmentation strategy during training, which maintains the aspect ratio of the input images. During validation, we maintain the original size of the COCO images. This approach differs from the scheme used in U3HS, where images are resized for both training and evaluation.

BDD100k: For the BDD100k [65] dataset, we follow the evaluation and training procedure employed for COCO. For the dataset split, we use the proposed class settings of BDD anomaly [27] and exclude the classes motorcycle and bicycle. In contrast to [27], we use the panoptic labels instead of the semantic labels to enable a panoptic evaluation. Note that besides the instance information, the panoptic labels additionally have an increased number of classes compared to the semantic labels, which increases complexity. We construct the closed-world training and validation sets and the open-world validation set similarly, but evaluate all scores based on the respective torchmetrics implementation. We restrict the evaluation of mIoU and PQ only on instances that cover more than 2,500 pixels. This is because the depth values in the dataset show significant variability, resulting in many small anomalies that cover only a few pixels. Including these small anomalies in the evaluation does not meaningfully contribute to understanding the models' ability to recognize the held-out classes.

Cityscapes: For open and world segmentation, we follow the setup described by [20] and use the standard training setup as suggested by M2F for all M2F-based baselines,

[1] https://github.com/COCOdataset/panopticapi

[2] https://lightning.ai/docs/torchmetrics/stable/

including P2F. For U3HS [20], we use the setup described in their paper and use the respective evaluation settings. For the Lost & Found evaluation, we use a resized version as done by [20] and a full-size version to maintain a fair evaluation.

A.4. Implementation Details

We train P2F and M2F on **Cityscapes** using a series of augmentations. Specifically, we apply a random zoom crop with a crop size of $(1024, 512)$ and a zoom range of $(1.0, 2.0)$. Additionally, we utilize a color jitter augmentation with a brightness delta of 32, a hue delta of 18, contrast adjustments in the range of $(0.5, 1.5)$, and saturation changes within $(0.5, 1.5)$. We also include a random horizontal flip with a probability of 0.5. The color jittering is necessary because we use the models trained on Cityscapes for evaluation on SMIYC and Oodis. Without this augmentation, different lighting conditions, especially of the sky, could result in high anomaly scores.

We train the models with a batch size of 16, a learning rate of 0.0001, and a learning rate of 0.00001 for the backbone, using the AdamW optimizer [38] with a weight decay of 0.05. A polynomial learning rate scheduler with a power of 0.9 is employed. Gradients are clipped at 0.01 according to the L2 norm. The training is conducted for 450 epochs, which corresponds to approximately $90,000$ iterations. Hence, we follow the training as suggested for M2F [12] on Cityscapes, with the only difference of adding the color jittering for the more robust uncertainty estimation.

For the **COCO** dataset, we employ a random zoom crop that keeps the aspect ratio fixed with a zoom range of $(0.1, 2.0)$ and a crop size of $(512, 512)$. No color jittering is used. The optimizer and scheduler configurations, random horizontal flip, and gradient clipping remain the same, but the batch size is increased to 32, and the training is performed over 50 epochs. In summary, we use a training scheme closely matching the M2F [12] configuration for COCO. However, we changed the crop size from $(1024, 1024)$ to $(512, 512)$ and the batch size from 16 to 32 for known settings, which improved the convergence speed.

For the **BDD** dataset, M2F [12] does not provide a configuration. Hence, we primarily follow the training on Cityscapes as a related dataset. We scale the images within the range of $(1.0, 2.0)$ and crop both dimensions by half to a size of $(640, 360)$, similar to the cropping applied in Cityscapes. No color jittering is used. All other parameters are consistent with those used for Cityscapes.

Across **all** datasets, we train using a no-object loss coefficient of 0.1, a class loss weight $\lambda_{\text{cls}} = 2.0$, a symmetric dice loss weight $\lambda_{\text{sDice}} = 5.0$, and an evidential loss weight $\lambda_{\text{evi}} = 0.1$ for P2F. The evidential loss weight is tuned to ensure that the loss values of the symmetric dice and evidential loss have similar absolute values. We use 200 object queries for both the baselines and P2F, which is necessary to generate a sufficient number of valid mask predictions on the SMIYC Road Anomaly and SMIYC Road Obstacle datasets. All other hyperparameters for model building and training remain consistent with those used in the official M2F repository [3].

For **postprocessing** the predictions of P2F, we set the object-mask-threshold to 0.5, compared to 0.8 in the original M2F model. This threshold on the mask prediction probability determines if an object is present in the panoptic prediction. However, since the beta prior predictions of P2F are more restrictive, we choose a lower threshold value.

To create the anomaly instance segmentation, we introduce a threshold t to the uncertainty estimates of P2F. The corresponding feature vectors of the pixel embedding F_E are then clustered using DBSCAN [17] with parameters eps and min-samples. The mask correspondence is determined using a scalar product between the predicted mask features and the pixel embedding. Hence, it is natural to use one minus the cosine similarity instead of an Euclidean distance for the DBSCAN, as it is the normalized scalar product of the two vectors. For the submission on OoDIS, the uncertainty threshold t is set to 2 times the standard deviation away from the mean uncertainty on the training set $t = -0.6$ for the L&F split. The overall uncertainty on SMIYC Road Anomaly and SMIYC Road Obstacle is greater since the images cover uncommon scenarios and conditions. Hence, we increase the threshold to 3.5 times the standard deviation, i.e., $t = -0.4$. For COCO and BDD, we follow the same approach of determining the threshold and, therefore, set $t = -0.55$ and $t = -0.6$. The parameter eps strongly influences the granularity of the clustering algorithms. Evaluating the embedding space, we use 0.04 in all experiments. Except for COCO, we set eps $= 0.1$ due to the wider variety of semantic classes and instances. The min-samples parameter is set to 17, however, clustering the embedding is robust against changes in this parameter, with values ranging from 10 to 23 being effective.

B. Additional Experiments

Besides the experiments provided in Sec. 5, we report additional experiments on closed-world segmentation, open-world semantics segmentation, the OoDIS validation set, as well as a further ablation study of our uncertainty metric.

B.1. Closed World Segmentation

To compare P2F in closed-world segmentation, we compare it to the vanilla M2F and a naive Dirichlet Prior network (DPN) [39] for M2F for Cityscapes. We follow the training settings provided in [12] and use the Cityscapes [14] script for evaluation. Besides the mIoU, we also report the

[3] https://github.com/facebookresearch/Mask2Former

Model	mIoU ↑	cIoU ↑	PQ ↑	SQ ↑	RQ ↑
M2F	77.3	90.2	60.29	81.28	73.15
M2F DPN	64.5	88.4	50.29	67.20	61.70
P2F [ours]	**77.0**	**89.1**	**59.41**	**80.74**	**72.34**

Table 6. Comparison of M2F with different evidential heads on Cityscapes on Panoptic Segmentation. Best evidential heads are marked in bold.

	BDD Anomaly		COCO	
Method	Closed	Open	Closed	Open
U3HS [20]	29.16	16.32	33.19	22.77
P2F [ours]	**35.73**	**29.12**	**46.00**	**33.56**

Table 7. Open- and closed-world semantic segmentation comparison using **mIoU** metric for BDD Anomaly on COCO.

	Closed-W.			Open-W.		
Method	PQ↑	SQ ↑	RQ↑	PQ ↑	SQ ↑	RQ↑
U3HS [20]	**36.82**	80.93	16.32	17.81	75.59	23.63
P2F [ours]	32.67	**81.68**	**29.0**	**29.10**	**79.43**	**36.66**

Table 8. Closed and open-world evaluation on the BDD100k anomaly dataset [65].

category-wise IoU (cIoU). In Tab. 6, it can be seen that P2F performs similarly to M2F. This contrasts with the massive performance drop of the DPN head.

B.2. Open-World and Closed-World Segmentation

Besides the results reported on panoptic segmentation in Sec. 5, we study open- and closed-world semantic segmentation. In Tab. 7, we show the mIoU results for BDD100k [65] anomaly and COCO [35] with the left-out classes listed in Appendix A.3. Like in panoptic segmentation, P2F ranks the highest in all settings.

In Tab. 8 we report the panoptic quality metric of U3HS and P2F on BDD100k [65] anomaly dataset. U3HS shows a strong closed-world performance in PQ and RQ. However, P2F achieves the highest scores in the open-world setting as well as for SQ in the closed-world setting.

B.3. Further Experiments on L&F

To further compare using the setting of unseen L&F data as introduced by [20], we compare the reduced size evaluation with the baselines reported by [20] and an M2F confidence uncertainty. It can be seen that the additional baselines struggle with this task. The confidence baseline of M2F shows a surprisingly strong performance.

In addition to the setting reported by [20], we report the performance of P2F in comparison to other masked-based uncertainty methods in Tab. 10 using the full resolution of the L&F dataset. We train M2A without OOD data. We further report M2F using a classical confidence measure as uncertainty, which diverges for this task. We also report the uncertainty measures of RbA [41] and EAM [22]. These results show that all these methods profit heavily from the increased resolution. Nevertheless P2F maintains the highest scores for PQ and RQ.

Method	Assumptions	PQ ↑	SQ ↑	RQ ↑
EOPSN [28]	data, void	0*	0*	0*
OSIS [60]	data, void	1.45	65.11	2.23
U3HS [20]	none	7.94	64.24	12.37
M2A* [46]	none	9.91	73.45	13.49
M2F*	none	9.02	**75.34**	11.98
P2F [ours]	none	**11.22**	74.47	**15.06**

Table 9. Results on the Lost&Found (unseen) dataset with the settings of [20]. [28] and [60], are taken from [20]. * Uses P2F postprocessing.

Method	PQ ↑	SQ ↑	RQ ↑
M2A [46] + P2F post-proc.	21.06	**73.14**	28.79
M2F	0.00	0.00	0.00
EAM [22] + P2F post-proc.	18.39	71.35	25.78
RbA [41] + P2F post-proc.	15.54	70.65	21.99
P2F [ours]	**22.07**	69.54	**31.73**

Table 10. Unseen L&F performance metrics on full-size resolution. Semantic approaches require P2F post-processing.

B.4. OoDIS Validation Set

In addition to the official OoDIS [43] benchmark scores, we present results on the validation set of OoDIS comprising 100 L&F [44] instance anomaly labels. We report the scores for the OOD-free and extra-model-free U3HS and P2F in Tab. 11. For U3HS, we experimented with 4 different uncertainty thresholds and reported the best. The results of both models are slightly better compared to the benchmark evaluation on a larger L&F, while their ranking remains unchanged.

		No Aux.	No OOD		
Backbone	Model	Models	Data	AP ↑	AP50 ↑
R50	U3HS [20]	✓	✓	0.61	2.04
R50	P2F [ours]	✓	✓	**8.17**	**16.13**

Table 11. Comparison of different Anomaly Segmentation Methods on the validation set of OoDIS.

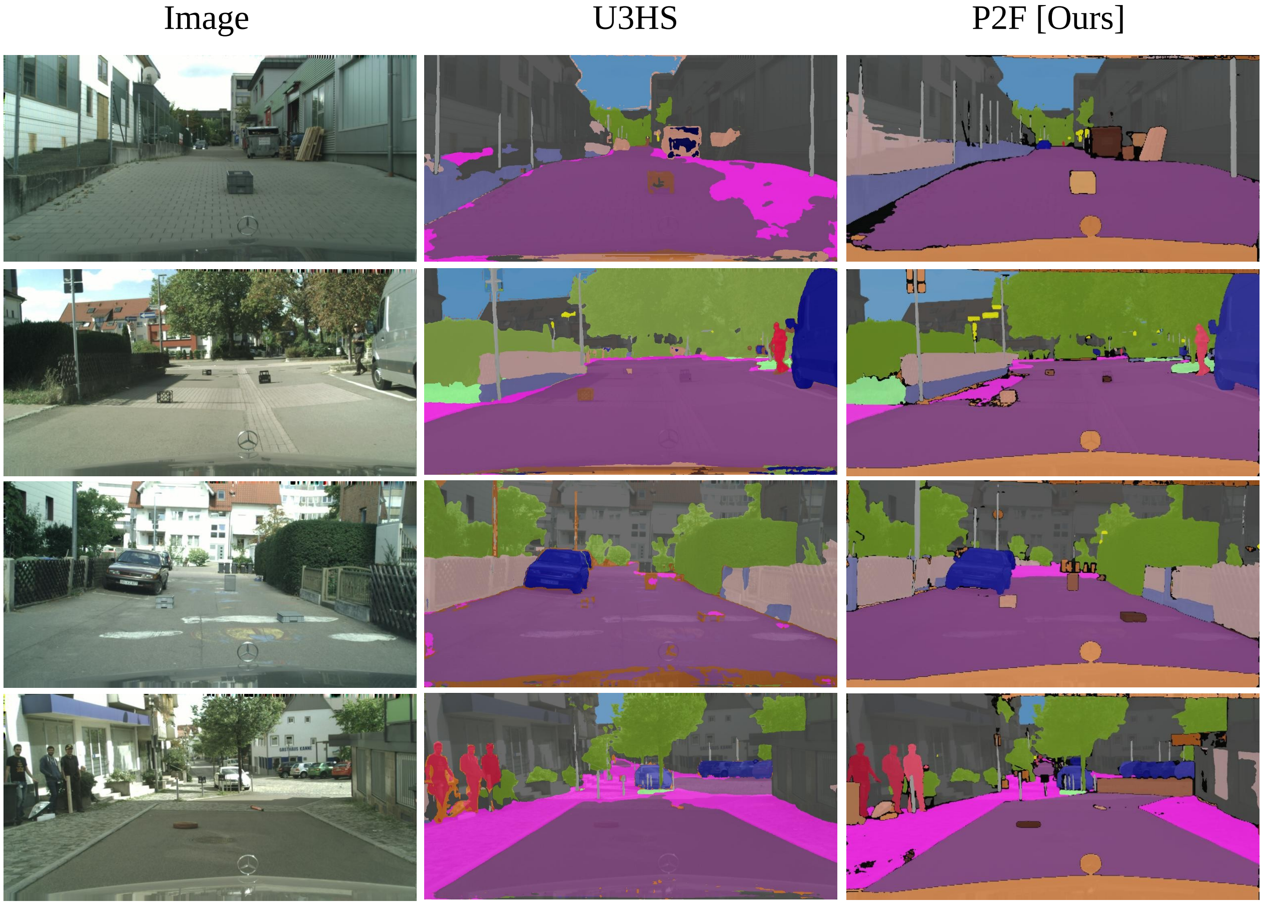

Figure 7. Open-World panoptic segmentation on L&F test set after training on Cityscapes [14].

B.5. Panic Open-Set Panoptic Segmentation

For additional evaluation of open-set panoptic segmentation, we utilized the novel PANIC benchmark [55]. The benchmark contains images of different resolutions from different cities in Germany and evaluates the PQ, SQ, and RQ scores of open-set objects. In Tab. 12, we show the official benchmark results of the open-set panoptic segmentation task. It can be seen that P2F significantly leads the benchmark introduced with the Con2Mav approach, with P2F delivering nearly double scores on PQ and RQ.

Method	PQ ↑	SQ ↑	RQ ↑
Con2Mav [55]	21.6	72.4	28.4
P2F [ours]	**52.9**	**87.1**	**56.7**

Table 12. Leaderboard of the PANIC [55] Open-set Panoptic Segmentation benchmark. Results from the test set.

B.6. Uncertainty Ablation Study

In this section, we conduct further ablation studies in addition to the reported study in Sec. 5. In Tab. 13, we compare the predictive uncertainty suggested by M2F [12] with the uncertainty of P2F. It can be seen that the P2F uncertainty is more effective on AP for the Road Anomaly split of SMIYC and provides a major improvement in the FPR of both datasets. This underlines the benefit of the evidential mask selection in our uncertainty. Overall, our P2F shows strong uncertainty statistics. Nevertheless, for minor semantics shifts like L&F, the mask filtering seems to be less important.

In Tab. 14, we compare our P2F uncertainty against three further uncertainty variants for anomaly segmentation and open-world panoptic segmentation. σ-unc. applies our combined uncertainty concept with mask matching according to Eq. (11a) with σ of the vanilla M2F model. Naive Beta describes the vanilla evidential uncertainty according to Eq. (6) and M2F* the vanilla M2F uncertainty with our postprocessing. It can be seen that our combination concept

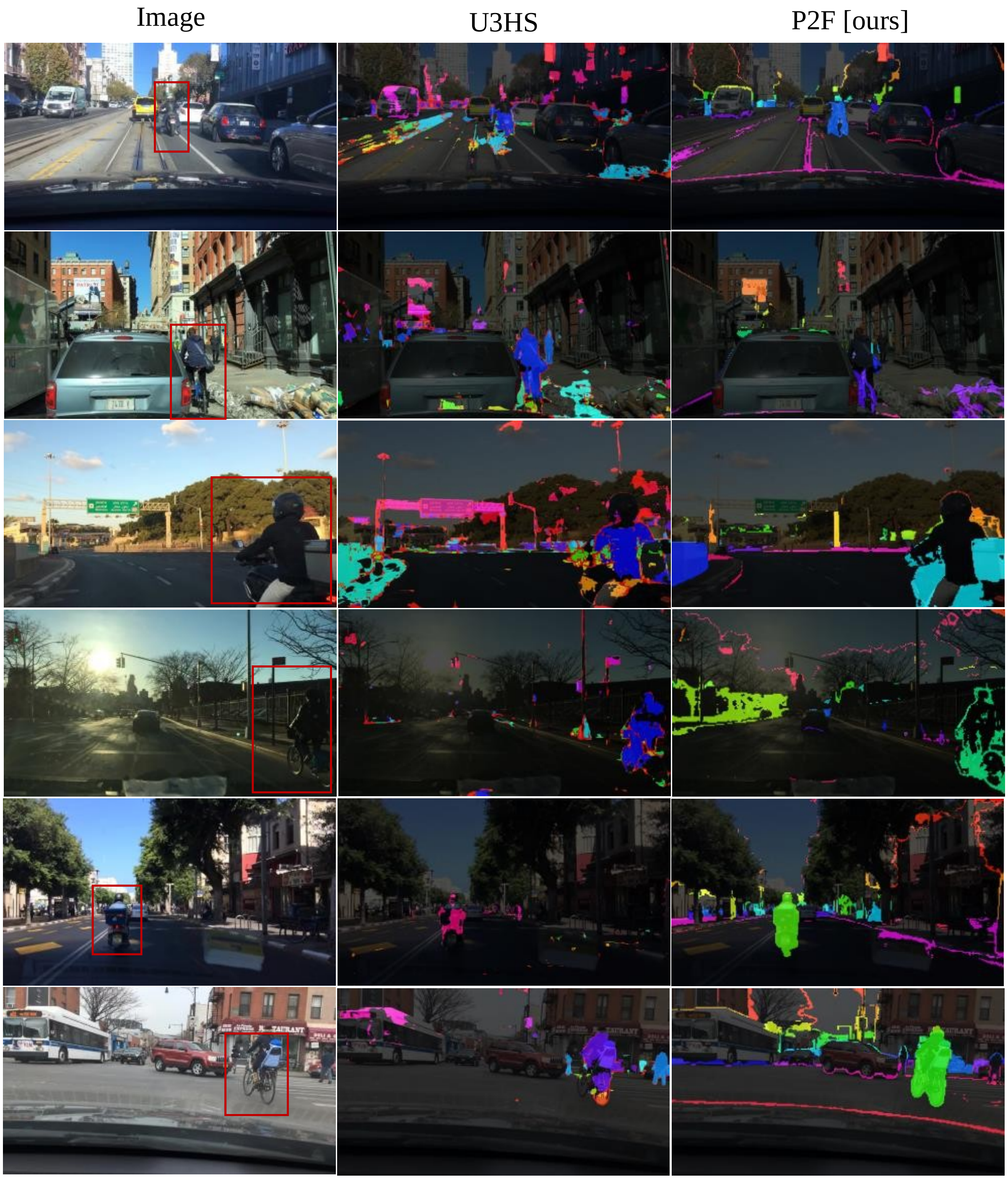

Figure 8. Visual comparison of anomaly instance segmentation on held-out classes on BDD [65], marked with a red box on the input image (left). The high diversity and unbalanced class distribution seem to confuse the DPN-based U3HS [20]. Nonetheless, U3HS showed a strong segmentation quality for the rare classes, like the traffic sign poles. P2F, managed to detect the unknown motorcycle or bicycles more precisely, given its less class imbalance affected Beta's prior approach.

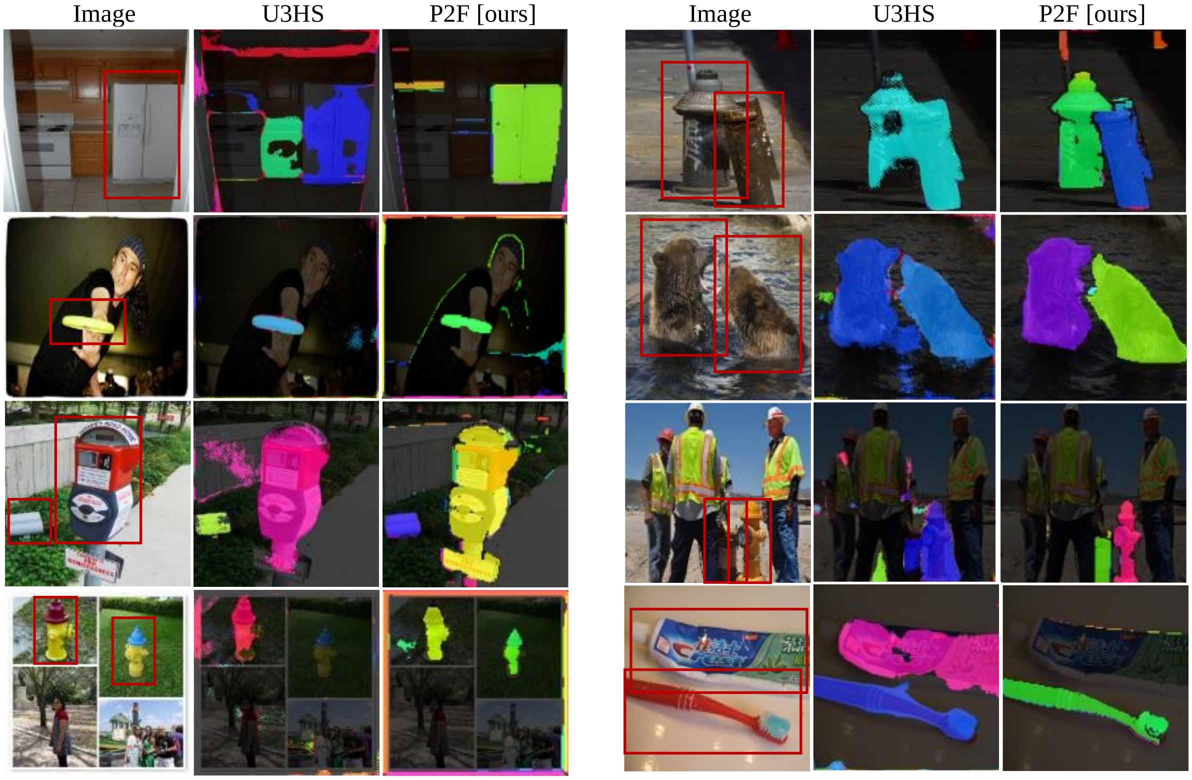

Figure 9. Visual comparison of anomaly instance segmentation on held-out classes on COCO [35], marked with a red box on the input image (left). It can be seen that U3HS shows a strong performance for individual and larger objects, but can get confused with related objects, like different kitchen objects in the top left. Additionally, it suppresses the combined uncertainty and distance approach uncertainty at the object border. With the direct mask supervision and Beta prior P2F provides a more robust separation of different instances, while improving the segmentation quality.

Method	Road Anomaly		FS L&F Obstacle	
	AP ↑	FPR95 ↓	AP ↑	FPR95 ↓
Prediction Uncertainty	48.6	62.2	61.2	88.4
P2F uncertainty [ours]	**58.6**	**46.3**	**66.6**	**16.8**

Table 13. Uncertainty comparison of P2F using the classical prediction uncertainty and P2F uncertainty definition.

of mask selection and evidential uncertainty outperforms our concept on M2F and the vanilla evidential uncertainty.

Method	FS L&F Obstacle		Panoptic L&F [19]		
	AP ↑	FPR95 ↓	PQ ↑	SQ ↑	RQ ↑
σ-unc.	22.78	27.04	6.17	**78.06**	7.91
naive Beta	12.81	23.04	0.94	62.94	1.50
M2F	48.60	62.20	9.02	75.34	11.98
P2F [ours]	**66.58**	**16.84**	**11.22**	74.47	**15.06**

Table 14. Uncertainty comparison of P2F with other naive uncertainties.

In Tab. 15, we evaluate the influence of the individual components in Eq. (11d). We evaluate our mask part $p^*_M(h,w)$ and classification part $p^*_C(h,w)$ compared with our P2F uncertainty. It can be seen that the class part performs for AP on large shifts from Cityscapes like SMIYC Road Anomaly, but suffers from a high FPR and panoptic detection of L&F. P2F outperforms both for Panoptic L&F and shows the lowest FPR for Road Anomaly.

Method	Road Anomaly		Panoptic L&F [20]		
	AP ↑	FPR95 ↓	PQ ↑	SQ ↑	RQ ↑
mask part	52.02	62.96	4.89	76.43	6.40
class part	**62.94**	100.00	0.00	0.00	0.00
P2F [ours]	58.60	**46.32**	**11.22**	74.47	**15.06**

Table 15. Uncertainty comparison of P2F using the classical prediction uncertainty and P2F uncertainty definition.

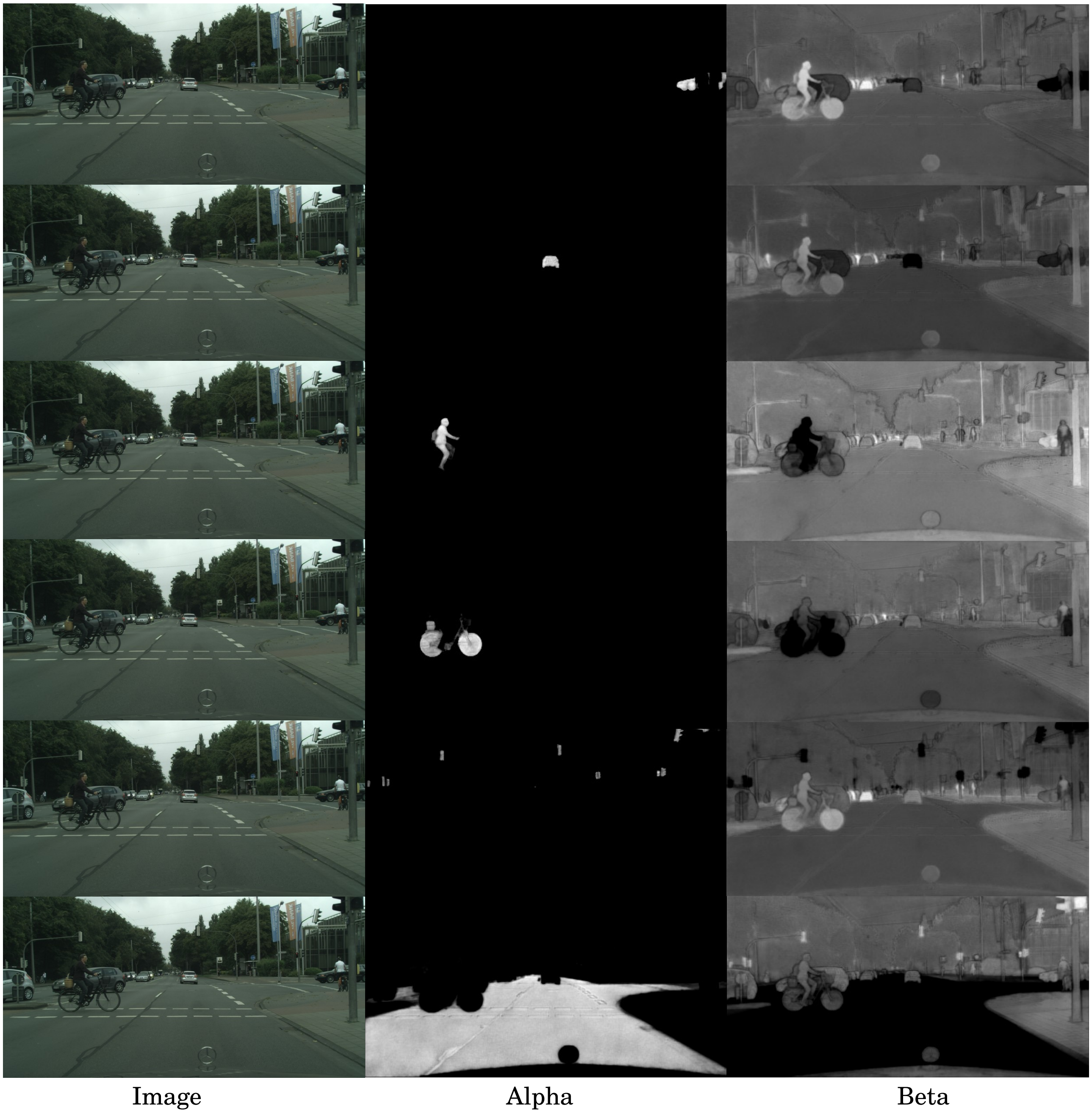

Figure 10. Mask Visualization of Alpha and Beta Masks on a single image from Cityscapes [14]. Alpha masks represent the positive correspondence of a pixel to a mask, while Beta masks emphasize the negative correspondence.

C. Additional Qualitative Assessment:

C.1. Lost and Found

In Fig. 7 we show visual results of open-world panoptic segmentation on the L&F test set. Anomaly predictions are marked in brown. For all predictions, including the anomaly prediction, different shades mark distinct instances. The L&F dataset includes the obstacles on the road as anomalies, but also other objects that have not been trained on during training on Cityscapes. These include trash cans, pipes, the back side of traffic signs, and pallets. For anomaly detection, U3HS uses a double threshold strategy where both the classification uncertainty and distances in the embedding are thresholded. For anomaly detection, both predictions need to surpass the individual thresholds. This can lead to not detecting anomalies, as in the second and third row of Fig. 7, but also results in less noise in the prediction. Further, the L&F features uncommon textures of the street, which results in confusion between sidewalk and road or incorrectly detected anomalies, as visible in row one for U3HS and in rows two and four for P2F.

C.2. Berkeley Deep Drive

We also provide further visual results of P2F and U3HS [20] on the BDD [65] dataset in Fig. 8 and on the COCO [35] Dataset in Fig. 9. Compared to Cityscapes, the BDD dataset comprises a higher variety of classes and different scenarios, making the anomaly instance segmentation task more challenging. This is demonstrated by the increased false positive rate of both models. Nonetheless, they are able to detect anomalies reasonably well in this challenging environment. However, the much smaller embedding dimension of U3HS has difficulties in identifying unseen instances, which results in cluttered predictions. This can be seen, for example, in rows one and three in Fig. 8. While the explicit regularization of the embeddings prevents uncertainty between instances, P2F might predict uncertainties between the transition or regions. Improve the uncertainty instance clustering to reject these remains for future work. Additionally, given the high-class imbalance of the dataset, U3HS tends to confuse rare classes with anomalies, for example, the traffic sign pole in row three. In contrast, P2F clearly detects the anomaly as its embeddings are high-dimensional and because of its additional mask supervision signal. Additionally, U3HS also marks the rider of the bicycles and motorcycles as an anomaly, whereas P2F separates well-visible humans, which is a natural phenomenon since human beings are present in the training set.

C.3. COCO

In comparison to BDD [65], the COCO [35] dataset features less cluttered images but includes a greater variety of classes. The total number of objects per image is lower, and the objects themselves are generally larger. In this setup, U3HS [65] and P2F perform well on simple images with only one anomaly as well as on more difficult images with several objects and anomaly instances. The distance thresholding of U3HS [20] and uncertainty thresholding suppresses uncertainty estimates at the edge of two classes. The pure uncertainty thresholding of P2F marks them as an anomaly, as visible in row 2 in Fig. 9. P2F consistently shows a superior object segmentation as well as instance separation compared to U3HS [20]. This underlines the effectiveness of our proposed Beta prior and the high quality of the embedding space learned by P2F.

C.4. Mask Visualization

In Appendix B.6, we visualize different prior masks $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ on a Cityscapes image. The first two rows show detections of individual cars. Notably, the $\boldsymbol{\beta}$-mask tends to suppress other cars less strongly than it does objects from different classes. This underlines the effective instance recognition of P2F for common classes. Rows three and four depict predictions for a human on a bicycle and the bicycle itself. The last two rows correspond to “stuff” classes; these often cover multiple instances within the same category, such as the traffic lights in row five. Finally, the last row shows the prediction for the street.

References

[1] Jan Ackermann, Christos Sakaridis, and Fisher Yu. Maskomaly: Zero-shot mask anomaly segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2023. 3

[2] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *Proceedings of the Internation Conference on Machine Learning (ICML)*. PMLR, 2020. 2

[3] Hermann Blum, Paul Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. The fishyscapes benchmark: Measuring blind spots in semantic segmentation. pages 3119–3135, 2021. 2, 6, 7, 9

[4] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. In *Proceedings of Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. 9

[5] 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 (ICCV)*, 2021. 2, 6, 7, 9, 10

[6] Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2020. 2

[7] Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural posterior network: Deep bayesian uncertainty for exponential family distributions. In *Proceedings of the Internation Conference on Machine Learning (ICML)*. PMLR, 2022. 2

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *Arxiv*, 1706.05587, 2017. 2

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 6, 9

[10] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[11] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3

[12] 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 Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 5, 9, 11, 13

[13] Hyunjun Choi, Hawook Jeong, and Jin Young Choi. Balanced energy regularization loss for out-of-distribution detection. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 8, 9, 11, 13, 16

[15] Anja Delić, Matej Grcić, and Siniša Šegvić. Outlier detection by ensembling uncertainty with negative objectness. *Proceedings of the British Machine Vision Conference (BMVC)*, 2024. 2, 3

[16] Arthur P Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968. 3

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226–231. AAAI Press, 1996. 6, 11

[18] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the Internation Conference on Machine Learning (ICML)*. PMLR, 2016. 2

[19] Stefano Gasperini, Jan Haug, Mohammad-Ali Nikouei Mahani, Alvaro Marcos-Ramiro, Nassir Navab, Benjamin Busam, and Federico Tombari. Certainnet: Sampling-free uncertainty estimation for object detection. *IEEE Robotics and Automation Letters*, 7(2):698–705, 2022. 2

[20] Stefano Gasperini, Alvaro Marcos-Ramiro, Michael Schmidt, Nassir Navab, Benjamin Busam, and Federico Tombari. Segmenting known objects and unseen unknowns without prior knowledge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 17

[21] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *Artif Intell Rev*, 56, 2023. 2

[22] Matej Grcić, Josip Šarić, and Siniša Šegvić. On advantages of mask-level recognition for outlier-aware segmentation. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 3, 6, 8, 9, 10, 12

[23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the Internation Conference on Machine Learning (ICML)*. PMLR, 2017. 3

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6, 9

[25] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions

far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019. 3

[26] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the Internation Conference on Learning and Representation (ICLR)*, 2017. 3, 9, 10

[27] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *Proceedings of the Internation Conference on Machine Learning (ICML)*. PMLR, 2022. 6, 10

[28] Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, and Bohyung Han. Exemplar-based open-set panoptic segmentation network. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 7, 12

[29] 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 (ICCV)*, 2021. 3, 6, 9

[30] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2017. 3

[31] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 7

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

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

[34] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Gmmseg: Gaussian mixture based generative semantic segmentation models. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2022. 2

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 7, 8, 10, 12, 15, 17

[36] Jeremiah Zhe Liu, Zi Lin, Google Research, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2020. 2

[37] Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2020. 3

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the Internation Conference on Learning and Representation (ICLR)*, 2019. 11

[39] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2018. 2, 11

[40] Jishnu Mukhoti, Andreas Kirsch, Joost Van Amersfoort, Philip H S Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4

[41] Nazir Nayal, Mısra Yavuz, João F. Henriques, and Fatma Güney. RbA: Segmenting unknown regions rejected by all. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 6, 8, 9, 10, 12

[42] Alexey Nekrasov, Alexander Hermans, Lars Kuhnert, and Bastian Leibe. Ugains: Uncertainty guided anomaly instance segmentation. In *German Conference on Pattern Recognition (GCPR)*, 2023. 1, 3, 7

[43] Alexey Nekrasov, Rui Zhou, Miriam Ackermann, Alexander Hermans, Bastian Leibe, and Matthias Rottmann. Oodis: Anomaly instance segmentation benchmark. *Arvix*, 2406.11835, 2024. 2, 7, 10, 12

[44] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: Detecting small road hazards for self-driving vehicles. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016. 1, 2, 6, 7, 10, 12

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the Internation Conference on Machine Learning (ICML)*. PMLR, 2021. 1

[46] Shyam Nandan Rai, Fabio Cermelli, Dario Fontanel, Carlo Masone, and Barbara Caputo. Unmasking anomalies in road-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 6, 7, 8, 9, 10, 12

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer Verlag, 2015. 2

[48] Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. Second-order uncertainty quantification: A distance-based approach. In *Proceedings of the Internation Conference on Machine Learning (ICML)*. PMLR, 2023. 4

[49] Sebastian Schmidt and Stephan Günnemann. Stream-based active learning by exploiting temporal properties in perception with temporal predicted loss. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2023. 2

[50] Sebastian Schmidt, Qing Rao, Julian Tatsch, and Alois Knoll. Advanced active learning strategies for object detection. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2020. 2

[51] Sebastian Schmidt, Leonard Schenk, Leonard Schwinn, and Stephan Günnemann. A unified approach towards active learning and out-of-distribution detection. *Arxiv*, 2405.11337, 2024. 2

[52] Sebastian Schmidt, Ludwig Stumpp, Diego Valverde, and Stephan Günnemann. Deep sensor fusion with constraint safety bounds for high precision localization. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12256–12262. IEEE, 2024. 1

[53] Sebastian Schmidt, Leonard Schenk, Leo Schwinn, and Stephan Günnemann. Joint out-of-distribution filtering and data discovery active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[54] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2018. 2, 3

[55] Matteo Sodano, Federico Magistri, Jens Behley, and Cyrill Stachniss. Open-World Panoptic Segmentation. *Arvix*, 2412.12740, 2024. 2, 8, 13

[56] Matteo Sodano, Federico Magistri, Lucas Nunes, Jens Behley, and Cyrill Stachniss. Open-world semantic segmentation including class similarity. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 6

[57] Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, and Stephan Günnemann. Graph posterior network: Bayesian predictive uncertainty for node classification. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2021. 2

[58] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248, 2017. 3

[59] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[60] Kelvin Wong, Shenlong Wang, Mengye Ren, Ming Liang, and Raquel Urtasun. Identifying unknown instances for autonomous driving. In *Proceedings of the Conference on Robot Learning (CoRL)*, pages 384–393. PMLR, 2020. 12

[61] C. Zach X. Liu, Y. Lochman. GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[62] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3

[63] Hai-Ming Xu, Hao Chen, Lingqiao Liu, and Yufei Yin. Dual decision improves open-set panoptic segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2022. 1, 2, 3, 7, 10

[64] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. Openood: Benchmarking generalized out-of-distribution detection. In *Proceedings of Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022. 2

[65] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2020. 10, 12, 14, 17

[66] Hao Zhang, Fang Li, Lu Qi, Ming-Hsuan Yang, and Narendra Ahuja. Csl: Class-agnostic structure-constrained learning for segmentation including the unseen. In *Proceeding of the AAAI Conference on Artificial Intelligence*, 2024. 2